

Permutation- and resampling-based inference for semi- and nonparametric effects in dependent data

Sarah Friedrich

Angaben zur Veröffentlichung / Publication details:

Friedrich, Sarah. 2017. Permutation- and resampling-based inference for semi- and nonparametric effects in dependent data. Augsburg: Universität Augsburg.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under the following conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publizieren>





ulm university universität
uulm

Permutation- and resampling-based inference for semi- and nonparametric effects in dependent data

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät für Mathematik und
Wirtschaftswissenschaften der Universität Ulm

Vorgelegt von

Sarah Jasmin Friedrich

aus Filderstadt

Ulm, Juli 2017

Amtierender Dekan:

Prof. Dr. Alexander Lindner

Gutachter:

Prof. Dr. Markus Pauly (Universität Ulm)

Prof. Dr. Werner Brannath (Universität Bremen)

Tag der Promotion:

01.12.2017

Abstract

In this thesis, we consider different resampling approaches for testing general linear hypotheses with dependent data. We distinguish between a repeated measures model, where subjects are repeatedly observed over time, and multivariate data, where outcomes may be measured on different scales. Furthermore, we consider semi-parametric approaches for metric data, where we test null hypotheses formulated in terms of means, as well as nonparametric rank-based models for ordinal data. In these settings, current state-of-the-art test statistics include the Wald-type statistic (WTS), which is asymptotically χ^2 -distributed, and the ANOVA-type statistic (ATS), which is no asymptotic pivot, but can be approximated by an F -distribution. To improve the small sample behavior of these test statistics in the described settings, we consider different resampling schemes. In the case of semi-parametric repeated measurements, a permutation procedure based on the WTS leads to astonishingly successful results in spite of the dependencies. In the nonparametric repeated measures design, a wild bootstrap procedure applied to the ATS yields the best simulation results and provides an asymptotic level α test. Finally, in the semi-parametric multivariate setting, we consider a modified ATS, which is invariant under scale transformations and can be applied to data with singular covariance matrices. Statistical inference (test decisions and the derivation of multivariate confidence regions) is based on quantiles of a parametric, nonparametric or wild bootstrap procedure.

We apply all resampling approaches to data examples from the life sciences. Furthermore, we analyze the small sample behavior of the tests in large simulation studies.

Zusammenfassung

Diese Arbeit beschäftigt sich mit Resampling-Verfahren für Hypothesentests bei abhängigen Daten. Wir unterscheiden zwischen repeated measurements, wobei die Individuen zu verschiedenen Zeitpunkten beobachtet werden, und multivariaten Daten, die auf unterschiedlichen Skalen gemessen sein können. Desweiteren betrachten wir ein semi-parametrisches Modell für metrische Daten, in welchem Hypothesen über Mittelwerte getestet werden, sowie ein nicht-parametrisches, rang-basiertes Modell für ordinale Daten. In der Literatur werden in solchen Situationen im Wesentlichen zwei Teststatistiken betrachtet: Eine Statistik vom Wald-Typ (WTS), die asymptotisch χ^2 -verteilt ist, sowie eine ANOVA-Typ Statistik (ATS), die durch eine F -Verteilung approximiert werden kann. Um das Verhalten dieser Statistiken für kleine Fallzahlen zu verbessern, werden verschiedene Resampling-Verfahren untersucht. Im Falle der semi-parametrischen repeated measures Daten führt ein auf der WTS basierender Permutationsansatz zu erstaunlich erfolgreichen Resultaten trotz der involvierten Zerstörung der Abhängigkeitsstrukturen. Im nicht-parametrischen repeated measures Modell liefert ein wild bootstrap der ATS die besten Simulationsergebnisse und einen asymptotischen Test zum Niveau α . Im multivariaten semi-parametrischen Modell schließlich betrachten wir eine Modifizierung der ATS, die invariant unter Einheitenwechsel ist und auch für Designs mit singulären Kovarianzmatrizen verwendet werden kann. Statistische Inferenz (Tests und die Herleitung multivariater Konfidenzregionen) basiert dann auf Quantilen eines parametrischen, nichtparametrischen oder wild bootstrap Ansatzes. Mit Hilfe der Verfahren werden verschiedene Datensätze aus den Lebenswissenschaften analysiert. Das Verhalten der Verfahren für kleine Fallzahlen wird darüber hinaus in umfangreichen Simulationsstudien untersucht.

Acknowledgments

First of all, I would like to thank my supervisor Markus Pauly for his support throughout the years of my dissertation, for the abundance of research topics he provided me with and for the opportunity to visit lots of interesting conferences and workshops! In this context I also acknowledge the financial support by the DFG. Furthermore, I would like to thank Werner Brannath for agreeing to be my second supervisor.

A special thank you goes to Jan Beyersmann for his encouragement to take up the position as doctoral student although he himself had no position to offer me and for his continued support of my academic career. Due to his great lectures on survival analysis I realized that biostatistical research is exactly what I always wanted to do.

Furthermore, I like to thank my co-authors Edgar Brunner and Frank Konietschke for helpful discussions that considerably improved this work. A special thank you goes to Frank for having me stay in Dallas, for teaching me how to write an R package (I am still sure I couldn't have done it without you) and for showing me everything Dallas and its surroundings have to offer.

Being a doctoral student would not have been half the fun if it wasn't for my colleagues at the Institute of Statistics. Thank you for all our discussions (be it topic-related or not), for the daily excursions to the cafeteria in the morning and the coffee breaks in the afternoon (although only Tobi ever drank coffee), for our Institutsstammtisch at the Barfüßer and for all the fun we had on various conferences. And a special thank you to Mia for proof-reading this work!

Finally, none of this would have been possible without the never-ending love and support of my family. Thank you for being there for me always! In particular, I want to thank my sister Lisa for her advice on English syntax and punctuation.

Last but not least, I would like to thank Nala for waiting patiently while I am away at work and for her warm smile and wagging tail that light up my every evening.

Contents

List of Publications	1
I Introduction	7
1 Motivation	9
2 Statistical Models	13
2.1 Semi-parametric repeated measures model	13
2.2 Nonparametric repeated measures model	15
2.3 Semi-parametric MANOVA model	18
3 Resampling procedures	21
3.1 Permutation procedure	22
3.2 Wild Bootstrap	23
3.3 Parametric Bootstrap	24
4 R packages and the EEG data example	27
4.1 The GFD package	27
4.2 MANOVA.RM and the EEG data example	28
5 Summary of the articles	33
5.1 Permuting longitudinal data in spite of the dependencies	33
5.2 A wild bootstrap approach for nonparametric repeated measurements . .	35
5.3 MATS : Inference for potentially singular and heteroscedastic MANOVA	37
6 Outlook: The fourth scenario	41
7 Discussion	45
Bibliography	47

II Publications	53
Appendix: The GFD package	157

List of Publications

This thesis is based on the following publications:

Article 1: Friedrich, S., Brunner, E. and Pauly, M. (2017). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, **153**, 255–265, DOI: 10.1016/j.jmva.2016.10.004.

Contribution of the author:

The author of this thesis implemented the extensive simulation studies and conducted the mathematical proofs under Prof. Pauly's guidance. Furthermore, she analyzed the data example under Prof. Brunner's advice and supervision.

Reprinted from Journal of Multivariate Analysis, Vol. 153, S. Friedrich, E. Brunner and M. Pauly, Permuting longitudinal data in spite of the dependencies, pages 255–265, Copyright (2017), with permission from Elsevier.

Article 2: Friedrich, S., Konietschke, F. and Pauly, M. (2017). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, **113**, 38–52, DOI: 10.1016/j.csda.2016.06.016.

Contribution of the author:

The author of this thesis had a leading role in the preparation and structuring of the manuscript. She mainly implemented the simulation studies and conducted the mathematical proofs as well as the analysis of the data example.

Reprinted from Computational Statistics and Data Analysis, Vol. 113, S. Friedrich, F. Konietschke and M. Pauly, A wild bootstrap approach for nonparametric repeated measurements, pages 38–52, Copyright (2017), with permission from Elsevier.

Article 3: Friedrich, S. and Pauly, M. (2017). MATS: Inference for potentially singular and heteroscedastic MANOVA. *arXiv preprint arXiv:1704.03731 (Revision submitted to Journal of Multivariate Analysis)*.

Contribution of the author:

The author of this thesis prepared and structured the manuscript mainly on her own. She conducted the simulation studies and the main part of the mathematical proofs. Furthermore, she chose and analyzed the data example independently.

Further publications:

Friedrich, S., Konietschke, F. and Pauly, M. (2017). GFD: An R package for the Analysis of General Factorial Designs. *Journal of Statistical Software, Code Snippets*, **79**(1), 1–18.

Bathke, A., Friedrich, S., Konietschke, F., Pauly, M., Staffen, W., Strobl, N. and Höller, Y. (2016). Using EEG, SPECT, and Multivariate Resampling Methods to Differentiate Alzheimer Patients from Others. *Revision submitted to Multivariate Behavioral Research, arXiv preprint arXiv:1606.09004*.

Dobler, D., Friedrich, S. and Pauly, M. (2017). Nonparametric MANOVA in Mann-Whitney effects. *Submitted to Journal of the American Statistical Association*

Friedrich, S, Konietschke, F, and Pauly, M (2017). Analysis of Multivariate Data and Repeated Measures Designs with the R Package MANOVA.RM. *Submitted to Computational Statistics and Data Analysis*.

R packages:

GFD: Tests for General Factorial Designs, R package version 0.2.4.

MANOVA.RM: Analysis of Multivariate Data and Repeated Measures Designs, R package version 0.2.1.

rankFD: Rank-Based Tests for General Factorial Designs, R package version 0.0.1.

rankMANOVA: Rank-Based Tests for Multivariate Data, *in preparation*.

Notation

Throughout the thesis, vectors and matrices are denoted by bold symbols, e.g., \mathbf{A} .

\mathbb{N}	natural numbers
$\mathbb{1}\{\cdot\}$	indicator function
\mathbf{A}'	the transpose of a matrix or (column) vector \mathbf{A}
\mathbf{A}^+	Moore-Penrose inverse ¹ of a matrix \mathbf{A}
\mathbf{I}_t	$t \times t$ identity matrix, $t \in \mathbb{N}$
$\mathbf{1}_t$	t -dimensional column vector of 1's, $t \in \mathbb{N}$
\mathbf{J}_t	$t \times t$ matrix of 1's, i.e., $\mathbf{J}_t = \mathbf{1}_t \mathbf{1}'_t$ for $t \in \mathbb{N}$
\mathbf{P}_t	t -dimensional centering matrix, $\mathbf{P}_t = \mathbf{I}_t - \frac{1}{t} \mathbf{J}_t$ for $t \in \mathbb{N}$
\otimes	Kronecker product
\oplus	direct sum
$\text{tr}()$	the trace of a square matrix
$\text{rank}()$	the rank of a matrix
\xrightarrow{P}	convergence in probability
\xrightarrow{D}	convergence in distribution

¹The Moore-Penrose inverse satisfies the following equations: $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$, $(\mathbf{A}\mathbf{A}^+)' = \mathbf{A}\mathbf{A}^+$ and $(\mathbf{A}^+\mathbf{A})' = \mathbf{A}^+\mathbf{A}$.

Part I

Introduction

1 Motivation

Factorial designs are widely used tools for modeling statistical experiments in many disciplines, for example in the life sciences. In the univariate case, inference is traditionally based on mean vectors and effects of the factors are tested using ANOVA F -tests. These tests, however, are derived under the assumptions of normally distributed errors and common variances between groups - two assumptions that are difficult to check and often not met in practice. If the assumptions are violated, type-I errors may be inflated, see e.g. the comments in Pauly et al. (2015).

To complicate matters, data are often not independent due to multivariate endpoints, which may be measured on different scales. Classical MANOVA methods and their extensions that still rely on the assumption of multivariate normality and covariance homogeneity are of limited use in practice, see e.g. the comments in Xu and Cui (2008); Suo et al. (2013); Konietschke et al. (2015).

