A wild bootstrap approach for nonparametric repeated measurements

Sarah Friedrich [a], Frank Konietschke [b], Markus Pauly [a,*]

[a] *Institute of Statistics, Ulm University, Helmholtzstr. 20, 89081 Ulm, Germany*
[b] *The University of Texas at Dallas, 0800 W. Campbell Road, Richardson, TX 75080-3021, USA*

1. Motivation and introduction

When planning experiments in behavioral, medical or psychological sciences repeated measures designs and split-plot plans are often preferred because fewer experimental units (subjects) are required to obtain 'sufficient' numbers of observations (Stevens, 2012; Howell, 2013; Hedeker and Gibbons, 2006; Davis, 2002). Such data are typically analyzed by mean-based multivariate analysis-of-variance methods (MANOVA), repeated measures ANOVA or linear mixed models requiring certain assumptions on the underlying parametric distributions, see e.g. the monographs of Davis (2002), Hedeker and Gibbons (2006) or Johnson and Wichern (2007). However, as e.g. pointed out by Kherad-Pajouh and Renaud (2015) "*it is likely that for this kind of data, the parametric assumptions are not satisfied*" so that the "*result of the methods (…) might not be reliable*", see also Xu and Cui (2008), Suo et al. (2013) or Konietschke et al. (2015) for related comments. Furthermore, parametric methods usually require a specific covariance structure of the data, e.g., compound symmetry, sphericity or equal covariance matrices across the different groups. The type of covariance matrix is hard to justify in real applications. If the assumed covariance matrix is mis-specified, the estimator of the covariance matrix is biased, which results in a liberal or conservative behavior of the test. In particular, Oberfeld and Franke (2013) point out that the "*covariance structure of the data*

* Corresponding author.
E-mail address: markus.pauly@uni-ulm.de (M. Pauly).

is important for the validity of the tests", see also Keselman et al. (2001) and the references cited therein. Therefore plenty of robustifications and/or approximations for more general mean-based analysis in various repeated measures designs have been proposed, see Huynh and Feldt (1976), Huynh (1978), Lecoutre (1991), Kenward and Roger (1997), Keselman et al. (2000), Pesarin (2001), Vallejo and Ato (2006), Xu and Cui (2008), Kenward and Roger (2009), Arnau et al. (2012), Chi et al. (2012), Pesarin and Salmaso (2012), Brombin et al. (2013), Konietschke et al. (2015), Pauly et al. (2015) or Friedrich et al. (2015), among others.

If count, ordinal, ordered categorial or score data are present, however, these approaches show their limits since means are neither meaningful nor adequate measures of deviations between groups or treatments. In such a situation, nonparametric rank-based methods are the preferred choice for making statistical inference. Such methods are robust, applicable to all kinds of data and the corresponding test results are invariant under monotone transformations of the data. In particular, Akritas and Arnold (1994), Akritas and Brunner (1997), Brunner et al. (1999), Brunner (2001), Brunner et al. (2002) and Akritas (2011) propose rank-based methods for testing nonparametric hypotheses formulated in terms of distribution functions for factorial longitudinal data. The procedures are valid for the analysis of metric, count, ordinal, score or even ordered categorial data in a unified way. The two proposed statistics therein, however, have drawbacks: The Wald-type statistic provides an asymptotically valid test, but very large sample sizes are required for accurate test decisions. The method tends to be very liberal in case of small and moderate numbers of observations, see e.g. Brunner (2001). Moreover, it is only applicable in case of regular covariance matrices. The latter drawback is not shared by the ANOVA-type statistic which turns out to be an approximation that is in general not asymptotically correct and results in rather conservative test decisions for small sample sizes, see Brunner (2001). Since sample sizes are often rather small compared to the number of time points in practical applications, it is thus the aim of the present paper to (i) enhance the small sample performances and (ii) the asymptotic properties of these testing procedures. To this end, we adopt a nonparametric wild bootstrap resampling technique which is already known for leading to the above desired enhancements in mean-based regression analyses, see Wu (1986), Liu (1988), Mammen (1993b), Flachaire (2005), Davidson and Flachaire (2008), Cameron et al. (2008) and Cameron and Miller (2015). Here its application to the above described statistics leads to our goals (i) and (ii) while preserving their general applicability in factorial repeated measures designs.

As a motivating example, we consider the shoulder tip pain trial reported by Lumley (1996). In this trial, the characteristic pain in the shoulder tip after laparoscopic surgery was observed in $N = 41$ patients during $t = 6$ time points. After randomization, $n_1 = 22$ patients (14 female and 8 male) received the active treatment (treatment = 'Yes') and $n_2 = 19$ (11 female and 8 male) patients belonged to the control group (treatment = 'No'). Thus, data was observed in an elaborate factorial design, with stratifying whole-plot factors *Treatment* and *Gender*, and sub-plot factor *Time*. For every patient enrolled in the trial, $t = 6$ possibly correlated repeated measurements were observed. The pain was measured on an ordinal scale ranging from 1 (low) to 5 (high). The lower the score, the better the clinical record. The observed score distribution is displayed in Fig. 1.

It can be readily seen from the boxplots displayed in Fig. 1 that the scores given under treatment tend to be lower than those under control. However, the investigation of statistical interactions between the factors *treatment, gender* and *time* are of major interest in this experiment. Since mean-based approaches are inappropriate for making statistical inference with ordered categorial data, nonparametric ranking methods are preferred.

The paper is organized as follows: In the next section, we state the statistical model as well as the hypotheses and test statistics considered. In Section 3 we describe the wild bootstrap procedure. Simulation results are displayed in Section 4 as well as in the supplementary material (see Appendix B) and a detailed analysis of the data example is given in Section 5. Finally, we discuss the results in Section 6. All proofs are deferred to Appendix A.

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

2. Statistical model, hypotheses and statistics

2.1. Statistical model and hypotheses

To establish the general nonparametric repeated measures model with a different groups and t different time points, let

$$\boldsymbol{X}_{ik} = (X_{ik1}, \ldots, X_{ikt})', \quad i = 1, \ldots, a, \ k = 1, \ldots, n_i,$$

denote the random vector belonging to the kth subject in group i. The $N = \sum_{i=1}^{a} n_i$ random vectors are assumed to be independent with marginals

$$X_{iks} \sim F_{is}, \quad i = 1, \ldots, a, \ k = 1, \ldots, n_i, \ s = 1, \ldots, t.$$

For convenience, we collect the observations $\boldsymbol{X}_{ik}$ in larger vectors

$$\boldsymbol{X}_i = (\boldsymbol{X}_{i1}', \ldots, \boldsymbol{X}_{in_i}')', \quad \text{and} \quad \boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_a')', \tag{2.1}$$

containing all the information of group i ($i = 1, \ldots, a$) and the pooled sample, respectively.

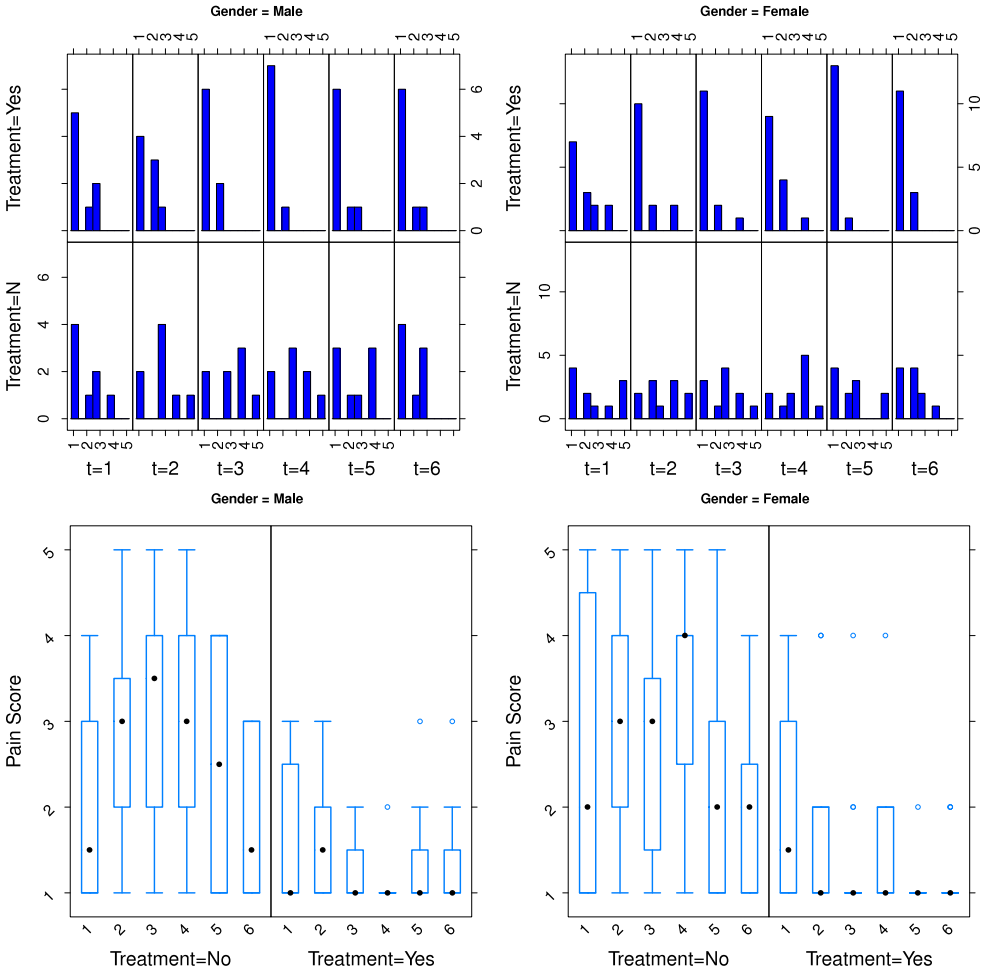

Fig. 1. Frequencies of the pain scores observed in the shoulder pain trial (Lumley, 1996).

In this set-up, null hypotheses are formulated by $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$, where $\boldsymbol{F} = (F_{11}, \ldots, F_{at})'$ denotes the vector of the distribution functions F_{is}, $i = 1, \ldots, a$, $s = 1, \ldots, t$ and $\boldsymbol{C}$ is a suitable hypothesis matrix. Rank-statistics for testing these hypotheses are derived by considering estimates of the relative marginal effects $\boldsymbol{p} = (p_{11}, \ldots, p_{at})'$, where $p_{is} = \int H_N dF_{is}$. Here, $H_N(x) = \frac{1}{t \cdot N} \sum_{i=1}^{a} \sum_{s=1}^{t} n_i F_{is}(x)$ denotes the (weighted) mean distribution function of the whole experiment. If $p_{is} < p_{is'}$ for some $s \neq s'$, then the (random) measurements in group i at time s tend to result in smaller values than those at time s'. If $p_{is} = p_{is'}$, no data tend to be smaller or larger. The effects p_{is} are estimated by

$$\widehat{p}_{is} = \frac{1}{tN} \left(\overline{R}_{i \cdot s} - \frac{1}{2} \right),$$

where R_{iks} denotes the (mid-)rank of X_{iks} among all tN observations and $\overline{R}_{i \cdot s} = \frac{1}{n_i} \sum_{k=1}^{n_i} R_{iks}$. A more detailed theoretical derivation of the relative treatment effects is given in the Appendix A. For convenience, the estimators are collected in the vector $\widehat{\boldsymbol{p}} = (\widehat{p}_{11}, \ldots, \widehat{p}_{at})'$. Assuming the usual sample size condition

$$\frac{n_i}{N} \to \kappa_i \in (0, 1), \quad \text{for all } i = 1, \ldots, a, \tag{2.2}$$

Akritas and Brunner (1997) have shown that $\sqrt{N}\boldsymbol{C}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ follows, asymptotically, as $N \to \infty$, a multivariate normal distribution with expectation $\boldsymbol{0}$ and covariance matrix $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'$ under the hypothesis H_0^F. Here, the matrix

$$\boldsymbol{\Sigma} = \bigoplus_{i=1}^{a} \kappa_i^{-1} \boldsymbol{V}_i \tag{2.3}$$

is the weighted block diagonal matrix of the covariance matrices $\boldsymbol{V}_i = Cov(\boldsymbol{Y}_{ik})$ of the random vectors $\boldsymbol{Y}_{ik} = (H(X_{ik1}), \ldots, H(X_{ikt}))'$ and $H = \frac{1}{t} \sum_{i=1}^{a} \sum_{s=1}^{t} \kappa_i F_{is}$ is the limit distribution function of H_N under (2.2).

2.2. Statistics and asymptotics

In this section, suitable test statistics for testing the null hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$ will be introduced. First, the Wald-type statistic (WTS) of Akritas and Arnold (1994) and Brunner and Puri (2001)

$$Q_N = N\widehat{\boldsymbol{p}}'\boldsymbol{C}'(\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}')^{+}\boldsymbol{C}\widehat{\boldsymbol{p}} \tag{2.4}$$

is considered, where $\boldsymbol{M}^{+}$ denotes the Moore–Penrose inverse of a matrix $\boldsymbol{M}$. Here, $\widehat{\boldsymbol{\Sigma}} = N\bigoplus_{i=1}^{a}\frac{1}{n_i}\widehat{\boldsymbol{V}}_i$ denotes the weighted direct sum of the empirical covariance matrices

$$\widehat{\boldsymbol{V}}_i = \frac{1}{(tN)^2(n_i-1)}\sum_{k=1}^{n_i}(\boldsymbol{R}_{ik}-\overline{\boldsymbol{R}}_{i\cdot})(\boldsymbol{R}_{ik}-\overline{\boldsymbol{R}}_{i\cdot})', \quad i=1,\ldots,a,$$

which is a consistent estimator of the limiting covariance matrix $\boldsymbol{V}_i$. The asymptotic distribution of the WTS is provided in the next theorem:

Theorem 2.1. *Assume (2.2) and $\boldsymbol{V}_i > 0$ for all $i = 1,\ldots,a$. Under the hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$, the WTS in (2.4) has, asymptotically as $N \to \infty$, a central χ_f^2-distribution with $f = \mathrm{rank}(\boldsymbol{C})$ degrees of freedom, i.e.,*

$$Q_N \xrightarrow{d} Q \sim \chi^2_{\mathrm{rank}(\boldsymbol{C})}. \tag{2.5}$$

Due to the weak performance of the WTS for small sample sizes and its restriction to non-singular covariance matrices, Brunner et al. (1997) and Brunner and Langer (2000) propose the so-called ANOVA-type test statistic (ATS). The idea is to first drop the estimated covariance matrix in (2.4), resulting in the following statistic:

$$A_N = N\widehat{\boldsymbol{p}}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{+}\boldsymbol{C}\widehat{\boldsymbol{p}} =: N\widehat{\boldsymbol{p}}'\boldsymbol{T}\widehat{\boldsymbol{p}}. \tag{2.6}$$

The asymptotic distribution of A_N is given in the next theorem (Brunner and Puri, 2001, Theorem 2.7):

Theorem 2.2. *Under the hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$ and assumption (2.2), the statistic A_N has, asymptotically as $N \to \infty$, the same distribution as a weighted sum of χ_1^2-distributed random variables, i.e.,*

$$A_N \xrightarrow{d} A \sim \sum_{i=1}^{a}\sum_{s=1}^{t}\lambda_{is}\xi_{is}, \tag{2.7}$$

where $\xi_{is} \overset{i.i.d.}{\sim} \chi_1^2$ and the weights λ_{is} are the eigenvalues of $\boldsymbol{T}\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is defined in (2.3).

Since the eigenvalues λ_{is} are unknown, the limiting distribution is approximated by a weighted $g\cdot\chi_f^2$ distribution, where g and f are estimated from the data such that the first two moments of the limiting distribution of the ATS and $g\cdot\chi_f^2$ coincide (Box, 1954). Finally, the distribution of the ANOVA-type statistic

$$F_N = \frac{N}{\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}})}\widehat{\boldsymbol{p}}'\boldsymbol{T}\widehat{\boldsymbol{p}}$$

can be approximated by a central $F(\hat{f},\infty)$-distribution with $\hat{f} = \frac{(\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}}))^2}{\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}}\boldsymbol{T}\widehat{\boldsymbol{\Sigma}})}$ degrees of freedom under the null hypothesis H_0^F, see Brunner et al. (1999). For testing the main effects of the whole-plot factors or interactions involving only whole-plot factors, the distribution of the ATS can be further approximated by an $F(\hat{f},\hat{f}_0)$ distribution with $\hat{f}_0$ as in Brunner et al. (1997). Compared to the WTS the corresponding ATS has the advantage of being applicable in case of a singular covariance matrix $\boldsymbol{\Sigma}$. The ATS is implemented in the R-package **nparLD** (Noguchi et al., 2012) for the analysis of factorial repeated measures designs. Furthermore, the rank-based ATS can be computed using SAS (SAS Institute Inc., 2003), e.g. SAS `PROC MIXED` using the option *ANOVAF*. Note that the ranks of the data are obtained via `PROC RANK` and used within the model statement.

Note that in contrast to the WTS, the corresponding ATS test provides in general no asymptotic level α test, which is a severe drawback of this procedure. The finite sample distributions of both the WTS and the ATS can be approximated by a wild bootstrap procedure, thus leading to more accurate statistical tests. This will be explained in the next section.