Another setting involving dependent data are repeated measures designs. Here, the same outcome is measured at different occasions, e.g., different time points, or at different parts of the subject (e.g., left and right hemisphere of the brain). The repeated measures are considered as levels of the sub-plot or within-subjects factor, whereas the observed groups are levels of the whole-plot or between-subjects factor. In the context of repeated measures or profile analysis, it is of interest to investigate whether a group effect, a non-constant time effect or different time profiles in the groups are present. Classical tests based on Hotelling's T^2 (Hotelling, 1931) or Wilk's Λ (Wilks, 1932) assume normally distributed errors and covariance homogeneity. Violation of these assumptions may again inflate type-I errors.

The main difference between these two approaches is that in a repeated measures design, comparisons between the response variables are meaningful. Therefore, it is of interest to formulate and test hypotheses about the sub-plot factors, e.g., time. Multivariate data, in contrast, consist of several endpoints, which are recorded per subject (or unit) and may be measured on different scales. In a multivariate setting, we test

whether the observed factors have an effect on the multivariate outcome vectors.

A data example demonstrating the similarities and differences between the two approaches is presented below and analyzed in detail in Chapter 4. Nevertheless, both settings lead to dependencies between observations from the same unit, thus complicating the statistical analysis.

In addition to the structure of the data, it is also important to choose adequate methods depending on the scaling of the measurements. Classical MANOVA and repeated measures approaches usually consider differences in terms of the mean vector. If ordinal, ordered categorical or count data are present, however, means are neither adequate nor meaningful measures of deviations between groups. We therefore consider two different approaches based on the scaling of the data: A semi-parametric model in case of metric data used in Friedrich et al. (2017a) and Friedrich and Pauly (2017), where effects are formulated in terms of means, and a nonparametric repeated measures model, where we base inference on rank-statistics of the relative effects (Friedrich et al., 2017d). The latter is valid for metric, ordinal, count, score or ordered categorical data in a unified way.

To motivate the procedures analyzed in this thesis, we will first provide some data examples from different fields of application:

EEG measurements in patients with Alzheimer’s disease

To illustrate the difference between a repeated measures setting and multivariate outcomes, we consider a study on EEG measurements in patients with Alzheimer’s disease (Bathke et al., 2016). This data set contains EEG measurements of 166 patients recorded at the University Hospital Salzburg. Beforehand, the patients were diagnosed with either Alzheimer’s disease (AD), mild cognitive impairment (MCI) or subjective cognitive complaints without clinically significant deficits (SCC) based on neurological examinations. For each patient, we consider six EEG measurements, which consist of z -scores for brain rate (Pop-Jordanova and Pop-Jordanova, 2005) and Hjorth complexity (Hjorth, 1970, 1975). Each of these measurements is averaged within frontal, temporal and central electrode positions. The data can be viewed in two different ways: First, we may consider the data as multivariate. Thus, each individual has a 6-dimensional response vector. We can then formulate and test the hypothesis of no difference between diagnoses. On the other hand, the EEG measurements might also be viewed as levels of the sub-plot factors ‘brain region’ (3 levels: frontal, temporal, central) and ‘feature’ (2 levels: brain rate, complexity). Therefore, we can formulate and test hypotheses in-

volving the sub-plot factors. In Section 4, we provide an analysis of this data example using the R package **MANOVA.RM** (Friedrich et al., 2017c).

Shoulder tip pain trial

In the shoulder tip pain trial (Lumley, 1996), the characteristic pain in the shoulder after laparoscopic surgery was observed in 41 patients at 6 time points. In this trial, $n_1 = 22$ (8 male and 14 female) patients received the active treatment, while $n_2 = 19$ (8 male and 11 female) patients belonged to the control group. Therefore, data was observed in a factorial design with whole-plot factors ‘treatment’ and ‘gender’ and sub-plot factor ‘time’. The pain was measured on an ordinal scale ranging from 1 (low) to 5 (high). A first analysis of the data reveals that pain scores under treatment seem to be lower than those observed for the control group (see Figure 1 in Friedrich et al., 2017d). In order to investigate effects and interactions of the factors involved in this trial, we need rank-based repeated measures models, since means provide no adequate measure for score data. Furthermore, the methods used have to be robust to account for the small sample size in the experiment. We analyze this data example in detail in Friedrich et al. (2017d).

Effect of gender on cardiologic measurements

In a cardiology study conducted at Ulm University Hospital, five cardiologic measurements were recorded in the left ventricle of 188 healthy patients. We wish to analyze whether these measurements (in terms of mean vectors) differ between male and female patients. Since the outcomes are measured on different scales (systolic and diastolic peak strain rate is measured in 1/s, whereas end systolic and diastolic volume as well as stroke volume is measured in ml), we are dealing with a multivariate outcome. Furthermore, stroke volume closely depends on end diastolic and end systolic volume, resulting in singular empirical covariance matrices. These issues complicate the statistical analysis, since we have to deal with unequal, singular covariance matrices in multivariate, i.e., dependent, data. We will therefore modify an existing test statistic in order to deal with all these issues (Friedrich and Pauly, 2017).

This thesis is organized as follows: In Chapter 2 and 3 we will briefly introduce the different models and resampling procedures considered in this thesis, respectively. Chapter 4 contains a description of the R packages **GFD** and **MANOVA.RM** as well as a

detailed analysis of the EEG data example using the latter. Summaries of the three articles this thesis is based on are given in Chapter 5. The fourth scenario in this context, a nonparametric multivariate setting which is still work in progress, is described in Chapter 6, while Chapter 7 contains an overall discussion of the results as well as some outlook to future research. The three articles are included in Part II, while the Appendix contains the JSS publication describing the R package **GFD**.

2 Statistical Models

2.1 Semi-parametric repeated measures model

We consider the following statistical model: Let

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

denote i.i.d. random vectors of individual k in treatment group i , where the outcome is observed at time points $1, \dots, t_i \in \mathbb{N}$, $i = 1, \dots, a$. We extend the classical repeated measures design in that we allow the number of time points to vary between groups. A setting like this might be possible in, e.g., psychology where questionnaires with different lengths are used in different groups. The classical setting with $t_i \equiv t$ is, however, included as well.

We assume existence of the group-specific expectation vector $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{it_i})' = E(\mathbf{Y}_{i1})$ and the covariance matrix $\mathbf{V}_i = \text{Cov}(\mathbf{Y}_{i1}) > 0$. Covariance matrices may differ between groups and we neither assume any special covariance structure nor any special underlying distribution of \mathbf{Y}_{ik} . For convenience, we consider the pooled vector of observations

$$\mathbf{Y} = (\mathbf{Y}'_{11}, \dots, \mathbf{Y}'_{an_a})'$$

as well as the pooled mean vector

$$\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \dots, \boldsymbol{\mu}'_a)'$$

We can easily incorporate a more complex factorial structure in this notation by splitting up the index i into sub-indices i_1, i_2, \dots according to the number of factors considered.

Let $N = \sum_{i=1}^a n_i$ denote the total number of individuals, $T = \sum_{i=1}^a t_i$ the number of time points as well as $\tilde{N} = \sum_{i=1}^a n_i t_i$ the total number of observations. In order to

derive asymptotic results, we will assume the following sample size condition:

$$\frac{n_i}{N} \rightarrow \kappa_i > 0, \quad i = 1, \dots, a \quad (2.1)$$

as $\min_i n_i \rightarrow \infty$. Furthermore, we will assume existence of fourth moments, i.e., $\sup_i E(\|\mathbf{Y}_{i1}\|^4) < \infty$.

In this set-up, hypotheses are formulated in terms of the mean vector as $H_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, where \mathbf{H} is a suitable contrast matrix, i.e., $\mathbf{H}\mathbf{1}_T = \mathbf{0}$. Instead of \mathbf{H} we can use the unique projection matrix $\mathbf{T} = \mathbf{H}'(\mathbf{H}\mathbf{H}')^+ \mathbf{H}$. The projection matrix \mathbf{T} is idempotent and symmetric and $\mathbf{T}\boldsymbol{\mu} = \mathbf{0}$ if and only if $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ (Brunner and Puri, 2001).

An estimator of $\boldsymbol{\mu}_i$ is given by the vector of pooled group means $\bar{\mathbf{Y}}_{i\cdot} = \frac{1}{n_i} \sum_{k=1}^a \mathbf{Y}_{ik}$, $i = 1, \dots, a$ and the covariance matrix \mathbf{V}_i in treatment group i is estimated by the sample covariance matrix

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

Thus, we estimate the covariance matrix $\boldsymbol{\Sigma}_N = \text{Cov}(\sqrt{N} \bar{\mathbf{Y}}_{\bullet}) = \text{diag}(\frac{N}{n_i} \mathbf{V}_i : 1 \leq i \leq a)$ by $\hat{\boldsymbol{\Sigma}}_N = \text{diag}(\frac{N}{n_i} \hat{\mathbf{V}}_i : 1 \leq i \leq a)$, where $\bar{\mathbf{Y}}_{\bullet} = (\bar{\mathbf{Y}}'_{1\cdot}, \dots, \bar{\mathbf{Y}}'_{a\cdot})'$.

As basis for our analyses we consider two test statistics, which are often used in the context of repeated measures designs with $t_i \equiv t$ repeated measures per group (Brunner, 2001), namely the Wald-type statistic (WTS) and the ANOVA-type statistic (ATS). We adapt these test statistics to our more involved design with different numbers of repeated measures per group and show that the small sample behavior of the WTS can be improved by a permutation procedure (Friedrich et al., 2017a).

First, the so-called Wald-type statistic (WTS) is given by

$$Q_N = N \bar{\mathbf{Y}}_{\bullet}' \mathbf{T} (\mathbf{T} \hat{\boldsymbol{\Sigma}}_N \mathbf{T})^+ \mathbf{T} \bar{\mathbf{Y}}_{\bullet}. \quad (2.2)$$

We show that under $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, Q_N has asymptotically, as $N \rightarrow \infty$, a χ^2_f -distribution with $f = \text{rank}(\mathbf{T})$ degrees of freedom (Friedrich et al., 2017a, Theorem 2). However, very large sample sizes are necessary to obtain a valid level α test based on the quantiles of the limiting χ^2 -distribution (e.g., Brunner, 2001; Konietzschke et al., 2015).

Another possible test statistic introduced in Brunner (2001) for repeated measures is the so-called ANOVA-type test statistic (ATS), where we drop the Moore-Penrose term

in (2.2). This leads to the following statistic:

$$T_N = N\bar{\mathbf{Y}}_{\bullet}'\mathbf{T}\bar{\mathbf{Y}}_{\bullet}$$

Under the null hypothesis, the ATS has, asymptotically, the same distribution as a weighted sum of independent χ_1^2 -distributed random variables, where the weights are the eigenvalues of $\mathbf{T}\boldsymbol{\Sigma}$ for $\boldsymbol{\Sigma} = \text{diag}(\kappa_i^{-1}\mathbf{V}_i)$ (Friedrich et al., 2017a, Theorem 1). Thus, the ATS is non-pivotal and the limit distribution has to be estimated, e.g., based on a Box-type approximation (Box, 1954; Brunner, 2001). This results in approximating the scaled ATS

$$\tilde{T}_N = \frac{N}{\text{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}})}\bar{\mathbf{Y}}_{\bullet}'\mathbf{T}\bar{\mathbf{Y}}_{\bullet}$$

by an $F(\hat{\nu}, \infty)$ -distribution with degree of freedom $\hat{\nu} = \text{tr}^2(\mathbf{T}\hat{\boldsymbol{\Sigma}}) / \text{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}}\mathbf{T}\hat{\boldsymbol{\Sigma}})$ (Brunner, 2001). For testing main effects of the whole-plot factors or interactions involving only whole-plot factors, the approximation can be improved by estimating a second degree of freedom $\hat{\nu}_0$, see Brunner et al. (2002) for details. This procedure leads to consistent test decisions for fixed alternatives, but is in general no asymptotic level α test under the null hypothesis, since, even in the asymptotic case, the $F(\hat{\nu}, \infty)$ -distribution is only an approximation of the true distribution of T_N under the null hypothesis (e.g., Brunner et al., 1997, 1999).

In order to improve the small sample behavior of the WTS, we investigate a permutation procedure, where we randomly permute the pooled observation vector (Friedrich et al., 2017a, Section 3).