3. The wild bootstrap procedure

Resampling techniques are widely known to induce *robust* inference procedures, even for small sample sizes, see e.g. their extensive treatment in Davison and Hinkley (1997), Davison et al. (2003), Good (2006) or Manly (2006). Typically, the idea of

the methods is as follows: Instead of computing the p-value (or critical value) from an approximate distribution of a statistic, the p-value is computed from a resampling distribution of the statistic. Thus, the resampling test can only be consistent, if both the distribution of the test statistic and its (conditional) resampling distribution coincide, at least asymptotically. In order to achieve this goal, several different resampling techniques have been explored in the literature: nonparametric bootstrap (randomly drawing with replacement), parametric bootstrap, permutation and randomization methods, cross validation and many more. Simulation studies indicate that the use of Efron's nonparametric bootstrap (Efron, 1979) results in liberal conclusions in the present setup. Therefore, we did not further investigate the conventional bootstrap. This result is in concordance with recent results for general MANOVA (Konietschke et al., 2015) in a semiparametric framework. For nonparametric bivariate data, Konietschke and Pauly (2012) investigated a studentized permutation approach based on rank statistics. Their bivariate model is included in ours by setting $a = 1$ and $t = 2$. Simulation results indicated that the resampling version greatly improves the classical rank-test for small sample sizes. The permutation method is based on randomly permuting the observed components X_{1k1} and X_{1k2} from subject k. Now, computing the differences $D_k = X_{1k1} - X_{1k2}$, it follows that their permutation approach is distributional identical to multiplying the differences with random signs ϵ_k, with $P(\epsilon_{ik} = 1) = P(\epsilon_{ik} = -1) = \frac{1}{2}$. This perception led to generalizing their method to our setting with general nonparametric factorial longitudinal data. Such resampling methods, which are based on multiplying the (fixed) data with random signs, i.e., using Rademacher distributed random weights (Davidson and Flachaire, 2008), are a specific *wild bootstrap technique*. Note that earlier wild bootstrap versions used different weights satisfying different moment conditions, see e.g. Wu (1986), Liu (1988) or Mammen (1993a). Typically, the choice of weights depends on the specific situation. In our nonparametric setting we found a specific preference for Rademacher weights in our simulation study. They have the additional advantage of leading to a finitely exact test if the multiplied random variables ($\boldsymbol{Z}_{ik}$ below) are 0-symmetric under the null, see e.g. Janssen (1999) or Lehmann and Romano (2005).

Furthermore, these resampling methods are motivated by the residual bootstrap commonly applied in regression analysis (Wu, 1986; Mammen, 1993b; Janssen, 1999; Flachaire, 2005; Davidson and Flachaire, 2008; Cameron et al., 2008), and in time-series testing problems (Kreiss and Paparoditis, 2011). It is also proposed in the context of survival analysis (Lin, 1997; Martinussen and Scheike, 2007; Pauly, 2011; Beyersmann et al., 2013; Dobler and Pauly, 2014; Dobler et al., 2015), and recently for the selection of biomarkers in early diagnostic trials (Zapf et al., 2015). The approach will be explained in the following.

Let $\boldsymbol{Z}_{ik} = \left(\boldsymbol{R}_{ik} - \overline{\boldsymbol{R}}_{i\cdot}\right),\ i = 1, \ldots, a;\ k = 1, \ldots, n_i$ denote the centered rank vectors and let ϵ_{ik} denote independent and identically distributed random signs; thus fulfilling $E(\epsilon_{11}) = 0$ and $Var(\epsilon_{11}) = 1$. We restrict ourselves to this specific kind of weights since they showed the best finite sample performance in the scenarios considered here (see also Davidson and Flachaire, 2008 for a similar observation in regression models). However, the subsequent results can easily be extended to other choices of weights fulfilling $E(\epsilon_{11}) = 0$ and $Var(\epsilon_{11}) = 1$. Now, consider the resampling vectors

$$\boldsymbol{Z}^*_{ik} = \epsilon_{ik} \cdot \boldsymbol{Z}_{ik}, \quad i = 1, \ldots, a;\ k = 1, \ldots, n_i,$$

which depict a conditional distribution of the centered rank vectors $\boldsymbol{Z}_{ik}$ around zero. The shape of the distribution depends on $\boldsymbol{Z}_{ik}$ and particularly on the shape of the distribution of the random weights. Since the ϵ'_{ik}s are random signs, the distribution of $\boldsymbol{Z}^*_{ik}$ is a symmetrization of the fixed vectors $\boldsymbol{Z}_{ik}$. Now, let

$$\hat{\boldsymbol{p}}^{\epsilon}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \frac{ik}{tN}\left(\boldsymbol{R}_{ik} - \overline{\boldsymbol{R}}_{i\cdot}\right) = \frac{1}{n_i}\sum_{k=1}^{n_i} \frac{1}{tN}\boldsymbol{Z}^*_{ik}$$

denote the resampling equivalent of the relative effect estimators $\widehat{\boldsymbol{p}}_i$; and let $\widehat{\boldsymbol{\Sigma}} = N\bigoplus_{i=1}^{a} \frac{1}{n_i}\widehat{\boldsymbol{V}}^{\epsilon}_i$ denote the direct sum of the empirical covariance matrices

$$\widehat{\boldsymbol{V}}^{\epsilon}_i = \frac{1}{(tN)^2(n_i-1)}\sum_{k=1}^{n_i} \boldsymbol{Z}^*_{ik} - \overline{\boldsymbol{Z}}^*_{i\cdot} \quad \boldsymbol{Z}_{ik} - \overline{\boldsymbol{Z}}^*_{i\cdot}{}', \quad i = 1, \ldots, a,$$

of the vectors $\boldsymbol{Z}^*_{ik}$, respectively. For convenience, the vectors $\hat{\boldsymbol{p}}^{\epsilon}_i$ are collected in the vector $\hat{\boldsymbol{p}}^{\epsilon} = (\hat{\boldsymbol{p}}^{\epsilon}_1, \ldots, \hat{\boldsymbol{p}}^{\epsilon}_a)'$. This bootstrap method corresponds to the wild cluster bootstrap proposed by Cameron et al. (2008) for semiparametric regression problems. In this sense we may also call our approach more specifically *nonparametric wild cluster bootstrap* of the individual rank vectors $\boldsymbol{R}_{ik}$. In the next theorem, the conditional multivariate distribution of $\sqrt{N}\hat{\boldsymbol{p}}^{\epsilon}$ will be examined.

Theorem 3.1. *The conditional distribution of $\sqrt{N}\hat{\boldsymbol{p}}^{\epsilon}$, given the data $\boldsymbol{X}$, is, asymptotically, as $N \to \infty$, the multivariate $N(0, \boldsymbol{\Sigma})$ distribution, in probability.*

Theorem 3.1 implies that both the distributions of $\sqrt{N}\boldsymbol{C}\hat{\boldsymbol{p}}^{\epsilon}$ and $\sqrt{N}\boldsymbol{C}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ are asymptotically identical under the hypothesis H_0^F. Furthermore, the asymptotic distribution of $\sqrt{N}\boldsymbol{C}\hat{\boldsymbol{p}}^{\epsilon}$ is independent from the distribution of the data $\boldsymbol{X}$. These results can now be used to derive the wild bootstrap versions of both the Wald-type statistic (WWTS)

$$Q^{\epsilon}_N = N(\hat{\boldsymbol{p}}^{\epsilon})'\boldsymbol{C}'(\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}^{\epsilon}\boldsymbol{C}')^{+}\boldsymbol{C}\hat{\boldsymbol{p}}^{\epsilon}, \tag{3.8}$$

and the ANOVA-type statistic (WATS)

$$F_N^\epsilon = \frac{N}{\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{\Sigma}})}(\hat{\boldsymbol{p}}^\epsilon)'\boldsymbol{T}\hat{\boldsymbol{p}}^\epsilon. \tag{3.9}$$

It will be shown in the subsequent theorems that both the conditional distributions of the statistics Q_N^ϵ and F_N^ϵ mimic the asymptotic null distributions of the WTS and the ATS given in Theorems 2.1 and 2.2, respectively.

Theorem 3.2. *Under Assumption (2.2) and $\boldsymbol{V}_i > 0$ for all $i = 1, \ldots, a$, the conditional distribution of Q_N^ϵ converges weakly to the central χ_f^2-distribution with $f = rank(\boldsymbol{C})$ degrees of freedom in probability for any underlying value $\boldsymbol{p} \in \mathbb{R}^{at}$, i.e. we have*

$$\sup_{x\in\mathbb{R}} |P_{\boldsymbol{p}}(Q_N^\epsilon \le x|\boldsymbol{X}) - P_{H_0}(Q_N \le x)| \xrightarrow{P} 0, \tag{3.10}$$

where $P_{H_0}(Q_N \le x)$ denotes the unconditional null distribution function of Q_N under H_0.

Theorem 3.3. *Under Assumption (2.2) the conditional distribution of F_N^ϵ converges weakly to the null distribution of F_N in probability for any underlying value $\boldsymbol{p} \in \mathbb{R}^{at}$, i.e. we have*

$$\sup_{x\in\mathbb{R}} |P_{\boldsymbol{p}}(F_N^\epsilon \le x|\boldsymbol{X}) - P_{H_0}(F_N \le x)| \xrightarrow{P} 0. \tag{3.11}$$

Remark 3.1. The corresponding wild bootstrap tests are given by $\varphi_{WTS}^\epsilon = \mathbb{1}\{Q_N > c_{WTS}^\epsilon\}$ and $\varphi_{ATS}^\epsilon = \mathbb{1}\{F_N > c_{ATS}^\epsilon\}$, where c_{WTS}^ϵ and c_{ATS}^ϵ denote the conditional $(1-\alpha)$-quantile of the wild bootstrap distribution of Q_N^ϵ and F_N^ϵ given the data, respectively. Properties (3.10) and (3.11) ensure that the wild bootstrap tests are of asymptotic level α under the null hypothesis and consistent for any fixed alternative. Moreover, it follows from Janssen and Pauls (2003) that they possess the same local power under contiguous alternatives as the original tests φ_{WTS} and φ_{ATS}, respectively.

4. Simulations

In the previous sections, nonparametric rank-based inference methods for the analysis of general factorial longitudinal data have been derived. The procedures are based on the asymptotic joint distribution of the vector $\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}$ under the hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$. As an approximate solution, wild bootstrap methods are proposed. All of the proposed approaches, however, are valid for large sample sizes. In order to investigate the accuracies of the procedures in terms of (i) controlling the pre-assigned type-1 error level under the null hypothesis, and (ii) their power to detect certain alternatives, extensive simulation studies were conducted. All simulations were performed with R environment, version 3.2.2. (R Core Team, 2010), each with 100,000 simulation and 999 bootstrap repetitions (Dufour and Khalaf, 2001; Racine and MacKinnon, 2007), respectively. Due to abundance of possible factorial longitudinal designs, we restrict the analysis to one-way designs with $a = 2$ independent groups of subjects, different numbers of time points $t \in \{4, 8\}$, underlying discrete and continuous data distributions (ordinal data, normal, and lognormal), and varying sample sizes $n_i \in \{10, 20\}$. Discrete data were simulated in order to investigate the impact of tied observations on the wild bootstrap tests. Both the WTS, ATS and their wild bootstrap versions are investigated to test the null hypothesis of "no main effect" (A), "no time effect" (T), as well as "no interaction" (A:T) between the main and time effect, respectively. The nominal type-1 error rate was set to 5%. More simulation results for different α levels ($\alpha = 1\%$ and $\alpha = 10\%$) can be found in the supplementary material (see Appendix B). The results and conclusions obtained are similar to the ones presented below. Throughout the simulations, random signs were used as weights for both the wild bootstrap methods. Results for standard normal, uniform or Mammen (1993a) weights lead to less accurate test decisions, and are therefore omitted.

4.1. Ordinal data

In order to imitate the underlying distributions of the grading scores given in the shoulder tip pain trial, a split-plot design with $a = 2$ groups, n_i subjects in group i and t repeated measures X_{iks} was simulated. The observations

$$X_{iks} = \frac{5(Z_{iks} + cY_{ik})}{c+1} + 1$$

were generated from independent observations $Y_{ik} \sim U[0, 1]$ and $Z_{iks} \sim U[0, 1]$, $i = 1, 2$, $k = 1, \ldots, n_i$ and $s = 1, \ldots, t$. The random variables X_{iks} take values between 1 and 5 as in the shoulder tip pain trial. The correlation between X_{iks} and $X_{iks'}$ can be regulated by choosing the constant c between 0 and ∞. Here, $c = 1$ has been chosen. Thus, the generated scores have a compound symmetric covariance structure. The type-1 error simulation results are displayed in Table 1.

It can be readily seen from Table 1 that the classical Wald-type test (WTS) tends to liberal conclusions. Roughly speaking, the liberality of the WTS can be explained by the non-consideration of the variability of the empirical covariance matrix by

Table 1
Simulation results ($\alpha = 5\%$) of the WTS, ATS and their wild bootstrap versions for testing the three different hypotheses (A, T, A:T) with ordinal data and varying sample sizes.

Hypothesis	$\boldsymbol{n}$	Method	$t = 4$		$t = 8$	
			ATS	WTS	ATS	WTS
A	$n_1 = n_2 = 10$	Classic	0.066	0.066	0.066	0.066
		Wild bootstrap	0.051	0.051	0.051	0.051
	$n_1 = 10, n_2 = 20$	Classic	0.067	0.067	0.066	0.066
		Wild bootstrap	0.054	0.054	0.053	0.053
T	$n_1 = n_2 = 10$	Classic	0.054	0.118	0.038	0.314
		Wild bootstrap	0.051	0.052	0.048	0.049
	$n_1 = 10, n_2 = 20$	Classic	0.053	0.111	0.039	0.277
		Wild bootstrap	0.051	0.055	0.049	0.059
A:T	$n_1 = n_2 = 10$	Classic	0.053	0.115	0.038	0.312
		Wild bootstrap	0.051	0.050	0.048	0.049
	$n_1 = 10, n_2 = 20$	Classic	0.055	0.113	0.038	0.274
		Wild bootstrap	0.052	0.055	0.049	0.058

its limiting χ^2 distribution. Its wild bootstrap version, however, greatly improves the type-1 error rate control of the WTS. This occurs, because the wild bootstrap distribution takes the variability of the empirical covariance matrix into account. Therefore, the actual sampling distribution of the WTS and its wild bootstrap version are similar when sample sizes are small. The same behavior can be seen for the ATS. The classical ATS is less liberal than the WTS, however, its empirical type-1 error rate is $\approx$7% when testing for the main effect. In the other situations (T and $A : T$), the method tends to be conservative in case of larger numbers of time points. Its wild bootstrap version, however, improves this behavior and tends to an accurate type-1 error rate control. Furthermore, it can be seen that unbalanced designs seem to not affect the accuracy of the wild bootstrap tests. Altogether, rejection rates for the wild bootstrap procedure vary between 0.048 and 0.059, with usually larger values for the WTS wild bootstrap test.

Next, continuous distributions and the impact of different covariance structures on the quality of the approximations will be investigated.

4.2. Continuous data

For the empirical investigation of the type-1 error rate control of the proposed methods, balanced and unbalanced split-plot design with $a = 2$ groups, sample sizes $n_i \in \{10, 20\}$, and $t = \{4, 8\}$ repeated measures X_{iks} was simulated. Data was generated by:

$$\boldsymbol{X}_{ik} = \boldsymbol{\Sigma}_i^{1/2} \widetilde{\boldsymbol{X}}_{ik},$$

where $\boldsymbol{\Sigma}_i$ either has an autoregressive structure (Setting 1) or a compound symmetric pattern (Setting 2):

Setting 1 (AR): $\boldsymbol{\Sigma}_i = \left(\rho^{|l-j|}\right)_{l,j \le t}$, $\rho = 0.6$ for $i = 1, 2$.
Setting 2 (CS): $\boldsymbol{\Sigma}_i = \boldsymbol{I}_t + \boldsymbol{J}_t$ for $i = 1, 2$.

The independent and identically distributed random vectors $\widetilde{\boldsymbol{X}}_{ik} = (\widetilde{X}_{ik1}, \ldots, \widetilde{X}_{ikt})$ were generated either from a standard normal distribution or from a standardized log-normal distribution.

The type-1 error simulation results for testing the hypotheses of *no main effect A*, *no time effect T* and *no main* $\times$ *time interaction* $A : T$ are displayed in Tables 2 and 3 for the normal and log-normal distribution, respectively.

It can be seen from Tables 2 and 3 that the shape of the underlying data distribution does not affect the type-1 error rate control of all four methods, and are similar for all three investigated distributions (ordinal, normal, and lognormal). Furthermore, the chosen dependency structures of the data do not impact the quality of the approximations. All of the investigated methods allow for an arbitrary covariance matrix. From Tables 2 and 3 it further follows that both the WTS and ATS show a liberal and conservative to slightly liberal behavior depending on the hypothesis and number of time points, respectively. Both the wild bootstrap methods show an accurate type-1 error rate control with rejection rates varying from 0.047 to 0.058 for the normal distribution and 0.044 to 0.068 for the log-normal distribution, and are therefore recommended for practical applications. Note, however, that the rejection frequencies of the wild bootstrap version of the WTS are sometimes statistically different from 5% for $t = 8$ time points and unequal sample sizes, e.g. 0.068 for testing T with compound symmetry in Table 3. Next, the power of the methods for the detection of certain alternatives will be investigated.

4.3. Power

To investigate the power of the tests a separate simulation study was performed in a one-sample repeated measures design utilizing multivariate normal distributions with expectation $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$, autoregressive covariance

Table 2
Type-1 error simulation results for normally distributed data with sample sizes $\boldsymbol{n}^{(1)} = (10, 10)$ and $\boldsymbol{n}^{(2)} = (10, 20)$.

	Cov. setting	$\boldsymbol{n}$	Method	$t = 4$		$t = 8$	
				ATS	WTS	ATS	WTS
A	AR	$\boldsymbol{n}^{(1)}$	Classic	0.064	0.066	0.066	0.067
			Wild bootstrap	0.050	0.050	0.052	0.052
		$\boldsymbol{n}^{(2)}$	Classic	0.063	0.065	0.067	0.068
			Wild bootstrap	0.052	0.052	0.054	0.054
	CS	$\boldsymbol{n}^{(1)}$	Classic	0.064	0.066	0.067	0.067
			Wild bootstrap	0.050	0.050	0.052	0.052
		$\boldsymbol{n}^{(2)}$	Classic	0.064	0.066	0.067	0.067
			Wild bootstrap	0.052	0.052	0.054	0.054
T	AR	$\boldsymbol{n}^{(1)}$	Classic	0.055	0.114	0.050	0.314
			Wild bootstrap	0.050	0.049	0.053	0.049
		$\boldsymbol{n}^{(2)}$	Classic	0.056	0.109	0.050	0.273
			Wild bootstrap	0.053	0.053	0.053	0.058
	CS	$\boldsymbol{n}^{(1)}$	Classic	0.052	0.115	0.037	0.314
			Wild bootstrap	0.049	0.050	0.047	0.048
		$\boldsymbol{n}^{(2)}$	Classic	0.052	0.108	0.037	0.270
			Wild bootstrap	0.051	0.053	0.048	0.056
A:T	AR	$\boldsymbol{n}^{(1)}$	Classic	0.055	0.115	0.049	0.315
			Wild bootstrap	0.051	0.050	0.052	0.049
		$\boldsymbol{n}^{(2)}$	Classic	0.057	0.109	0.049	0.273
			Wild bootstrap	0.053	0.053	0.053	0.058
	CS	$\boldsymbol{n}^{(1)}$	Classic	0.052	0.114	0.037	0.315
			Wild bootstrap	0.049	0.050	0.048	0.047
		$\boldsymbol{n}^{(2)}$	Classic	0.052	0.108	0.036	0.268
			Wild bootstrap	0.050	0.053	0.047	0.057

Table 3
Simulation results for log-normally distributed data with sample sizes $\boldsymbol{n}^{(1)} = (10, 10)$ and $\boldsymbol{n}^{(2)} = (10, 20)$.