2.2 Nonparametric repeated measures model

In order to cover situations with ordinal or ordered categorical data, where means provide no adequate measure, we consider a completely nonparametric model. That is, we assume arbitrary marginal distribution functions

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

In this setting, we assume the same number of time points t for all groups. Null hypotheses are formulated as $H_0^F : \mathbf{T}\mathbf{F} = \mathbf{0}$, where $\mathbf{F} = (F_{11}, \dots, F_{at})'$ denotes the vector of distribution functions and \mathbf{T} is the projection matrix defined above.

In order to estimate effects in such a setting, Mann and Whitney (1947) introduced the quantity $w = P(Y_{11} \leq Y_{21}) = \int F_1 dF_2$ for a univariate nonparametric two-sample design with independent observations $Y_{ik} \sim F_i, i = 1, 2, k = 1, \dots, n_i$. By replacing the distribution functions with their empirical counterparts, an estimator of w can be obtained. This effect has several desirable properties, provides a meaningful interpretation of the results and is widely accepted in practice, see e.g., Brumback et al. (2006); Fischer et al. (2014); De Neve et al. (2014); Rauch et al. (2014); Brückner and Brannath (2016). However, a generalization of w to more than two distributions or factorial designs is not obvious. The straightforward generalization to pairwise effects $w_{\ell i} = P(Y_{\ell 1} \leq Y_{i1}), \ell \neq i = 1, \dots, a$ can lead to paradox results in the sense of Efron's Dice, since these effects are not transitive, see e.g., Gardner (1970) and Brown and Hettmansperger (2002).

We therefore consider rank-statistics based on the relative effects $\mathbf{p} = (p_{11}, \dots, p_{at})'$, where

$$p_{is} = \int H_N dF_{is}$$

and $H_N(x) = \frac{1}{tN} \sum_{i=1}^a \sum_{s=1}^t n_i F_{is}(x)$ denotes the weighted mean distribution function. These relative effects avoid the problem of non-transitivity by comparing the distribution functions F_{is} to the same reference distribution H_N . The interpretation of the relative effects in a repeated measures design is as follows: If $p_{is} < p_{is'}$ for some $s \neq s'$, the random measures in group i at time s tend to smaller values than those at time s' .

Since these relative effects depend on the sample sizes, they are no fixed model constants and changing the sample sizes may change the results (Brunner et al., 2016). For testing hypotheses about the effects, one should therefore consider the unweighted mean of the distribution functions $G(x) = \frac{1}{at} \sum_{i=1}^a \sum_{s=1}^t F_{is}(x)$, resulting in the unweighted effects $q_{ij} = \int G dF_{is}$, see Friedrich et al. (2017d, page 50) and Brunner et al. (2016). In Dobler et al. (2017), we consider a wild bootstrap approach to unweighted effects in a nonparametric MANOVA setting, see Chapter 6 for details. Here, however, we focus on null hypotheses formulated in terms of the distribution functions and therefore consider the relative treatment effects p_{is} .

Denoting by R_{iks} the (mid-)rank of Y_{iks} among all tN observations and by $\bar{R}_{i.s} = \frac{1}{n_i} \sum_{k=1}^{n_i} R_{iks}$ the corresponding rank means, estimates of p_{is} are given by

$$\hat{p}_{is} = \frac{1}{tN} \left(\bar{R}_{i.s} - \frac{1}{2} \right).$$

Under $H_0^F : \mathbf{T}\mathbf{F} = \mathbf{0}$, $\sqrt{N}\mathbf{T}(\hat{\mathbf{p}} - \mathbf{p})$ follows asymptotically, as $N \rightarrow \infty$, a multivariate normal distribution with expectation $\mathbf{0}$ and covariance matrix $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}$, where $\boldsymbol{\Sigma} = \text{diag}(\kappa_i^{-1}\mathbf{V}_i)$, see Akritas and Brunner (1997). Here $\mathbf{V}_i = \text{Cov}(\mathbf{X}_{ik})$ denotes the covariance matrix of the random vectors $\mathbf{X}_{ik} = (H(Y_{ik1}), \dots, H(Y_{ikt}))'$ 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 assumption (2.1), see Friedrich et al. (2017d, page 40).

Thus, we calculate the WTS based on the relative effects as

$$Q_N^p = N\hat{\mathbf{p}}'\mathbf{T}(\mathbf{T}\hat{\boldsymbol{\Sigma}}\mathbf{T})^+\mathbf{T}\hat{\mathbf{p}},$$

where $\hat{\boldsymbol{\Sigma}} = \bigoplus_{i=1}^a \frac{N}{n_i} \hat{\mathbf{V}}_i$ and $\hat{\mathbf{V}}_i = \frac{1}{(tN)^2(n_i-1)} \sum_{k=1}^{n_i} (\mathbf{R}_{ik} - \bar{\mathbf{R}}_{i\cdot})(\mathbf{R}_{ik} - \bar{\mathbf{R}}_{i\cdot})'$ denotes the empirical covariance matrix in group i .

Analogous to the semi-parametric case in Section 2.1, the WTS has, under $H_0^F : \mathbf{T}\mathbf{F} = \mathbf{0}$ and if $\mathbf{V}_i > \mathbf{0}$ for all $i = 1, \dots, a$, asymptotically, as $N \rightarrow \infty$, a χ_f^2 -distribution. As in the semi-parametric case, the WTS requires large sample sizes to maintain the pre-assigned level α . Furthermore, it is only applicable in designs with non-singular covariance matrices.

Due to the restriction to non-singular covariance matrices and the weak performance of the WTS for small sample sizes, we also consider a nonparametric ATS, which is given by

$$T_N^p = N\hat{\mathbf{p}}'\mathbf{T}\hat{\mathbf{p}}$$

in this setting. Under the null hypothesis, T_N^p has, asymptotically, the same distribution as a weighted sum of χ_1^2 -distributed random variables, where the weights are the eigenvalues of $\mathbf{T}\boldsymbol{\Sigma}$ (Brunner et al., 2016). Since the eigenvalues are unknown, the limiting distribution has to be approximated again, e.g. by an F -distribution, see Brunner et al. (2016). Again, the corresponding ATS test provides in general no asymptotic level α test.

In Friedrich et al. (2017d), we investigate a wild bootstrap approach to improve the small sample behavior of both the WTS and the ATS in this nonparametric setting.

2.3 Semi-parametric MANOVA model

In this section, we consider multivariate data, which may be measured on different scales. Therefore, we consider d -dimensional random vectors

$$\mathbf{Y}_{ik} = (Y_{ik1}, \dots, Y_{ikd})', \quad i = 1, \dots, a, \quad k = 1, \dots, n_i$$

of individual k in treatment group i . We assume existence of the group-specific expectation vector $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{id})' = \mathbb{E}(\mathbf{Y}_{i1})$ and the covariance matrix $\mathbf{V}_i = \text{Cov}(\mathbf{Y}_{i1}) \geq 0$. Note that covariance matrices may differ between groups and we do not assume any special structure. In particular, we allow for singular covariance matrices in this setting. Our only distributional assumption is the existence of finite second moments, i.e., $0 < \text{Var}(Y_{iks}) =: \sigma_{is}^2 < \infty$, $i = 1, \dots, a$, $k = 1, \dots, n_i$, $s = 1, \dots, d$. We aggregate the observation vectors in

$$\mathbf{Y} = (\mathbf{Y}'_{11}, \dots, \mathbf{Y}'_{an_a})'$$

as well as

$$\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \dots, \boldsymbol{\mu}'_a)'$$

We consider hypotheses formulated in terms of the mean vector as $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, where \mathbf{T} denotes the unique projection matrix.

Analogous to Section 2.1, we estimate $\boldsymbol{\mu}_i$ again by the vector of pooled group means $\bar{\mathbf{Y}}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{Y}_{ik}$, $i = 1, \dots, a$ and the covariance matrix \mathbf{V}_i in treatment group i by the corresponding sample covariance matrix $\hat{\mathbf{V}}_i$ as well as $\hat{\boldsymbol{\Sigma}}_N = \text{diag}(\frac{N}{n_i} \hat{\mathbf{V}}_i : 1 \leq i \leq a)$.

The two commonly considered test statistics WTS and ATS have several drawbacks in this multivariate setting: First, the WTS does not provide a valid test in designs involving singular covariance matrices. The ATS, on the other hand, is in general not applicable to multivariate data, where the endpoints are measured on different scales, since it is not invariant under transformations of the data like change in units (e.g., $cm \mapsto m$ or $kg \mapsto g$).

We therefore propose a different test statistic, which we denote as MATS. It is motivated from the test statistic proposed by Srivastava and Kubokawa (2013) in the special context of a high-dimensional homoscedastic one-way layout. In particular, we con-

sider

$$M_N = N\bar{\mathbf{Y}}' \mathbf{T} (\mathbf{T} \hat{\mathbf{D}}_N \mathbf{T})^+ \mathbf{T} \bar{\mathbf{Y}},$$

where $\hat{\mathbf{D}}_N = \text{diag} \left(\frac{N}{n_i} \hat{\sigma}_{is}^2 \right)$, $i = 1, \dots, a$, $s = 1, \dots, d$ and $\hat{\sigma}_{is}^2$ denotes the empirical variance of component s in group i . Under $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, the MATS has asymptotically, as $N \rightarrow \infty$, the same distribution as a weighted sum of χ_1^2 -distributed random variables, where the weights are the eigenvalues of $\mathbf{T}(\mathbf{T}\mathbf{D}\mathbf{T})^+ \mathbf{T}\boldsymbol{\Sigma}$ for $\mathbf{D} = \text{diag}(\kappa_i^{-1} \sigma_{is}^2)$ (Friedrich and Pauly, 2017, Theorem 2.1).

The MATS has several advantages in the multivariate setting: First, it is invariant under scale transformations of the data, e.g., under change of units in one or more components. Second, in contrast to the WTS, we do not need to assume non-singular covariance matrices and only the existence of finite second moments is required. However, the limiting distribution of the MATS is again non-pivotal and we can not base inference on quantiles of this distribution directly. We therefore consider three bootstrap approaches in order to derive data-driven quantiles for test decisions. Furthermore, we derive confidence regions and simultaneous confidence intervals for contrasts of the mean vector based on the bootstrap quantiles (Friedrich and Pauly, 2017, Section 4.2).

3 Resampling procedures

Resampling methods are a class of inference procedures known for inducing robust results even for small sample sizes, see e.g., Efron and Tibshirani (1994); Davison and Hinkley (1997); Davison et al. (2003); Good (2006); Manly (2006); Konietschke et al. (2015); Pauly et al. (2015). The idea of the methods is to base inference on data-dependent critical values instead of critical values of the approximate distribution. The corresponding resampling test is (at least) asymptotically valid, if the distribution of the test statistic under the null and the conditional resampling distribution coincide asymptotically. Several different resampling approaches have been considered in the literature. The most well known, perhaps, is Efron's nonparametric bootstrap (Efron, 1979), which is based on randomly drawing with replacement from the data. However, simulation studies indicated that this bootstrap procedure may lead to liberal test decisions in several non- and semi-parametric setups, see e.g., Konietschke et al. (2015) for general univariate MANOVA. This result has been confirmed by our simulation studies in the context of semi-parametric repeated measures data, see Tables 6–11 in the supplementary material to Friedrich et al. (2017a).

In practice, the p -value based on the resampling distribution can be numerically calculated as follows:

1. Given the data \mathbf{Y} , calculate the test statistic of interest, e.g., the WTS Q_N .
2. Resample the data according to the resampling procedure of your choice, for example in case of the nonparametric bootstrap draw a random sample with replacement from the data¹.
3. Calculate the test statistic of interest based on the resampling sample and save its value in A_1 .
4. Repeat steps 2 and 3 a large number of times, e.g., $B = 10,000$ times, resulting in values A_1, \dots, A_B .

¹Details on the resampling procedures will be discussed below.

5. The p -value is calculated based on the empirical resampling distribution as



$$p\text{-value} = \frac{1}{B} \sum_{b=1}^B \mathbb{1}\{A_b \leq Q_N\}.$$

Note that for very small sample sizes, it is often possible (e.g., for permutation procedures) to calculate the resampling distribution exactly. However, this is even for sample sizes of, e.g., $N = 20$ computationally extremely intensive and therefore usually not feasible in practice.