	Cov. setting	$\boldsymbol{n}$	Method	$t = 4$		$t = 8$	
				ATS	WTS	ATS	WTS
A	AR	$\boldsymbol{n}^{(1)}$	Classic	0.065	0.066	0.066	0.066
			Wild bootstrap	0.051	0.051	0.052	0.052
		$\boldsymbol{n}^{(2)}$	Classic	0.064	0.065	0.067	0.068
			Wild bootstrap	0.052	0.052	0.055	0.055
	CS	$\boldsymbol{n}^{(1)}$	Classic	0.065	0.066	0.067	0.067
			Wild bootstrap	0.051	0.051	0.052	0.052
		$\boldsymbol{n}^{(2)}$	Classic	0.064	0.065	0.068	0.068
			Wild bootstrap	0.052	0.052	0.054	0.054
T	AR	$\boldsymbol{n}^{(1)}$	Classic	0.059	0.121	0.055	0.324
			Wild bootstrap	0.053	0.055	0.054	0.054
		$\boldsymbol{n}^{(2)}$	Classic	0.060	0.118	0.056	0.281
			Wild bootstrap	0.056	0.058	0.055	0.062
	CS	$\boldsymbol{n}^{(1)}$	Classic	0.051	0.122	0.034	0.334
			Wild bootstrap	0.048	0.056	0.044	0.059
		$\boldsymbol{n}^{(2)}$	Classic	0.051	0.116	0.035	0.283
			Wild bootstrap	0.049	0.059	0.046	0.068
A:T	AR	$\boldsymbol{n}^{(1)}$	Classic	0.057	0.116	0.054	0.316
			Wild bootstrap	0.051	0.050	0.053	0.048
		$\boldsymbol{n}^{(2)}$	Classic	0.058	0.111	0.054	0.275
			Wild bootstrap	0.054	0.055	0.054	0.059
	CS	$\boldsymbol{n}^{(1)}$	Classic	0.052	0.115	0.036	0.314
			Wild bootstrap	0.049	0.049	0.047	0.045
		$\boldsymbol{n}^{(2)}$	Classic	0.052	0.111	0.036	0.268
			Wild bootstrap	0.050	0.056	0.046	0.054

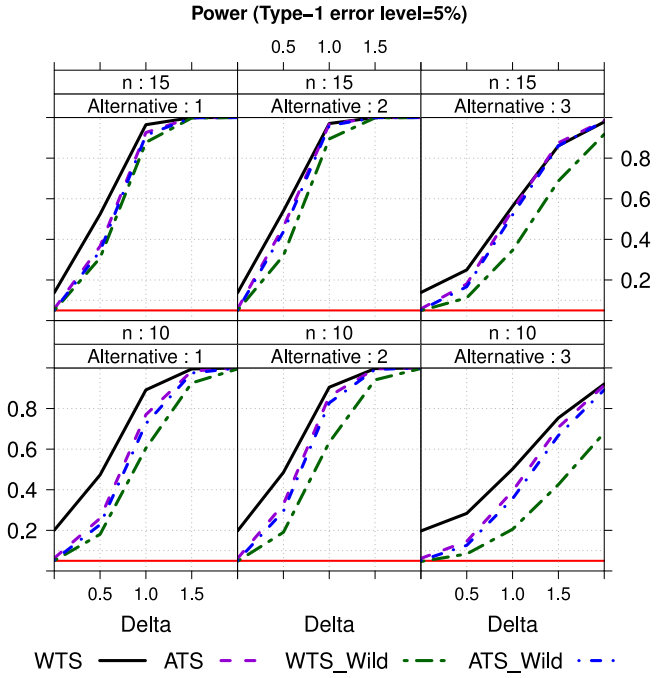

Fig. 2. Power simulation results (type-1 error level $\alpha = 5\%$) of the four investigated methods to detect the *Alternatives* 1–3 defined above with varying sample sizes $n = 10$ and $n = 15$, respectively.

structure $\boldsymbol{V}_{ij} = (0.6)^{|i-j|}$, $t = 4$ dimensions and sample size $n \in \{10, 15\}$. The aim of the simulation study is to investigate and compare the empirical power of the tests to detect the three chosen alternatives

Alternative 1: $\boldsymbol{\mu}^{(1)} = (0, 0, 0, \delta)$ for varying $\delta \in \{0, 0.5, 1, 1.5, 2\}$,
Alternative 2: $\boldsymbol{\mu}^{(2)} = (0, 0, \delta, \delta)$ for varying $\delta \in \{0, 0.5, 1, 1.5, 2\}$,
Alternative 3: $\boldsymbol{\mu}^{(3)} = \delta \cdot (1/4, 1/2, 3/4, 1)$ for varying $\delta \in \{0, 0.5, 1, 1.5, 2\}$.

They are chosen to represent frequently appearing alternatives in practical applications. The *Alternatives* 1–3 represent a 1-point, 2-point and a trend alternative, respectively. The power simulation results are displayed in Fig. 2.

Although the Wald-type statistic Q_N tends to be highly liberal when small sample sizes like $n = 10$ or 15 are present, the statistic has been included in Fig. 2 for illustration purposes. However, because of these issues, the method will not be viewed as a competitor to the other three methods. It can be seen from Fig. 2 that the ATS has the highest power to detect all three chosen alternatives when sample size is very small ($n = 10$). However, this method is slightly liberal when $n = 10$, and therefore the conclusion that the ATS is head and shoulders above the rest is questionable. In particular, with increasing sample sizes ($n = 15$) both the power of the ATS and its wild bootstrap version are similar while the latter keeps the prescribed α level more accurate. The wild bootstrap version of the WTS has the lowest power to detect all of the three alternatives. It has a slightly lower power than the wild bootstrap version of the ATS in the first two scenarios while a considerable power loss (compared to the ATS and its wild bootstrap version) to detect trend alternatives is apparent.

5. Application: analysis of the data example

We now re-analyze the data of the shoulder tip pain trial (Lumley, 1996). It turns out that the given scores for the treated male patients given under time point 5 and 6 are identical, thus, the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is singular. Therefore, the WTS cannot be used for data analysis, and only the ATS will be used for making inference. First, data will be descriptively analyzed. Since data was observed in an elaborate factorial design, the relative marginal effects are computed for each factor combination separately. The results for the joint analysis of all possible treatment × gender × time combinations along with 95% point-wise confidence intervals are displayed in Fig. 3, which was generated using the R-package **nparLD**.

It can be readily seen from Fig. 3 that the time responses between the treated and non-treated patients differ. This is most apparent at time point $t = 3$, where the confidence intervals between the treatment groups do not overlap. At all time points, the estimated effects are smaller under treatment than under control. Thus, the scores seemingly tend to be smaller under treatment. Over time, the effects of the treated male patients tend to decrease until time $t = 4$ before they slightly increase and stabilize at the end. For the non-treated male patients the time profile is contrary: The effects rise until $t = 3$ and decline thereafter. Compared to the male patients, the time profile of the female patients show a similar behavior in both groups with slightly larger effects at the beginning. The ATS as well as its wild bootstrap version can now be used to

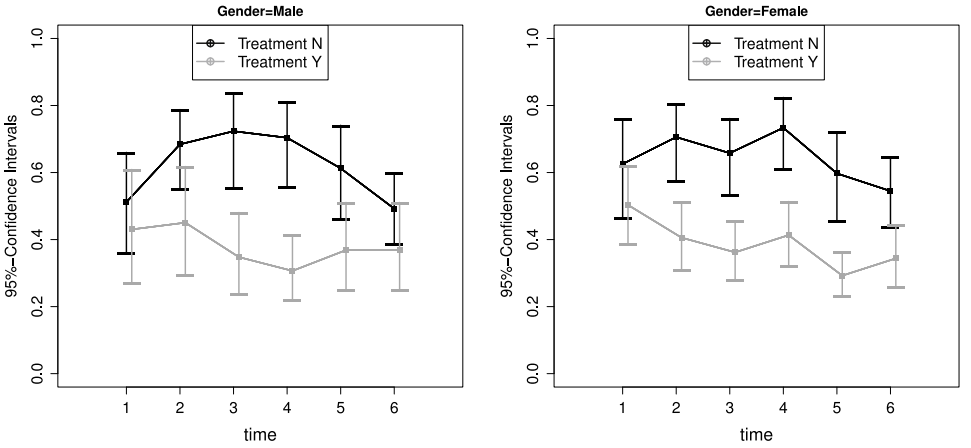

Fig. 3. Joint analysis of the data example: Treatment specific plots of the relative effects with 95%-confidence intervals—Gender: Male (left) and Female (right).