In the following, we will describe the three resampling procedures that provided the best results in the settings considered in this thesis.

3.1 Permutation procedure

In the semi-parametric repeated measures design described in Section 2.1, we consider a permutation procedure for the WTS, which is based on a random permutation \mathbf{Y}^π of all elements of the pooled sample \mathbf{Y} . Here, Y_{iks}^π denotes the (i, k, s) -entry of the permuted vector \mathbf{Y}^π . Obviously, the original data vector \mathbf{Y} and the permuted vector \mathbf{Y}^π only have the same distribution, if the components of \mathbf{Y} are exchangeable. However, this is often not even the case in univariate two- and higher-way layouts and becomes more untenable in the context of repeated measures.

Our arguments generalize the idea of Pauly et al. (2015), where a permutation approach is applied in the context of univariate factorial designs. This approach is implemented in the R package **GFD**, see Section 4.1 below as well as the Appendix (Friedrich et al., 2017b).

Thus, following the idea of Janssen (1997, 2005); Chung and Romano (2013) and Pauly et al. (2015), we consider a studentized test statistic. Therefore, we calculate the permutation Wald-type statistic (WTPS) based on the mean vectors $\bar{\mathbf{Y}}_\bullet^\pi$ and empirical covariance matrices $\hat{\Sigma}_N^\pi$ of the permuted observations as

$$Q_N^\pi = N(\bar{\mathbf{Y}}_\bullet^\pi)' \mathbf{T} (\mathbf{T} \hat{\Sigma}_N^\pi \mathbf{T})^+ \mathbf{T} \bar{\mathbf{Y}}_\bullet^\pi. \quad (3.1)$$

Due to the involved studentization, this approach is asymptotically correct, that is, it always mimics the null distribution of the WTS. In particular, the following theorem holds (Friedrich et al., 2017a, Theorem 3):

THEOREM 3.1 *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 $f = \text{rank}(\mathbf{T})$.*

Theorem 3.1 states that for any underlying parameter $\boldsymbol{\mu} \in \mathbb{R}^T$ and $\boldsymbol{\mu}_0 \in H_0(\mathbf{T})$ with $\mathbf{T}\boldsymbol{\mu}_0 = \mathbf{0}$ we have convergence in probability

$$\sup_{x \in \mathbb{R}} |P_{\boldsymbol{\mu}}(Q_N^\pi \leq x | \mathbf{Y}) - P_{\boldsymbol{\mu}_0}(Q_N \leq x)| \rightarrow 0,$$

where $P_{\boldsymbol{\mu}}(Q_N^\pi \leq x | \mathbf{Y})$ and $P_{\boldsymbol{\mu}}(Q_N \leq x)$ denote the conditional and unconditional distribution function of Q_N^π and Q_N , respectively, under the assumption that $\boldsymbol{\mu}$ is the true underlying parameter. As described earlier, this is the crucial criterion for obtaining a valid resampling approach. In particular, we obtain an asymptotic level α test by comparing the original test statistic Q_N to the conditional $(1 - \alpha)$ -quantile of the permutation distribution. It follows that the permutation test asymptotically keeps the pre-assigned level α under the null hypothesis and is consistent for fixed alternatives. Moreover, the WTS and the WTPS are asymptotically equivalent and the relative efficiency of the WTPS compared to the WTS is 1. In addition, the permutation test based on the WTS is finitely exact under exchangeability of the data (Friedrich et al., 2017a, page 259).

Note that this permutation procedure does not work for the ATS, since it would result in an incorrect covariance structure. For similar reasons, it is also problematic in the more general nonparametric case $H_0^F : \mathbf{T}\mathbf{F} = \mathbf{0}$.

3.2 Wild Bootstrap

The wild bootstrap approach is based on multiplying the fixed (often centered) data with random weights. To this end, let W_{ik} denote i.i.d. random variables with $E(W_{ik}) = 0$, $\text{Var}(W_{ik}) = 1$ and $\sup_{i,k} E(W_{ik}^4) < \infty$, which are independent of \mathbf{Y} . Depending on the situation, different choices of weights are possible, some satisfying different moment conditions (Wu, 1986; Liu, 1988; Mammen, 1993a). In our applications, we focus on random signs, that is, Rademacher distributed random variables, as well as standard normal weights. Such resampling methods have been successfully applied in the context of regression analysis (e.g., Wu, 1986; Mammen, 1993b; Davidson and Flachaire, 2008), in time-series problems (e.g., Kreiss and Paparoditis, 2011; Jentsch and Pauly, 2015) and in survival analysis (e.g., Lin, 1997; Martinussen and Scheike,

2007; Beyersmann et al., 2013; Dobler and Pauly, 2014). We apply the wild bootstrap approach both in the nonparametric repeated measures setting (Friedrich et al., 2017d) and in the semi-parametric setting with multivariate data (Friedrich and Pauly, 2017). In this section, we will focus on the latter. The approach in the nonparametric case is similar and will be explained in more detail in Section 5.2 below.

In the semi-parametric setting, we obtain a wild bootstrap sample as

$$\mathbf{Y}_{ik}^* = W_{ik}(\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_{i\cdot}), \quad i = 1, \dots, a, \quad k = 1, \dots, n_i. \quad (3.2)$$

Based on these bootstrap variables we can calculate the test statistic of interest, in this case the MATS, as

$$M_N^* = N(\bar{\mathbf{Y}}_{\bullet}^*)' \mathbf{T} (\mathbf{T} \hat{\mathbf{D}}_N^* \mathbf{T})^+ \mathbf{T} \bar{\mathbf{Y}}_{\bullet}^*,$$

where $\bar{\mathbf{Y}}_{\bullet}^*$ denotes the (empirical) mean vector of the wild bootstrap sample and, similarly, $\hat{\mathbf{D}}_N^*$ is calculated based on the empirical variances of the wild bootstrap sample.

We obtain a wild bootstrap test by comparing the original test statistic to the conditional $(1 - \alpha)$ -quantile of its wild bootstrap version. This test is asymptotically valid and consistent for fixed alternatives (Friedrich and Pauly, 2017, Theorem 3.2).

In the nonparametric repeated measures setting, the wild bootstrap is applied to both WTS and ATS and leads to asymptotically valid tests in both cases. Furthermore, the wild bootstrap tests have the same local power under contiguous alternatives as the original tests (Friedrich et al., 2017d, page 43).

3.3 Parametric Bootstrap

The parametric bootstrap approach, also known as asymptotic model based bootstrap (Konietschke et al., 2015; Bathke et al., 2016), is typically applied for parametric models. However, although it originates from the assumption of multivariate normality, the parametric bootstrap yields valid results in our semi-parametric setting. The approach is based on an application of the multivariate central limit theorem: For any fixed $i = 1, \dots, a$, the central limit theorem implies that $\sqrt{n_i}(\bar{\mathbf{Y}}_{i\cdot} - \boldsymbol{\mu}_i)$ is asymptotically normal with mean zero and covariance matrix \mathbf{V}_i . Thus, given the data, we generate a parametric bootstrap sample as

$$\mathbf{Y}_{i1}^*, \dots, \mathbf{Y}_{in_i}^* \sim N(\mathbf{0}, \hat{\mathbf{V}}_i), \quad i = 1, \dots, a.$$

Here and in (3.2), we chose the subscripts $*$ and $*$ to match the notation in Friedrich and Pauly (2017, Section 3). Since the parametric bootstrap variables mimic the covariance structure of the data, we obtain an accurate finite sample approximation by calculating the test statistic (e.g., M_N^*) based on the bootstrap variables \mathbf{Y}_{ik}^* . Again, we can show that the conditional distribution of M_N^* given the data weakly converges to the null distribution of M_N in probability under both the null hypothesis and the alternative (Friedrich and Pauly, 2017, Theorem 3.1). Thus, the parametric bootstrap also provides an asymptotically valid test, which is consistent for fixed alternatives.

4 *R* packages and the EEG data example

4.1 The GFD package

The permutation approach investigated in Pauly et al. (2015) for general univariate factorial designs is implemented in the R package **GFD** (Friedrich et al., 2017b, see Appendix). In this setting, we assume the following model of the univariate observations

$$Y_{ik} = \mu_i + \varepsilon_{ik}, i = 1, \dots, a, k = 1, \dots, n_i.$$

The error terms ε_{ik} are assumed to be i.i.d. with $E(\varepsilon_{i1}) = 0$ and $\text{Var}(\varepsilon_{i1}) = \sigma_i^2 > 0$. As in the multivariate models we neither assume normality nor equal variances or sample sizes across groups. Null hypotheses are formulated in terms of the mean vector as $H_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_a)'$ and \mathbf{H} is a suitable contrast matrix. Pauly et al. (2015) showed that a permutation procedure of the WTS yields accurate type-I error control and an asymptotically valid test in this context. This permutation procedure is implemented in **GFD** along with the asymptotic χ^2 -approximation of the WTS and the F -approximation of the ATS. All methods can be used to test hypotheses about the main and interaction effects of the involved factors. Application and output of the function `GFD()` are similar to the `lm()` or `aov()` function in R. In particular, `summary()`, `print()` and `plot()` methods are implemented for an object of class 'GFD'. The plotting routine displays the calculated mean values along with asymptotic $(1 - \alpha)$ confidence intervals based on quantiles of a t -distribution. Furthermore, the package is equipped with an optional graphical user interface in order to facilitate application for a wide range of users.

The methods implemented in **GFD** can be applied to arbitrary crossed or hierarchically nested designs with up to three factors. A detailed description of the package including many examples to demonstrate its use for different designs is published in Friedrich et al. (2017b), which is included in the Appendix of this thesis.

4.2 MANOVA.RM and the EEG data example

The methods for the semi-parametric models analyzed in Friedrich et al. (2017a) and Friedrich and Pauly (2017) are implemented in the R package **MANOVA.RM** (Friedrich et al., 2017c). The package consists of two parts: the function `RM()` calculates test statistics and p -values for repeated measures designs (data must be provided in long format), while `MANOVA()` and `MANOVA.wide()` provide functions for multivariate data in long and wide format, respectively. Since the permutation approach is not feasible for multivariate data, the implemented resampling methods differ with respect to the different types of data: For multivariate data, the package provides a parametric bootstrap and a wild bootstrap based on Rademacher weights for the MATS (Friedrich and Pauly, 2017) and the WTS. In case of repeated measures, we additionally implemented the permutation approach for the WTS as in Friedrich et al. (2017a) for a common number of time points $t_i \equiv t$ in each group as well as the ATS with F -approximation and the two bootstrap approaches.

To demonstrate the performance of the package and to illustrate the difference between a repeated measures setting and multivariate outcomes, we consider the EEG data example described in Section 1 and in Bathke et al. (2016). This data set contains EEG measurements of 166 patients diagnosed with either Alzheimer's disease (AD), mild cognitive impairment (MCI) or subjective cognitive complaints without clinically significant deficits (SCC). For each patient, we consider six EEG measurements, which consist of z -scores for brain rate (Pop-Jordanova and Pop-Jordanova, 2005) and Hjorth complexity (Hjorth, 1970, 1975) averaged within frontal, temporal and central positions.

Since comparisons between the variables are meaningful here, the data can be viewed in two different ways: First, we may consider it as multivariate, i.e., each individual has a 6-dimensional response vector. We can then formulate and test the hypothesis of no difference between diagnoses (3 levels)

$$H_0 : \{(\mathbf{P}_3 \otimes \mathbf{I}_6)\boldsymbol{\mu} = \mathbf{0}\}: \text{No effect of diagnosis.}$$

Here, $\boldsymbol{\mu}$ is the pooled mean vector as in Section 2.3. The analysis of this data example can be conducted using the R package **MANOVA.RM** (Friedrich et al., 2017c). The EEG data set is included in the package and evaluation of the above hypothesis is carried out as follows:

```
R > library(MANOVA.RM)
R > data(EEG)
```

4.2 MANOVA.RM and the EEG data example

```
R > multi <- MANOVA(resp ~ diagnosis, data = EEG, subject = "id",
  iter = 10000, resampling = "paramBS", seed = 123,
  dec = 2)
R > summary(multi)
```

This results in the following output, where we have rounded the results to 2 digits (parameter 'dec').

```
Call:
resp ~ diagnosis

Descriptive:
  diagnosis  n  Means  Means  Means  Means  Means  Means
1         AD 36 -0.52 -0.44 -0.53 -0.57 -0.34 -0.58
2         MCI 57 -0.28 -0.26 -0.27 -0.17 -0.15 -0.07
3         SCC 67  0.51  0.46  0.51  0.45  0.31  0.37

Wald-Type Statistic (WTS):
Test statistic    df    p-value
      53.55         12    3.28e-07

modified ANOVA-Type Statistic (MATS):
  Test statistic
      193.62

p-values resampling:
paramBS (WTS)  paramBS (MATS)
      2e-04         0e+00
```

In the descriptive part, the mean vectors for the different diagnoses are reported along with the corresponding sample sizes in the groups. We find a highly significant effect of diagnosis on the 6-dimensional EEG measurements, a finding shared by all testing procedures including the parametric bootstrap approach. Since the ATS is not invariant under scale transformations in the multivariate setting, it has not been implemented.

On the other hand, the EEG measurements might also be viewed as levels of the sub-plot factors 'brain region' (3 levels: frontal, temporal, central) and 'feature' (2 levels: brain rate, complexity). Therefore, we can formulate and test hypotheses involving the sub-plot factors, for example

1. $H_0^{(1)} : \{(P_3 \otimes \frac{1}{3}1'_3 \otimes \frac{1}{2}1'_2)\mu = \mathbf{0}\}$: No effect of diagnosis (whole-plot factor)
2. $H_0^{(2)} : \{(\frac{1}{3}1'_3 \otimes \frac{1}{3}1'_3 \otimes P_2)\mu = \mathbf{0}\}$: No effect of feature (second sub-plot factor)

4 R packages and the EEG data example

3. $H_0^{(3)} : \{(\frac{1}{3}\mathbf{1}'_3 \otimes \mathbf{P}_3 \otimes \mathbf{P}_2)\boldsymbol{\mu} = \mathbf{0}\}$: No two-fold interaction between feature and brain region
4. $H_0^{(4)} : \{(\mathbf{P}_3 \otimes \mathbf{P}_3 \otimes \mathbf{P}_2)\boldsymbol{\mu} = \mathbf{0}\}$: No three-fold interaction
5. ...

This analysis may be conducted with **MANOVA.RM** as follows:

```
R > rm <- RM(resp ~ diagnosis * region * feature, data = EEG,
             subject = "id", no.subf = 2, iter = 10000,
             resampling = "Perm", seed = 456, CI.method = "t-quantile",
             dec = 2)
R > summary(rm)
```

The number of sub-plot factors considered must be specified in the RM function via `no.subf`. The returned output is displayed below:

Call:

```
resp ~ diagnosis * region * feature
```

Descriptive:

	diagnosis	region	feature	n	Means	Lower	Upper	95% CI
1	AD	central	brainrate	36	-0.53	-1.38	0.32	
10	AD	central	complexity	36	-0.58	-1.84	0.68	
4	AD	frontal	brainrate	36	-0.44	-1.31	0.42	
13	AD	frontal	complexity	36	-0.34	-1.29	0.60	
7	AD	temporal	brainrate	36	-0.52	-1.29	0.25	
16	AD	temporal	complexity	36	-0.57	-1.58	0.44	
2	MCI	central	brainrate	57	-0.27	-0.65	0.12	
11	MCI	central	complexity	57	-0.07	-0.34	0.20	
5	MCI	frontal	brainrate	57	-0.26	-0.65	0.13	
14	MCI	frontal	complexity	57	-0.15	-0.60	0.31	
8	MCI	temporal	brainrate	57	-0.28	-0.66	0.12	
17	MCI	temporal	complexity	57	-0.17	-0.56	0.23	
3	SCC	central	brainrate	67	0.51	0.24	0.79	
12	SCC	central	complexity	67	0.37	0.25	0.50	
6	SCC	frontal	brainrate	67	0.46	0.17	0.75	
15	SCC	frontal	complexity	67	0.31	0.07	0.55	
9	SCC	temporal	brainrate	67	0.51	0.20	0.82	
18	SCC	temporal	complexity	67	0.45	0.25	0.64	

Wald-Type Statistic (WTS):

	Test statistic	df	p-value
diagnosis	42.59	2	5.66-10

4.2 MANOVA.RM and the EEG data example

region	0.35	2	0.84
diagnosis:region	6.08	4	0.19
feature	0.01	1	0.91
diagnosis:feature	5.60	2	0.06
region:feature	0.35	2	0.84
diagnosis:region:feature	8.45	4	0.08

ANOVA-Type Statistic (ATS):

	Test statistic	df1	df2	p-value
diagnosis	13.33	1.47	2088.03	2.50e-05
region	0.13	1.73	Inf	0.85
diagnosis:region	1.21	2.41	Inf	0.30
feature	0.01	1.00	Inf	0.91
diagnosis:feature	1.66	1.70	Inf	0.19
region:feature	0.08	1.57	Inf	0.88
diagnosis:region:feature	1.09	1.98	Inf	0.34

p-values resampling:

	Perm (WTS)	Perm (ATS)
diagnosis	0.00	NA
region	0.84	NA
diagnosis:region	0.21	NA
feature	0.91	NA
diagnosis:feature	0.06	NA
region:feature	0.85	NA
diagnosis:region:feature	0.09	NA

None of the p -values and confidence intervals are automatically adjusted for multiple testing. Since the permutation approach is not appropriate for the ATS, the program returns no resampled p -value. In addition, for tests involving only the whole-plot factors (here: diagnosis), a second degree of freedom $\hat{f}_0 = \text{tr}^2(\mathbf{T}\hat{\Sigma}) / \text{tr}(\mathbf{D}_T^2 \hat{\Sigma}^2 \mathbf{\Lambda})$ is calculated for the F -approximation, similar to Brunner et al. (1997, 2002) and the SAS PROC MIXED procedure (SAS Institute Inc., 2003). Here $\mathbf{D}_T = \text{diag}(\mathbf{T})$ denotes the matrix consisting of the diagonal entries in \mathbf{T} and $\mathbf{\Lambda} = \text{diag}(1/(n_1 - 1), \dots, 1/(n_a - 1))$.

We find a significant effect of the whole-plot factor ‘diagnosis’ (again shared by all test statistics), but none of the sub-plot or interaction effects are significant at 5% level.

The output demonstrates the differences between the two analyses: In the repeated measures model, we have one mean value per factor level combination, which is additionally equipped with 95% confidence intervals based on either t -quantiles or the corresponding resampling quantiles. In the MANOVA setting, on the other hand, we have 6-dimensional mean vectors for each level of the factor ‘diagnosis’. Furthermore,

we implemented a function for calculation of confidence regions in the multivariate setting (`conf.reg()`), thus replacing the univariate confidence intervals in the repeated measures setting with multivariate confidence regions. Furthermore, the MATS is calculated in the MANOVA setting instead of the ATS along with the corresponding resampling-based p -values. Since the permutation procedure is not meaningful in the context of multivariate data, it is not available in the `MANOVA()` function. The `RM()` function is equipped with a plotting routine similar to the one implemented in **GFD**, which displays mean values along with $(1 - \alpha)$ confidence intervals for a specified factor (combination) of interest. In the case of two-dimensional multivariate data, it is possible to plot the corresponding confidence regions. An optional graphical user interface is available for both the `RM()` and the `MANOVA()` function.

Note that the two-sided view on the data demonstrated here only makes sense in situations where the response variables are commensurate in the sense that comparisons between them are meaningful (Bathke et al., 2016). Usually we distinguish between either of the two approaches.

5 Summary of the articles

5.1 Article 1: ‘Permuting longitudinal data in spite of the dependencies’ (JMVA, 2017)

We consider the semi-parametric model for general repeated measures designs with potentially non-normal and/or heteroscedastic data described in Section 2.1. In such situations, the WTS provides an asymptotically valid procedure. However, for small to moderate sample sizes, test decisions based on the asymptotic χ^2 -quantiles become very liberal.

As an extension to the standard repeated measures setting, we allow the number of time points to differ between groups. We generalize all theorems on the asymptotic behavior of the WTS and the ATS to this more involved setting and state the power behavior of the considered methods (Theorems 1 and 2).

To improve the small sample behavior of the WTS, we propose a novel permutation approach, where we randomly permute all elements of the pooled sample. We calculate the permutation Wald-type statistic (WTPS) based on the mean vectors and empirical covariance matrices of the permuted observations, i.e.,

$$Q_N^\pi = N(\overline{\mathbf{Y}}_\bullet^\pi)' \mathbf{T} (\mathbf{T} \widehat{\Sigma}_N^\pi \mathbf{T})^+ \mathbf{T} \overline{\mathbf{Y}}_\bullet^\pi.$$

Due to the dependencies in the repeated measures data, the idea of how to permute is more involved here than in the case of independent univariate observations (Pauly et al., 2015). Heuristically, an explanation of why the above approach works is as follows: When multiplied by a contrast matrix, the permuted mean vector always mimics the null situation, because the permuted components unconditionally have the same mean (page 259).

More precisely, we prove that the conditional distribution of the WTPS always approx-

imates the null distribution of the WTS given the data (Theorem 3).

In a large simulation study, we analyze the behavior of the WTPS for different distributions, sample sizes and covariance settings and compare it to the WTS and the ATS (pages 260–263). In accordance with the literature (e.g., Brunner et al., 1997; Brunner, 2001; Pauly et al., 2015; Konietzschke et al., 2015; Smaga, 2017), the test based on the WTS considerably exceeds the nominal level, reaching type-I error rates of almost 50% in some scenarios. The ATS, in contrast, leads to conservative test decisions in case of non-normal data and a large number of repeated measurements. The WTPS keeps the pre-assigned level in most scenarios. However, in settings with negative (positive) pairing, the permutation test shows a slightly conservative (liberal) behavior. Similar problems have been noted before for permutation tests, see, e.g., Pauly et al. (2015). In terms of power, the ATS has a slightly higher power than the WTPS in situations with normally distributed data. For skewed distributions, however, the WTPS has the highest power.

As a real data example we consider a study on the oxygen consumption of leukocytes in the presence and absence of inactivated staphylococci. This is a repeated measures design with whole-plot factor ‘treatment’ and sub-plot factors ‘time’ and ‘staphylococci’. Questions of interest include the effects of the three factors as well as interactions between them. Since there are only 12 observations per treatment group, the test based on the WTS is not reliable. The analysis of the data example with the three test statistics considered shows a significant effect of all three factors as well as a significant interaction between treatment and time. The three test statistics all lead to the same conclusions in this example.

To sum up, we have considered a permutation test for semi-parametric repeated measurements, where we randomly permute the pooled sample. Despite the somewhat counterintuitive destruction of the dependency structure, the approach turns out to perform very well in simulation studies. We have furthermore rigorously proven the asymptotic correctness of the permutation procedure generalizing arguments from Pauly et al. (2015) using results from Pauly (2011) to the more involved situation with dependent data.

5.2 Article 2: ‘A wild bootstrap approach for nonparametric repeated measurements’ (CSDA, 2017)

As a motivating data example, we consider the shoulder tip pain trial described in Section 1 of this thesis. In such a setting with score data, classical repeated measures ANOVA models show their limits, since means are no adequate measure of deviations between the groups. We therefore consider a nonparametric repeated measures model, where inference is based on ranks, and improve the small sample behavior of the WTS and the ATS by a wild bootstrap approach.

The wild bootstrap approach considered is based on multiplying the fixed data with random signs W_{ik} , i.e., Rademacher distributed weights. We consider the centered rank vectors $\mathbf{Z}_{ik} = (\mathbf{R}_{ik} - \overline{\mathbf{R}}_{i.})$ as well as independent and identically distributed random signs W_{ik} . The resampling version of the estimated treatment effects $\hat{\mathbf{p}}_i$ is given by

$$\hat{\mathbf{p}}_i^* = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{tN} W_{ik} \mathbf{Z}_{ik}.$$

Since the distributions of $\sqrt{N}\mathbf{T}\hat{\mathbf{p}}^*$ and $\sqrt{N}\mathbf{T}(\hat{\mathbf{p}} - \mathbf{p})$ are asymptotically identical under the null hypothesis $H_0^F : \mathbf{T}\mathbf{F} = \mathbf{0}$ (Theorem 3.1), we can derive wild bootstrap versions of the WTS and ATS by plugging in the corresponding wild bootstrap estimates as well as the corresponding covariance matrix, resulting in

$$(Q_N^p)^* = N(\hat{\mathbf{p}}^*)' \mathbf{T} (\mathbf{T} \hat{\Sigma}^* \mathbf{T})^{-1} \mathbf{T} \hat{\mathbf{p}}^*$$

for the WTS as well as

$$(T_N^p)^* = N(\hat{\mathbf{p}}^*)' \mathbf{T} \hat{\mathbf{p}}^*$$

for the ATS.