Table 4
Analysis of the shoulder tip pain trial using the ATS as given in (2.6) as well as the wild bootstrap ATS defined in (3.9).

Effect	Statistic	*df*	*p*-value (ATS)	*p*-value (WATS)
Treatment	16.401	1.000	<0.001	<0.001
Gender	0.046	1.000	0.832	0.827
Time	3.382	2.701	0.021	0.021
Treatment:gender	0.036	1.000	0.852	0.847
Treatment:time	3.711	2.701	0.014	0.013
Gender:time	1.144	2.701	0.327	0.325
Treatment:gender:time	0.438	2.701	0.705	0.736

Table 5
Treatment specific results for the shoulder tip pain trial using the ATS as given in (2.6) as well as the wild bootstrap ATS defined in (3.9).

Effect	Treatment = Yes				Treatment = No			
	Statistic	*df*	*p*-value (ATS)	*p*-value (WATS)	Statistic	*df*	*p*-value (ATS)	*p*-value (WATS)
Gender	0.007	1.000	0.932	0.931	0.046	1.000	0.834	0.828
Time	1.893	2.663	0.136	0.151	5.580	2.696	0.001	<0.001
Gender:time	0.959	2.663	0.403	0.432	0.926	2.696	0.419	0.424

test if significant main effects and interactions among the three factors treatment, gender and time are apparent. The results are presented in Tables 4 and 5. Here, the values of the test statistics, degrees of freedom of the classical F-approximation of the ATS and p-values for both the ATS and its wild bootstrap version (WATS) introduced in Section 3 are displayed. For the wild bootstrap 10,000 bootstraps were conducted.

It turns out that the ATS as well as its wild bootstrap version tend to result in similar conclusions. Overall, p-values obtained by the ATS, however, are slightly larger than those by the WATS (except for the threefold interaction). It turns out that the interaction between treatment and time is significant at 5% level of significance. Therefore, data is further split by the factor *treatment* and the above analysis is repeated for each treatment group separately. We note that this changes the estimates and confidence intervals since ranks are no longer calculated from the pooled sample but separately for both (independent) groups. The results are given in Table 5.

It can be seen from Table 5 that in both treatment groups data do not provide the evidence for a gender $\times$ time interaction. Similarly, a significant gender effect does not seem to exist at 5% level in both groups. However, under treatment, the scores do not change significantly over time (WATS p-value of 0.151), while a significant time effect is apparent under placebo. The corresponding relative effect estimators with 95%-confidence intervals are displayed in Fig. 4. The significant time effect under control can be readily seen from Fig. 4. For both the male and female patients, the pain score is significantly smaller at time point 6 compared to time point 3.

5.1. Sensitivity analysis

In the data example, the WTS cannot be used due to the singularity of the covariance matrix. In order to apply both ATS and WTS, we have dropped time point 6 from the following analysis, yielding a non-singular covariance matrix. The resulting p-values of this analysis are displayed in Tables 6 and 7. It can be seen that all methods still detect a significant effect of the

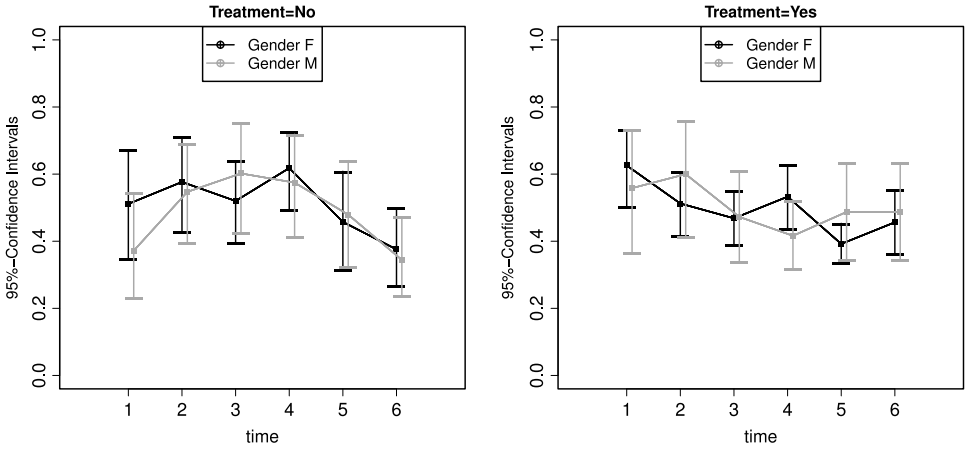

Fig. 4. Separate analysis of the data example per treatment: Plots of the relative effects with 95%-confidence intervals—Treatment: No (left) and Yes (right).

Table 6
Analysis of the shoulder tip pain trial (without time point 6) using the ATS, WTS as well as the wild bootstrap ATS and WTS.

Effect	*p*-value (ATS)	*p*-value (WATS)	*p*-value (WTS)	*p*-value (WWTS)
Treatment	<0.001	<0.001	<0.001	<0.001
Gender	0.8006	0.8033	0.8006	0.8033
Time	0.1356	0.1398	0.0344	0.0755
Treatment:gender	0.9151	0.9137	0.9151	0.9137
Treatment:time	0.0168	0.0189	0.0102	0.0338
Gender:time	0.2409	0.2419	0.0227	0.0591
Treatment:gender:time	0.6721	0.7028	0.3022	0.3867

Table 7
Treatment specific results for the shoulder tip pain trial using the ATS and the WTS as well as their wild bootstrap versions.

Effect	Treatment = Yes				Treatment = No			
	p-value (ATS)	*p*-value (WATS)	*p*-value (WTS)	*p*-value (WWTS)	*p*-value (ATS)	*p*-value (WATS)	*p*-value (WTS)	*p*-value (WWTS)
Gender	0.9314	0.9311	0.9314	0.9311	0.8306	0.8250	0.8306	0.8250
Time	0.1356	0.1413	0.0763	0.2400	0.0013	0.0006	<0.0001	0.0043
Gender:time	0.4032	0.4215	0.0237	0.1444	0.4193	0.4346	0.0520	0.2251

treatment as well as a significant interaction between treatment and time. The WTS furthermore detects a significant time effect as well as a significant interaction between gender and time. These findings are not supported by the other procedures (the WWTS finds borderline significance in both cases) and are probably due to the liberality of the WTS. To further analyze the results, we again consider the two treatment groups separately (see Table 7). Here, only the WTS detects a significant gender × time interaction in both treatment groups (borderline significant for the placebo group). All other procedures do not provide evidence for such an interaction. Furthermore, a significant gender effect does not seem to exist in both groups. However, a significant time effect seems to be present only in the placebo group, a finding shared by all four procedures again.

Overall, these findings are similar to the ones obtained above with the exception of the significant results only detected by the WTS, which are consistent with the liberal behavior of the WTS seen in the simulation studies in Section 4.

6. Conclusions and discussion

Ranking methods for the analysis of factorial longitudinal data provide a robust and powerful tool for making statistical inference. The considered Wald- and ANOVA-type statistic of Akritas and Brunner (1997) can be seen as the current state of the art. It turns out, however, that the Wald-type statistic tends to be quite liberal, while the ANOVA-type statistic tends to rather conservative or even liberal conclusions when small sample sizes are apparent. In this paper, a wild bootstrap method has been introduced. It was shown that the conditional distributions of the wild bootstrap statistics mimic the (asymptotic) distributions of the corresponding test statistics in both cases. Thus, the resampling versions are (at least) asymptotically valid, a desirable property that is not shared by the classical ATS. The empirical type-1 error rate control of the methods has been investigated for ordinal, symmetric as well as skewed continuous distributions with different covariance matrices in different balanced and unbalanced designs. The studies show that the wild bootstrap approximations of both the WTS and

ATS improve their finite sample behavior. Both resampling tests improve the type-1 error control of their non-bootstrap versions in all considered scenarios, whereof the wild bootstrap version of the ATS is more accurate in most instances. Regarding the power of the resampling tests, it could be seen that the wild bootstrap version of the ATS has higher power to detect the chosen alternatives when sample sizes are small ($n = 10, 15$). The power simulations have further shown, that the power of the ATS and its resampling version are asymptotically equivalent, i.e., both tests have the same power to detect certain alternatives when sample sizes are large. The findings via the simulation study give rise to recommend the wild bootstrap version of the ATS for practical applications. Different to both WTS procedures, this method is further applicable when the estimated variance covariance matrix is singular as in the presented data example.

The considered nonparametric hypotheses are formulated in terms of the distribution functions. The interpretation of the hypotheses can be challenging, particularly in factorial designs. The extension of the methods for testing hypotheses formulated in terms of the relative marginal effects by $H_0^p : \boldsymbol{Cp} = \boldsymbol{0}$ will be part of future research.

Acknowledgments

The authors would like to thank Edgar Brunner for providing the data example.
This work was supported by the German Research Foundation project DFG-PA 2409/3-1.

Appendix A

In a nonparametric setting as described in Section 2, the relative treatment effects are defined as

$$p_{is} = \int H_N dF_{is}, \quad i = 1, \ldots, a, \; s = 1, \ldots, t, \tag{A.12}$$

where again $H_N(x) = \frac{1}{tN} \sum_{i=1}^{a} \sum_{s=1}^{t} n_i F_{is}(x)$ denotes the weighted average of all marginal distribution functions.

Estimators of $H_N(x)$ and p_{is} are derived by replacing the distribution functions in (A.12) with the empirical distribution functions

$$\widehat{F}_{is}(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{2} \left[\mathbb{1}(x > X_{iks}) + \mathbb{1}(x \geq X_{iks}) \right], \tag{A.13}$$

resulting in $\widehat{H}(x) = \frac{1}{tN} \sum_{i=1}^{a} \sum_{s=1}^{t} n_i \widehat{F}_{is}(x)$ and a rank estimator of the relative effects

$$\widehat{p}_{is} = \int \widehat{H} d\widehat{F}_{is} = \frac{1}{n_i} \sum_{k=1}^{n_i} \widehat{H}(X_{iks}). \tag{A.14}$$

For each summand on the right hand side of (A.14) we write $\widehat{Y}_{iks} = \widehat{H}(X_{iks}) = \frac{R_{iks} - \frac{1}{2}}{tN}$ and set its limit variable to $Y_{iks} := H(X_{iks})$. It follows from the *Asymptotic Equivalence Theorem* (Akritas and Brunner, 1997) that under H_0^F, $\sqrt{N}\boldsymbol{C}\overline{\boldsymbol{Y}}_{\cdot}$ and $\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}$ asymptotically have the same distribution. Since $\overline{\boldsymbol{Y}}_{i\cdot}$ are means of independent random vectors and $\boldsymbol{Cp} = \boldsymbol{0}$ under H_0^F, it is easily established that $\sqrt{N}\boldsymbol{C}\overline{\boldsymbol{Y}}_{\cdot} \xrightarrow{d} N(0, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$ under the null hypothesis. Thus, we even have

$$\sqrt{N}\boldsymbol{C}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{d} N(0, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}') \tag{A.15}$$

under H_0^F. Here, $\boldsymbol{\Sigma} = \oplus_{i=1}^{a} \frac{1}{\kappa_i} \boldsymbol{V}_i$ and $\boldsymbol{V}_i = Cov(\boldsymbol{Y}_{i1})$. Note that the covariance matrices $\boldsymbol{V}_i$ may not be equal, even if a homoscedastic model is assumed for $\boldsymbol{X}$, since $H(\cdot)$ is a nonlinear transformation.

Proof of Theorem 2.1. The result is also stated in Section 1.5.1 of Brunner and Puri (2001). For completeness we shortly present its proof here: From (A.15) it follows that

$$\tilde{Q}_N = N\widehat{\boldsymbol{p}}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')^{+}\boldsymbol{C}\widehat{\boldsymbol{p}}$$

has asymptotically a central $\chi^2_{rank(\boldsymbol{C})}$ distribution under H_0^F. Finally, the result follows by replacing $\boldsymbol{\Sigma}$ with $\widehat{\boldsymbol{\Sigma}}$ by applying the multivariate Slutsky Theorem and noting that the involved Moore–Penrose inverse is continuous since $\boldsymbol{\Sigma} > 0$ by assumption.

Proof of Theorem 2.2. The proof can be found in Brunner and Puri (2001, THEOREM 1.8).

Proof of Theorem 3.1. Due to conditional independence of the random variables $\hat{\boldsymbol{p}}_i^{\epsilon}$, $i = 1, \ldots, a$, we can study them separately. Applying Theorem A.1 in Beyersmann et al. (2013), see also Theorem 4.1 in Pauly (2011), it remains to show the following convergences in probability

$$\max_{1 \leq i \leq a} \frac{\sqrt{N} \| \widehat{\boldsymbol{Y}}_{ik} - \widehat{\boldsymbol{Y}}_{i\cdot} \|}{n_i} \xrightarrow{P} 0, \quad N \to \infty, \tag{A.16}$$

as well as

$$\frac{N}{n_i^2}\sum_{k=1}^{n_i}(\widehat{\boldsymbol{Y}}_{ik}-\widehat{\boldsymbol{Y}}_{i\cdot})(\widehat{\boldsymbol{Y}}_{ik}-\widehat{\boldsymbol{Y}}_{i\cdot})' \xrightarrow{P} \frac{1}{\kappa_i}\boldsymbol{V}_i, \tag{A.17}$$

where $\widehat{\boldsymbol{Y}}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i}\widehat{\boldsymbol{Y}}_{ik}$. The first convergence (A.16) follows due to $|\widehat{Y}_{iks}| \leq 1$ and the second one (A.17) from

$$\begin{aligned}\frac{N}{n_i^2}\sum_{k=1}^{n_i}(\widehat{\boldsymbol{Y}}_{ik}-\widehat{\boldsymbol{Y}}_{i\cdot})(\widehat{\boldsymbol{Y}}_{ik}-\widehat{\boldsymbol{Y}}_{i\cdot})' &= \frac{(n_i-1)N}{n_i}\frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}(\widehat{\boldsymbol{Y}}_{ik}-\widehat{\boldsymbol{Y}}_{i\cdot})(\widehat{\boldsymbol{Y}}_{ik}-\widehat{\boldsymbol{Y}}_{i\cdot})' \\ &\xrightarrow{P} \frac{1}{\kappa_i}\boldsymbol{V}_i.\end{aligned}$$

Thus, we can conclude convergence in distribution

$$\frac{\sqrt{N}}{n_i}\sum_{k=1}^{n_i}{}_{ik}(\widehat{\boldsymbol{Y}}_{ik}-\overline{\boldsymbol{Y}}_{i\cdot}) \xrightarrow{d} N\left(0, \frac{1}{\kappa_i}\boldsymbol{V}_i\right) \qquad (i=1,\ldots,a)$$

given the data $\boldsymbol{X}$ in probability and the stated weak convergence of the conditional distribution of $\sqrt{N}\hat{\boldsymbol{p}}^{\epsilon}$ to an $N(0, \boldsymbol{\Sigma})$-distributed random variable as well as of $\sqrt{N}\boldsymbol{C}\hat{\boldsymbol{p}}^{\epsilon}$ to the right hand side of (A.15) in probability follows.

Proof of Theorem 3.2. The statement follows directly from Theorem 3.1 if we prove consistency of $\widehat{\boldsymbol{\Sigma}}$. Therefore, consider

$$\widehat{p}_{is}^{\epsilon} = \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{{}_{ik}}{Nt}(R_{iks}-\overline{R}_{i\cdot s}) =: \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{{}_{ik}}{Nt}Z_{iks}.$$

First, it holds that

$$E(\widehat{p}_{is}^{\epsilon}|\boldsymbol{X}) = E\left(\frac{1}{n_i}\sum_{k=1}^{n_i}\frac{{}_{ik}}{Nt}Z_{iks}|\boldsymbol{X}\right) = \frac{1}{Ntn_i}\sum_{k=1}^{n_i}E(\epsilon_{ik})\cdot E(Z_{iks}|\boldsymbol{X}) = 0.$$

Moreover, due to conditional independence of $\epsilon_{ik}Z_{iks}$ given $\boldsymbol{X}$, the corresponding conditional variances also converge to zero in probability as $n_i/N \to \kappa_i$:

$$\begin{aligned}\mathrm{Var}(\widehat{p}_{is}^{\epsilon}|\boldsymbol{X}) &= \frac{1}{(Ntn_i)^2}\sum_{k=1}^{n_i}\mathrm{Var}(\epsilon_{ik})Z_{iks}^2 \\ &= \frac{1}{(Ntn_i)^2}\sum_{k=1}^{n_i}Z_{iks}^2 \leq \frac{1}{(Ntn_i)^2}n_i(N-1)^2 \to 0.\end{aligned}$$

Because of Tschebyscheff's inequality this implies $\widehat{p}_{is}^{\epsilon} \xrightarrow{P} 0$ for all $i = 1,\ldots,a$ and thus the asymptotic equivalence of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}$. Since $\widehat{\boldsymbol{\Sigma}}$ is consistent, this completes the proof.

Proof of Theorem 3.3. The result follows from Theorems 2.2 and 3.1 and an application of Lemma 1 in Janssen and Pauls (2003) by noting that the limit distribution of A_N in (2.7) is continuous.

Remark. Note that the relative effects depend on the sample sizes n_i. To avoid this dependence one may replace the function $H(x)$ by the unweighted mean of the distribution functions $G(x) = \frac{1}{at}\sum_{i=1}^{a}\sum_{s=1}^{t}F_{is}(x)$. This results in unweighted relative effects $q_{is} = \int G dF_{is}$, see e.g. Puri and Hall (2003). A wild bootstrap version thereof may be defined analogously to the relative effects considered above and the asymptotic results follow analogous to $\hat{\boldsymbol{p}}^{\epsilon}$, if we consider $\widehat{Z}_{iks} = \widehat{G}(X_{iks})$ instead of $\widehat{Y}_{iks}$, i.e. let $\widehat{\boldsymbol{q}}^{\epsilon} = \frac{1}{n_i}\sum_{k=1}^{n_i}{}_{ik}(\widehat{\boldsymbol{Z}}_{ik}-\widehat{\boldsymbol{Z}}_{i\cdot})$ for $\widehat{\boldsymbol{Z}}_{i\cdot} = n_i^{-1}\sum_{k=1}^{n_i}\widehat{\boldsymbol{Z}}_{ik}$. Given the data $\boldsymbol{X}$, we have conditional convergence in distribution

$$\sqrt{N}\widehat{\boldsymbol{q}}^{\epsilon} \xrightarrow{d} N(0, \widetilde{\boldsymbol{\Sigma}})$$

in probability, where $\widetilde{\boldsymbol{\Sigma}} = \oplus\frac{1}{\kappa_i}\widetilde{\boldsymbol{V}}_i$ and $\widetilde{\boldsymbol{V}}_i = Cov(G(X_{iks}))$, analogous to the proof of Theorem 3.1.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.06.016.