Both conditional distributions mimic the corresponding null distribution of the WTS and ATS, respectively, thus leading to asymptotic level α tests, which are consistent for fixed alternatives (Theorem 3.2 and 3.3).

We analyze the behavior of the test statistics in a simulation study, where we simulate a one-way repeated measures design with two groups and $t \in \{4, 8\}$ repeated measures for underlying discrete and continuous distributions (pages 43–46). The WTS tends to

liberal decisions in all scenarios, a behavior that is greatly improved by its wild bootstrap version. Roughly speaking, the wild bootstrap WTS takes the variability of the covariance matrix into account and is therefore closer to the actual sampling distribution of the WTS when sample sizes are small. We observe a similar behavior for the ATS, which is also slightly liberal when testing for the main effect. The wild bootstrap ATS tends to an accurate type-I error control. The wild bootstrap WTS has lower power than the wild bootstrap version of the ATS, especially under a trend alternative, while the power is comparable for both versions of the ATS.

Due to the results of the simulation study and since the wild bootstrap ATS can even be applied in situations with singular covariance matrices (see the data example below), we recommend this procedure for practical applications.

When analyzing the data example, it turns out that the estimated covariance matrix is singular. Thus, the WTS cannot be used for the analysis. The ATS and its wild bootstrap version lead to the same conclusions, namely a significant treatment and time effect as well as an interaction between the two. A further analysis in the data set split by treatment reveals a significant time trend under placebo, while there is no significant effect in the treatment group. In a sensitivity analysis, we drop time point 6 from the analysis, leading to a non-singular covariance matrix. We are therefore able to calculate the WTS and its wild bootstrap version. Due to its liberality, the WTS detects several significant effects which are not supported by the other test statistics. In particular, the time effect is no longer significant in the analysis of the complete data set. Only when split up according to treatment, all four procedures again find a significant effect of time in the placebo group (pages 46–48).

In summary, we have considered a wild bootstrap approach to improve the small sample behavior of both the WTS and the ATS in a nonparametric repeated measures design, where null hypotheses are formulated in terms of distribution functions. The test statistics are based on ranks, thus providing methods to deal with metric as well as ordinal, count or score data in a unified way. Extensive simulations have shown that the wild bootstrap approach indeed improves the small sample behavior of both test statistics. We have furthermore proven that the method is asymptotically valid and has the same local power under contiguous alternatives as the original tests based on the WTS and ATS, respectively.

5.3 Article 3: ‘MATS: Inference for potentially singular and heteroscedastic MANOVA’ (2017)

We consider the data example on cardiologic measurements recorded at Ulm University Hospital. Remember that in addition to (empirical) covariance heterogeneity, the covariance matrices are also singular in this example. Furthermore, the cardiologic outcomes are measured on different scales. We can therefore neither apply the WTS, because it relies on the assumption of non-singular covariance matrices, nor the ATS, since it is not invariant under scale transformations of the data. Thus, we introduce a new test statistic, which we denote as MATS:

$$M_N = N\bar{\mathbf{Y}}_{\bullet}' \mathbf{T} (\mathbf{T} \hat{\mathbf{D}}_N \mathbf{T})^+ \mathbf{T} \bar{\mathbf{Y}}_{\bullet}.$$

Srivastava and Kubokawa (2013) proposed a similar test statistic for a specific homoscedastic one-way layout in the context of high-dimensional data. The MATS modifies and extends their test statistic to general factorial MANOVA designs, including heteroscedastic models. In particular, our only distributional assumption is the existence of second moments, thereby incorporating designs with singular covariance matrices and relaxing the usual assumption on finite fourth moments of the data (see, e.g., Pauly et al., 2015; Konietzschke et al., 2015; Friedrich et al., 2017a).

We derive the asymptotic distribution of the MATS in Theorem 2.1. Since it is non-pivotal, we analyze three different resampling techniques to base inference upon. The first approach is a parametric bootstrap procedure as proposed by Konietzschke et al. (2015) for the WTS in general MANOVA designs. The second approach is a wild bootstrap, which has already been successfully applied in the context of nonparametric repeated measures (Friedrich et al., 2017d) and cluster data (Cameron et al., 2008; Cameron and Miller, 2015). In our simulation study, we focus on standard normally distributed weights in the wild bootstrap approach, since they yielded similar results as random signs in the simulations. The asymptotic results are, however, valid for all choices of weights W_{ik} satisfying $E(W_{ik}) = 0$, $\text{Var}(W_{ik}) = 1$ and $\sup_{i,k} E(W_{ik}^4) < \infty$. Finally, we consider a nonparametric bootstrap, where for each group $i = 1, \dots, a$ we randomly draw with replacement n_i independent selections $\mathbf{Y}_{ik}^{\dagger}$ from the i -th sample.

All bootstrap procedures lead to asymptotically valid level α tests, which are consistent for fixed alternatives (Theorems 3.1–3.3). Moreover, the asymptotic relative efficiency of the bootstrap tests compared to the test based on the asymptotic distribution is 1. In addition to statistical testing decisions, we also use the bootstrap quantiles to derive

confidence regions and simultaneous confidence intervals for contrasts of the mean vector (pages 7–8).

In a large simulation study, we analyze the behavior of the bootstrap procedures in different one- and two-way layouts with balanced and unbalanced designs, for $d = 4$ and $d = 8$ dimensions, as well as in settings with singular covariance matrices (pages 9–15). As a competitor to the MATS, we consider the parametric bootstrap of the WTS as proposed by Konietschke et al. (2015), since it turned out to perform better than other resampling approaches in their simulations. We find that the parametric bootstrap of the MATS yields better results than the parametric bootstrap of the WTS, especially in situations with negative pairing. Furthermore, the parametric bootstrap WTS yields no valid level α test in situations with singular covariance matrices, while the MATS is still valid in these settings.

To our surprise, the bootstrap approaches of the MATS improve with growing number of dimensions. Therefore, it seems worth to investigate this approach also in high-dimensional settings where the dimension d is larger than the sample size N . The parametric bootstrap for the MATS yields the best results in most scenarios. In particular, it is less liberal than the wild bootstrap and the parametric bootstrap of the WTS and less conservative than the nonparametric bootstrap. Furthermore, it has a higher power to detect fixed alternatives than the nonparametric bootstrap. However, it shows a slightly liberal behavior in situations with negative pairing and skewed distributions.

Since the singularity of the covariance matrix in the data example on cardiologic measurements is somewhat artificial, we additionally consider a data example on the 2016 presidential elections in the USA. Our aim is to investigate whether 7 demographic factors differ between the 43 states under consideration. In this example, the empirical covariance matrix is computationally singular. We therefore apply the parametric bootstrap of the MATS, which reveals a significant difference between the states with respect to the 7 demographic factors considered (page 16).

To sum up, we have considered a novel test statistic which is invariant under scale transformations in multivariate data and is applicable to settings with singular covariance matrices. We have proven the asymptotic distribution of the MATS as well as the asymptotic validity of the bootstrap tests. These proofs are more involved than in, e.g., Konietschke et al. (2015), since we relax the assumption of finite fourth moments to only assuming finite second moments and we do not assume positive definite covariance matrices. Furthermore, we have constructed confidence regions and simultaneous confidence intervals for contrasts of the mean vector, which provide additional insight into statistical analyses. Finally, the simulation results indicated that the MATS might

5.3 MATS: Inference for potentially singular and heteroscedastic MANOVA

be suitable in high-dimensional multivariate settings. This extension, however, requires different techniques and will be part of future research of the working group in Ulm.

6 Outlook: The fourth scenario

So far, we have considered a semi-parametric model (Friedrich et al., 2017a) as well as a nonparametric model (Friedrich et al., 2017d) for repeated measures designs. Furthermore, we analyzed semi-parametric multivariate data in Friedrich and Pauly (2017). The fourth scenario in this context thus is a nonparametric model for multivariate data. In this chapter we give a brief overview of the corresponding working paper Dobler et al. (2017).

For the d -dimensional observation vectors $\mathbf{Y}_{ik} = (Y_{i1k}, \dots, Y_{idk})'$, we consider the following model:

$$Y_{ijk} \sim F_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, d, \quad k = 1, \dots, n_i,$$

i.e., we assume arbitrary marginal distribution functions F_{ij} .

The relative effects considered in Friedrich et al. (2017d) depend on the sample sizes n_i , since they are based on weighted means of the empirical distribution functions. This means that they are no fixed model constants and changing the sample sizes may change the results, see Brunner et al. (2016) for an example. It is therefore sensible to consider the unweighted relative effects proposed by Brunner and Puri (2001). In the multivariate setting, the unweighted treatment effects for group i and dimension j are given by

$$q_{ij} = \int G_j dF_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, d.$$

Here, $G_j = \frac{1}{a} \sum_{i=1}^a F_{ij}$ denotes the unweighted mean of the distribution functions in dimension j . In comparison to the nonparametric treatment effects for the repeated measures, comparisons to the overall mean distribution $G = \frac{1}{ad} \sum_{i=1}^a \sum_{j=1}^d F_{ij}$ are not meaningful in the multivariate context, since variables may be measured on different scales. Interpretation of the effects is again rather simple: An effect $q_{ij} < 0.5$ means that the observations from component j in group i tend to smaller values than those from the reference distribution G_j .

We consider null hypotheses formulated in terms of these unweighted relative effects as $H_0^q : \mathbf{T}\mathbf{q} = \mathbf{0}$, where $\mathbf{q} = (q_{11}, \dots, q_{ad})'$. These hypotheses are less restrictive than $H_0^F : \mathbf{T}\mathbf{F} = \mathbf{0}$, since $\mathbf{T}\mathbf{F} = \mathbf{0}$ implies $\mathbf{T}\mathbf{q} = \mathbf{0}$ but not vice versa (Brunner et al., 2016). Procedures based on H_0^F have the advantage of a rather simple covariance structure of $\mathbf{T}\overline{\mathbf{R}}_{\bullet}$, see Akritas and Brunner (1997); Akritas et al. (1997) for details in the univariate case. Here, $\overline{\mathbf{R}}_{\bullet} = (\overline{\mathbf{R}}'_1, \dots, \overline{\mathbf{R}}'_a)'$ denotes the vector of rank means. Under the hypothesis $H_0^q : \mathbf{T}\mathbf{q} = \mathbf{0}$ the covariance structure is much more involved, see Puri (1964) for a derivation of the general covariance matrix of $\overline{\mathbf{R}}_{\bullet}$ in the univariate case as well as Brunner et al. (2016) for the covariance structure of $\mathbf{T}\overline{\mathbf{R}}_{\bullet}$ in a general univariate factorial design as well as for repeated measures. In our current working paper, we derive this covariance structure in the case of multivariate data.

We use empirical process theory to prove that \mathbf{q} can be consistently estimated by $\widehat{\mathbf{q}}$, where

$$\widehat{q}_{ij} = \frac{1}{a} \sum_{\ell=1}^a \widehat{w}_{\ell ij}$$

and

$$\widehat{w}_{\ell ij} = \frac{1}{n_{\ell}} \left(\overline{R}_{ij \cdot}^{(\ell i)} - \frac{n_i + 1}{2} \right) \text{ for } i, \ell = 1, \dots, a, j = 1, \dots, d.$$

Here, $R_{ijk}^{(\ell i)}$ denotes the (mid-)rank of observation Y_{ijk} in dimension j among the $(n_i + n_{\ell})$ observations in the pooled sample $Y_{\ell j 1}, \dots, Y_{\ell j n_{\ell}}, Y_{ij 1}, \dots, Y_{ij n_i}$ from treatment groups i and ℓ .