References

Akritas, M.G., 2011. Nonparametric models for anova and ancova designs. In: International Encyclopedia of Statistical Science. Springer, pp. 964–968.
Akritas, M.G., Arnold, S.F., 1994. Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. J. Amer. Statist. Assoc. 89 (425), 336–343.
Akritas, M.G., Brunner, E., 1997. A unified approach to rank tests for mixed models. J. Statist. Plann. Inference 61 (2), 249–277.
Arnau, J., Bono, R., Blanca, M.J., Bendayan, R., 2012. Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. Behav. Res. Methods 44 (4), 1224–1238.
Beyersmann, J., Termini, S.D., Pauly, M., 2013. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. Scand. J. Statist. 40 (3), 387–402.
Box, G.E., 1954. 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 (2), 290–302.
Brombin, C., Midena, E., Salmaso, L., 2013. Robust non-parametric tests for complex-repeated measures problems in ophthalmology. Stat. Methods Med. Res. 22 (6), 643–660.
Brunner, E., 2001. Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity. In: Mathematical Statistics with Applications in Biometry.
Brunner, E., Dette, H., Munk, A., 1997. Box-type approximations in nonparametric factorial designs. J. Amer. Statist. Assoc. 92 (440), 1494–1502.
Brunner, E., Domhof, S., Langer, F., 2002. Nonparametric Analysis of Longitudinal Data in Factorial Experiments. Wiley, New York, USA.
Brunner, E., Langer, F., 2000. Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. Biom. J. 42 (6), 663–675.
Brunner, E., Munzel, U., Puri, M.L., 1999. Rank-score tests in factorial designs with repeated measures. J. Multivariate Anal. 70 (2), 286–317.
Brunner, E., Puri, M.L., 2001. Nonparametric methods in factorial designs. Statist. Papers 42 (1), 1–52.
Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. Rev. Econ. Stat. 90 (3), 414–427.
Cameron, A.C., Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. J. Hum. Resour. 50 (2), 317–372.
Chi, Y.-Y., Gribbin, M., Lamers, Y., Gregory, J.F., Muller, K.E., 2012. Global hypothesis testing for high-dimensional repeated measures outcomes. Stat. Med. 31 (8), 724–742.
Davidson, R., Flachaire, E., 2008. The wild bootstrap, tamed at last. J. Econometrics 146 (1), 162–169.
Davis, C.S., 2002. Statistical Methods for the Analysis of Repeated Measurements. Springer Science & Business Media.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application, Vol. 1. Cambridge University Press.
Davison, A.C., Hinkley, D.V., Young, G.A., 2003. Recent developments in bootstrap methodology. Statist. Sci. 141–157.
Dobler, D., Beyersmann, J., Pauly, M., 2015. Non-strange weird resampling for complex survival data, arXiv preprint arXiv:1507.02838.
Dobler, D., Pauly, M., 2014. Bootstrapping Aalen-Johansen processes for competing risks: Handicaps, solutions, and limitations. Electron. J. Stat. 8 (2), 2779–2803.
Dufour, J.-M., Khalaf, L., 2001. Monte Carlo test methods in econometrics. In: Companion to Theoretical Econometrics. In: Blackwell Companions to Contemporary Economics, Basil Blackwell, Oxford, UK, pp. 494–519.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 1–26.
Flachaire, E., 2005. Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. Comput. Statist. Data Anal. 49 (2), 361–376.
Friedrich, S., Brunner, E., Pauly, M., 2015. Permuting longitudinal data despite all the dependencies, arXiv preprint arXiv:1509.05570.
Good, P.I., 2006. Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer Science & Business Media.
Hedeker, D., Gibbons, R.D., 2006. Longitudinal Data Analysis, Vol. 451. John Wiley & Sons.
Howell, D., 2013. Fundamental Statistics for the Behavioral Sciences. Cengage Learning.
Huynh, H., 1978. Some approximate tests for repeated measurement designs. Psychometrika 43 (2), 161–175.
Huynh, H., Feldt, L.S., 1976. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. J. Educ. Behav. Stat. 1 (1), 69–82.
Janssen, A., 1999. Nonparametric symmetry tests for statistical functionals. Math. Methods Statist. 8 (3), 320–343.
Janssen, A., Pauls, T., 2003. How do bootstrap and permutation tests work? Ann. Statist. 768–806.
Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis. Pearson.
Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 983–997.
Kenward, M.G., Roger, J.H., 2009. An improved approximation to the precision of fixed effects from restricted maximum likelihood. Comput. Statist. Data Anal. 53 (7), 2583–2595.
Keselman, H., Algina, J., Kowalchuk, R.K., 2001. The analysis of repeated measures designs: a review. British J. Math. Statist. Psych. 54 (1), 1–20.
Keselman, H., Kowalchuk, R.K., Algina, J., Lix, L.M., Wilcox, R.R., 2000. Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. British J. Math. Statist. Psych. 53 (2), 175–191.
Kherad-Pajouh, S., Renaud, O., 2015. A general permutation approach for analyzing repeated measures anova and mixed-model designs. Statist. Papers 56, 947–967.
Konietschke, F., Bathke, A., Harrar, S., Pauly, M., 2015. Parametric and nonparametric bootstrap methods for general manova. J. Multivariate Anal. 140, 291–301.
Konietschke, F., Pauly, M., 2012. A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. Electron. J. Stat. 6, 1358–1372.
Kreiss, J.-P., Paparoditis, E., 2011. Bootstrap methods for dependent data: A review. J. Korean Stat. Soc. 40 (4), 357–378.
Lecoutre, B., 1991. A correction for the ε approximate test in repeated measures designs with two or more independent groups. J. Educ. Behav. Stat. 16 (4), 371–372.
Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses. In: Springer Texts in Statistics.
Lin, D., 1997. Non-parametric inference for cumulative incidence functions in competing risks studies. Stat. Med. 16 (8), 901–910.
Liu, R.Y., 1988. Bootstrap procedures under some non-iid models. Ann. Statist. 16 (4), 1696–1708.
Lumley, T., 1996. Generalized estimating equations for ordinal data: a note on working correlation structures. Biometrics 354–361.
Mammen, E., 1993a. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist. 255–285.
Mammen, E., 1993b. When Does Bootstrap Work? Asymptotic Results and Simulations. Springer Science & Business Media.
Manly, B.F., 2006. Randomization, Bootstrap and Monte Carlo Methods in Biology, Vol. 70. CRC Press.
Martinussen, T., Scheike, T.H., 2007. Dynamic Regression Models for Survival Data. Springer Science & Business Media.
Noguchi, K., Gel, Y.R., Brunner, E., Konietschke, F., 2012. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. J. Stat. Softw. 50 (12), 1–23.
Oberfeld, D., Franke, T., 2013. Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. Behav. Res. Methods 45 (3), 792–812.
Pauly, M., 2011. Weighted resampling of martingale difference arrays with applications. Electron. J. Stat. 5, 41–52.
Pauly, M., Ellenberger, D., Brunner, E., 2015. Analysis of high-dimensional one group repeated measures designs. Statistics 49, 1243–1261.
Pesarin, F., 2001. Multivariate Permutation Tests: With Applications in Biostatistics, Vol. 240. Wiley, Chichester.
Pesarin, F., Salmaso, L., 2012. A review and some new results on permutation testing for multivariate problems. Stat. Comput. 22 (2), 639–646.
Puri, M., Hall, P., 2003. Nonparametric methods in statistics and related topics. In: Puri, Madan Lal, Hall, Peter G., Hallin, Marc, Roussas, George G. (Eds.), Selected Collected Works. De Gruyter.

R Core Team (2010). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. http://www.R-project.org/.
Racine, J.S., MacKinnon, J.G., 2007. Simulation-based tests that can use any number of simulations. Comm. Statist. Simulation Comput. 36 (2), 357–365.
SAS Institute Inc. 2003. SAS Software, Version 9.1. Cary, NC.
Stevens, J.P., 2012. Applied Multivariate Statistics for the Social Sciences. Routledge.
Suo, C., Toulopoulou, T., Bramon, E., Walshe, M., Picchioni, M., Murray, R., Ott, J., 2013. Analysis of multiple phenotypes in genome-wide genetic mapping studies. BMC Bioinform. 14 (1), 151.
Vallejo, G., Ato, M., 2006. Modified brown–forsythe procedure for testing interaction effects in split-plot designs. Multivariate Behav. Res. 41 (4), 549–578.
Wu, C.-F.J., 1986. Jackknife, bootstrap and other resampling methods in regression analysis. Ann. Statist. 1261–1295.
Xu, J., Cui, X., 2008. Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. Bioinformatics 24 (8), 1056–1062.
Zapf, A., Brunner, E., Konietschke, F., 2015. A wild bootstrap approach for the selection of biomarkers in early diagnostic trials. BMC Med. Res. Methodol. 15 (1), 43.