Inference is based on a wild bootstrap approach and a group-wise, nonparametric bootstrap (Efron, 1979), where we randomly draw with replacement from the observation vectors \mathbf{Y}_{ik} , resulting in the bootstrapped observation vectors \mathbf{Y}_{ik}^* , which are then used to build the bootstrapped empirical distribution functions F_{ij}^* . Finally, the bootstrapped treatment effects are obtained via

$$q_{ij}^* = \int G_j^* dF_{ij}^*.$$

The validity of this bootstrap approach again follows from arguments of empirical process theory (van der Vaart and Wellner, 1996).

A bootstrap test is finally obtained by comparing the ANOVA-type test statistic

$$T_N^q = N \widehat{\mathbf{q}}' \mathbf{T} \widehat{\mathbf{q}}$$

to the conditional $(1 - \alpha)$ -quantile of the bootstrap distribution.

In simulation studies, we analyze the type-I error control of the proposed bootstrap ap-

proaches in several settings including various continuous as well as ordinal distributions with different covariance structures and varying sample sizes. We even consider a heteroscedastic setting where H_0^q is satisfied but H_0^F is not. The wild bootstrap approach shows an accurate type-I error control across all settings. Only in the heteroscedastic case with small sample sizes, it turns out to be very conservative. This is the only scenario, where the group-wise, nonparametric bootstrap provides better results.

In summary, we consider a wild bootstrap approach to unweighted nonparametric treatment effects in a multivariate setting. These effects allow for transitive ordering and do not depend on the sample sizes. Null hypotheses are formulated in terms of the treatment effects instead of distribution functions, which allows for, e.g., derivation of confidence intervals. We are currently working on implementing these procedures in an R package called **rankMANOVA**, which will be released on CRAN soon.

7 Discussion

In this thesis, we considered different resampling procedures in order to analyze dependent data with small sample sizes. We have distinguished between a semi-parametric model, where inference is based on means, and a nonparametric model based on rank-statistics. The latter is even applicable to designs with ordinal, score or count data. Furthermore, we considered two different kinds of dependent data, namely repeated measurements, where we additionally test for underlying sub-plot factors like time, and multivariate data, where several outcomes, potentially measured on different scales, are recorded per subject.

In all situations, the small sample behavior of the state-of-the-art test statistics WTS and ATS was greatly improved by the resampling approaches. In particular, the permutation procedure of the WTS provided a valid and accurate test decision in the semi-parametric repeated measures design (Friedrich et al., 2017a). In the nonparametric case, a wild bootstrap procedure based on the ATS yielded the best results (Friedrich et al., 2017d). And in the semi-parametric MANOVA setting, we developed a new test statistic which, equipped with a parametric bootstrap routine, allowed us to develop asymptotically valid tests with accurate finite sample properties and to construct multivariate confidence regions (Friedrich and Pauly, 2017).

None of the approaches considered in this thesis relies on the assumption of multivariate normality, homoscedasticity or balanced designs. Thus, they are applicable to a wide range of factorial designs with dependent data as illustrated by the data examples considered in the articles and in Section 1. We have rigorously proven that the resampling approaches approximate the null distribution of the corresponding underlying test statistic and can thus be used for calculating data-dependent critical values. In particular, the corresponding resampling tests are asymptotically valid and provide the same local power as the original tests under contiguous alternatives.

Since the relative effects considered in Friedrich et al. (2017d) depend on the sample sizes n_i , we are currently working on a bootstrap procedure in the nonparametric mul-

tivariate situation, where we formulate hypotheses in terms of the unweighted relative effects q defined in Chapter 6. In particular, we use empirical process theory to analyze the asymptotic behavior of \hat{q} as well as its bootstrap version \hat{q}^* .

The simulation studies in Friedrich et al. (2017a) showed that the permutation procedure leads to slightly liberal (conservative) test decisions in situations with positive (negative) pairing, respectively. This is a well-known problem for permutation tests and improvement of this behavior will be part of future research. An idea is the approach considered by Smaga (2015), where a $\{2\}$ -inverse (Getson and Hsuan, 2012) is used instead of the Moore-Penrose inverse in the WTS. These tests are asymptotically valid, but only consistent for a smaller class of fixed alternatives (Smaga, 2015, 2017).

The resampling procedures for the semi-parametric setting for both multivariate data and repeated measures are implemented in the R package **MANOVA.RM** (Friedrich et al., 2017c).

The permutation procedure for the univariate case as in Pauly et al. (2015) is implemented in the R package **GFD** (Friedrich et al., 2017b). Nonparametric univariate methods are implemented in **rankFD** (Konietschke et al., 2016) and we are currently working on the corresponding extensions to the multivariate setting (package **rankMANOVA**). By providing freely available software packages, the newly derived methods are placed at the disposal of a general audience.

Bibliography

- Akritis, M. G., Arnold, S. F., and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437):258–265.
- Akritis, M. G. and Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference*, 61(2):249–277.
- Bathke, A., Friedrich, S., Konietschke, F., Pauly, M., Staffen, W., Strobl, N., and Höller, Y. (2016). Using EEG, SPECT, and multivariate resampling methods to differentiate between Alzheimer’s and other cognitive impairments. *arXiv preprint arXiv:1606.09004*.
- Beyersmann, J., Termini, S. D., and Pauly, M. (2013). Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 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. *The Annals of Mathematical Statistics*, 25(2):290–302.
- Brown, B. M. and Hettmansperger, T. P. (2002). Kruskal–Wallis, multiple comparisons and Efron dice. *Australian & New Zealand Journal of Statistics*, 44(4):427–438.
- Brückner, M. and Brannath, W. (2016). Sequential tests for non-proportional hazards data. *Lifetime Data Analysis*, 23(3):1–14.
- Brumback, L. C., Pepe, M. S., and Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25(4):575–590.
- Brunner, E. (2001). Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity. *Mathematical Statistics with Applications in Biometry*.

- Brunner, E., Dette, H., and Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, 92(440):1494–1502.
- Brunner, E., Domhof, S., and Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. Wiley, New York, USA.
- Brunner, E., Konietschke, F., Pauly, M., and Puri, M. L. (2016). Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Brunner, E., Munzel, U., and Puri, M. L. (1999). Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis*, 70(2):286–317.
- Brunner, E. and Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers*, 42(1):1–52.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.
- Cameron, A. C. and Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.
- Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.
- Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.
- Davison, A. C., Hinkley, D. V., and Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science*, 18(2):141–157.
- De Neve, J., Meys, J., Ottoy, J.-P., Clement, L., and Thas, O. (2014). unified-WMWqPCR: the unified Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data in R. *Bioinformatics*, 30(17):2494–2495.
- Dobler, D., Friedrich, S., and Pauly, M. (2017). Nonparametric MANOVA in Mann-Whitney effects. In preparation.

- Dobler, D. and Pauly, M. (2014). Bootstrapping Aalen-Johansen processes for competing risks: Handicaps, solutions, and limitations. *Electronic Journal of Statistics*, 8(2):2779–2803.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Fischer, D., Oja, H., Schleutker, J., Sen, P. K., and Wahlfors, T. (2014). Generalized Mann-Whitney type tests for microarray experiments. *Scandinavian Journal of Statistics*, 41(3):672–692.
- Friedrich, S., Brunner, E., and Pauly, M. (2017a). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153:255–265.
- Friedrich, S., Konietschke, F., and Pauly, M. (2017b). GFD: An R package for the analysis of general factorial designs. *Journal of Statistical Software, Code Snippets*, 79(1):1–18.
- Friedrich, S., Konietschke, F., and Pauly, M. (2017c). *MANOVA.RM: Analysis of Multivariate Data and Repeated Measures Designs*. R package version 0.1.1.
- Friedrich, S., Konietschke, F., and Pauly, M. (2017d). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, 113:38–52.
- Friedrich, S. and Pauly, M. (2017). MATS: Inference for potentially singular and heteroscedastic MANOVA. *arXiv preprint arXiv:1704.03731*.
- Gardner, M. (1970). Paradox of nontransitive dice and elusive principle of indifference. *Scientific American*, 223(6):110.
- Getson, A. J. and Hsuan, F. C. (2012). *{2}-inverses and their statistical application*. Springer Science & Business Media.
- Good, P. I. (2006). *Permutation, parametric, and bootstrap tests of hypotheses*. Springer Science & Business Media.
- Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310.

- Hjorth, B. (1975). Time domain descriptors and their relation to a particular model for generation of EEG activity. In Dolce, G. and Kunkel, H., editors, *CEAN Computerized EEG Analysis*, pages 3–8. Gustav Fischer.
- Hotelling, H. (1931). A generalization of Student's ratio. *The Annals of Mathematical Statistics*, 2:360–378.
- Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem. *Statistics & Probability Letters*, 36(1):9–21.
- Janssen, A. (2005). Resampling Student's t -type statistics. *Annals of the Institute of Statistical Mathematics*, 57(3):507–529.
- Jentsch, C. and Pauly, M. (2015). Testing equality of spectral densities using randomization techniques. *Bernoulli*, 21(2):697–739.
- Konietschke, F., Bathke, A., Harrar, S., and Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, 140:291–301.
- Konietschke, F., Friedrich, S., Brunner, E., and Pauly, M. (2016). *rankFD: Rank-Based Tests for General Factorial Designs*. R package version 0.0.1.
- Kreiss, J.-P. and Paparoditis, E. (2011). Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, 40(4):357–378.
- Lin, D. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16(8):901–910.
- Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708.
- Lumley, T. (1996). Generalized estimating equations for ordinal data: a note on working correlation structures. *Biometrics*, 52(1):354–361.
- Mammen, E. (1993a). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):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*. CRC Press.

- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.
- Martinussen, T. and Scheike, T. H. (2007). *Dynamic regression models for survival data*. Springer Science & Business Media.
- Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics*, 5:41–52.
- Pauly, M., Brunner, E., and Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):461–473.
- Pop-Jordanova, N. and Pop-Jordanova, J. (2005). Spectrum-weighted EEG frequency (“brainrate”) as a quantitative indicator of mental arousal. *Prilozi*, 26(2):35–42.
- Puri, M. L. (1964). Asymptotic efficiency of a class of c-sample tests. *The Annals of Mathematical Statistics*, pages 102–121.
- Rauch, G., Jahn-Eimermacher, A., Brannath, W., and Kieser, M. (2014). Opportunities and challenges of combined effect measures based on prioritized outcomes. *Statistics in Medicine*, 33(7):1104–1120.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- SAS Institute Inc. (2003). *SAS Software, Version 9.1*. Cary, NC.
- Smaga, Ł. (2015). Wald-type statistics using $\{2\}$ -inverses for hypothesis testing in general factorial designs. *Statistics & Probability Letters*, 107:215–220.
- Smaga, Ł. (2017). Bootstrap methods for multivariate hypothesis testing. *Communications in Statistics-Simulation and Computation*, pages 1–14.
- Srivastava, M. S. and Kubokawa, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis*, 115:204–216.
- Suo, C., Touloupoulou, T., Bramon, E., Walshe, M., Picchioni, M., Murray, R., and Ott, J. (2013). Analysis of multiple phenotypes in genome-wide genetic mapping studies. *BMC Bioinformatics*, 14(1):151.

Bibliography

- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.
- Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24(3-4):471–494.
- Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295.
- Xu, J. and Cui, X. (2008). Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. *Bioinformatics*, 24(8):1056–1062.

Part II

Publications

Article 1

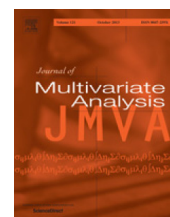
Friedrich, S., Brunner, E. and Pauly, M. (2017). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, **153**, 255–265, DOI: 10.1016/j.jmva.2016.10.004.

Reprinted from Journal of Multivariate Analysis, Vol. 153, S. Friedrich, E. Brunner and M. Pauly, Permuting longitudinal data in spite of the dependencies, pages 255–265, Copyright (2017), with permission from Elsevier.



Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

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

ARTICLE INFO

Article history:

Received 22 June 2016

Available online 25 October 2016

AMS subject classifications:

primary 62G10

secondary 62G09

62H15

62P10

Keywords:

Permutation tests

Longitudinal data

Quadratic forms

Repeated measures

ABSTRACT

For general repeated measures designs the Wald-type statistic (WTS) is an asymptotically valid procedure allowing for unequal covariance matrices and possibly non-normal multivariate observations. The drawback of this procedure is its poor performance for small to moderate samples, i.e., decisions based on the WTS may become quite liberal. It is the aim of the present paper to improve the small-sample behavior of the WTS by means of a novel permutation procedure. In particular, it is shown that a permutation version of the WTS inherits its good large-sample properties while yielding a very accurate finite-sample control of the type-I error as shown in extensive simulations. Moreover, the new permutation method is motivated by a practical data set of a split plot design with a factorial structure on the repeated measures.

© 2016 Elsevier Inc. All rights reserved.

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 Konietzschke 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 $\mathbf{V}_i = \text{cov}(\mathbf{Y}_{i1}) > 0$ 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

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

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 $\mu_{is} = E(Y_{i1s}), i \in \{1, \dots, a\}, s \in \{1, \dots, t_i\}$ 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:

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

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

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

by an $\mathcal{F}(\hat{\nu}, \infty)$ -distribution with estimated degree of freedom $\hat{\nu} = \text{tr}^2(\hat{\mathbf{T}}\hat{\boldsymbol{\Sigma}})/\text{tr}(\hat{\mathbf{T}}\hat{\boldsymbol{\Sigma}})^2$, where $\hat{\boldsymbol{\Sigma}} = N \bigoplus_{i=1}^a 1/n_i \hat{\mathbf{V}}_i$. The corresponding ATS test $\varphi_{ATS} = \mathbf{1}\{\tilde{Q}_N > \mathcal{F}_\alpha(\hat{\nu}, \infty)\}$, 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^2_{t-1} -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

$$Q_N = N\bar{\mathbf{Y}} \cdot \mathbf{H}^\top (\mathbf{H}\hat{\Sigma}\mathbf{H}^\top)^+ \mathbf{H}\bar{\mathbf{Y}} \cdot \tag{2.5}$$

Here $(\mathbf{H}\hat{\Sigma}\mathbf{H}^\top)^+$ denotes the Moore–Penrose inverse of $(\mathbf{H}\hat{\Sigma}\mathbf{H}^\top)$. 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.,

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

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 $\bar{\mathbf{Y}}^\pi$ denote the vector of the means under this permutation and $\hat{\Sigma}^\pi = \bigoplus_{i=1}^a N/n_i \hat{\mathbf{V}}_i^\pi$ 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

$$Q_N^\pi = N(\bar{\mathbf{Y}}_\bullet^\pi)^\top \mathbf{H}^\top (\mathbf{H}\hat{\Sigma}^\pi \mathbf{H}^\top)^+ \mathbf{H}\bar{\mathbf{Y}}_\bullet^\pi. \tag{3.1}$$

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., $\mathbf{H}\mathbf{E}(\bar{\mathbf{Y}}_\bullet^\pi) = \mathbf{0}$ 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.

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

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., $\mathbb{E}_{\mathcal{H}} \mathbf{1}(\varphi_{WTPS}) \rightarrow \Pr(Z > \chi_{f,1-\alpha}^2)$ 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 $\boldsymbol{\mu}_i = E(\mathbf{Y}_{i1})$ 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

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

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*

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

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: $\mathbf{V}_i = \left(\rho_i^{|\ell-j|} \right)_{\ell, j \leq t}$, $(\rho_1, \rho_2, \rho_3) = (0.6, 0.5, 0.4)$ 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.

<i>T</i>		<i>t</i> = 4			<i>t</i> = 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^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

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

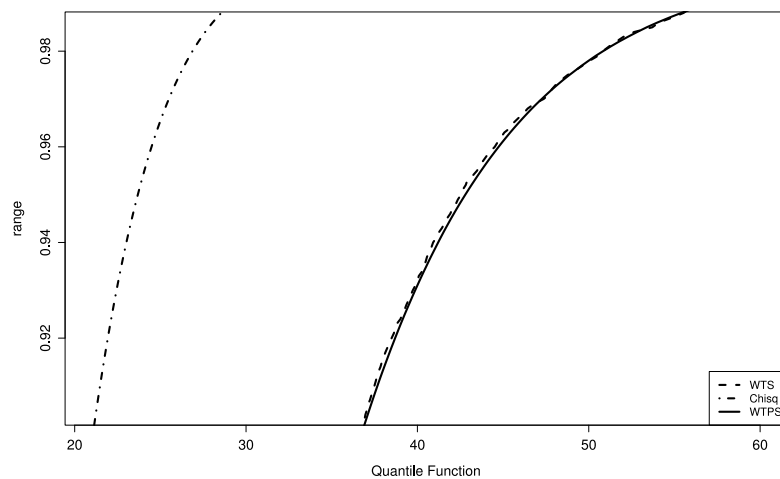
in order to compare the distance between the quantile function F_N^{-1} and the limiting quantile function F^{-1} (KQS) with the distance between F_N^{-1} and $(F_N^\pi)^{-1}$, the quantile functions of the test statistic and its permuted version (KQS^π), respectively. We have calculated these distances 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 4Results of the simulation studies for the hypothesis of no group \times time interaction.

CT	Cov. setting	$t = 4$			$t = 8$		
		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 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

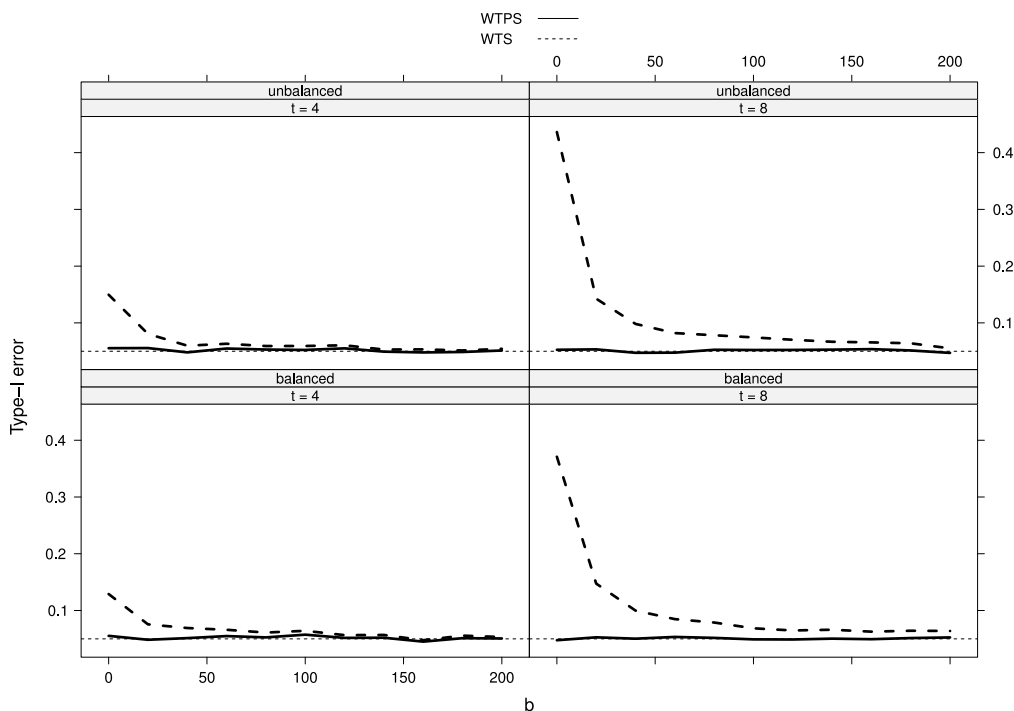


Fig. 2. Type-I error rates under the interaction hypothesis for the WTS and the WTPS, where sample size was increased by adding $b\mathbf{1}_3$, $b = \ell 20$ for all $\ell \in \{0, \dots, 10\}$ to the sample size vectors in a balanced (lower panel) and unbalanced (upper panel) design with $t = 4$ (left 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 $\mathcal{F}(\hat{v}, \infty)$ -quantile, the WTS in (2.5) to the asymptotic χ^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 Konietzschke 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. Midena, 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. Konietzschke, 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 ϵ 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.

Supplement to Permuting Longitudinal Data in spite of the Dependencies

Sarah Friedrich*, Edgar Brunner** and Markus Pauly*

October 28, 2016

Abstract

In this supplementary material to the authors' paper 'Permuting longitudinal data in spite of the dependencies' we present all technical details together with additional simulation results and consider further resampling techniques, namely a nonparametric and a parametric bootstrap approach. Furthermore, we analyze the power of the WTPS and compare it to the power of the ATS. Finally, we present more details about the the data example on O_2 -consumption of leukocytes and also present an analysis, where we applied all considered approaches.

* Ulm University, Institute of Statistics, Germany
email: sarah.friedrich@uni-ulm.de

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

8 Proofs

Proof of Theorem 1: First note that $\mathbf{T} = \mathbf{T}^\top$ as well as $\mathbf{T}^2 = \mathbf{T}$. Let $\mathbf{T}\boldsymbol{\mu} = \mathbf{T}\boldsymbol{\nu}/\sqrt{N}$ for $\boldsymbol{\nu} \in \mathbb{R}^T$, i.e., for $\boldsymbol{\nu} = \mathbf{0}$ we are working under \mathcal{H}_0 . It holds that $\sqrt{N}(\bar{\mathbf{Y}}_\bullet - \boldsymbol{\mu})$ has, asymptotically, a multivariate normal distribution with mean $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\Sigma}$. Thus, it follows that

$$N\bar{\mathbf{Y}}_\bullet^\top \mathbf{T} \bar{\mathbf{Y}}_\bullet \rightarrow \mathbf{Z}^\top \mathbf{T} \mathbf{Z},$$

with $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$. If additionally $\boldsymbol{\Sigma} > 0$, we may write $\mathbf{Z} = \boldsymbol{\Sigma}^{1/2} \tilde{\mathbf{Z}}$ where $\tilde{\mathbf{Z}} \sim \mathcal{N}(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\nu}, \mathbf{I})$ and thus $\mathbf{Z}^\top \mathbf{T} \mathbf{Z} = \sum_{i=1}^a \sum_{s=1}^{t_i} \lambda_{is} X_{is}$ where λ_{is} are the eigenvalues of $\mathbf{T}\boldsymbol{\Sigma}$ and $X_{is} \sim \chi_1^2(\delta)$ for $\delta = \boldsymbol{\nu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\nu}$. \square

Proof of Theorem 2: The null distribution of the WTS follows analogous to the proof of Theorem 2.1 in [3]. Obviously, φ_{WTS} is an asymptotic level α test and consistent for fixed alternatives $\mathbf{H}\boldsymbol{\mu} \neq \mathbf{0}$.

Under $\mathcal{H}_1 : \mathbf{H}\boldsymbol{\mu} = 1/\sqrt{N} \cdot \boldsymbol{\nu}$, it holds that $\sqrt{N}\mathbf{H}\bar{\mathbf{Y}}_\bullet$ has, asymptotically, an $\mathcal{N}(\mathbf{H}\boldsymbol{\nu}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^\top)$ distribution. Thus, the WTS has asymptotically a non-central $\chi_f^2(\tilde{\delta})$ distribution with $f = \text{rank}(\mathbf{H})$ degrees of freedom and non-centrality parameter $\tilde{\delta} = (\mathbf{H}\boldsymbol{\nu})^\top (\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^\top)^+ \mathbf{H}\boldsymbol{\nu}$. \square

We will now proof Theorem 3. For notational convenience, we introduce

$$\mathbf{Z} = (Z_{N,1}, \dots, Z_{N,\tilde{N}}) = (Y_{111}, Y_{121}, \dots, Y_{1n_11}, Y_{112}, \dots, Y_{an_a t_a})$$

for the pooled sample. Since $\mathbf{H}\mathbf{1} = \mathbf{0}$ we can rewrite the permuted test statistic as

$$Q_N^\pi = \sqrt{N}(\bar{\mathbf{Y}}_\bullet^\pi - \bar{\mathbf{Y}}_{\dots})^\top \mathbf{H}^\top (\mathbf{H}\hat{\boldsymbol{\Sigma}}^\pi \mathbf{H}^\top)^+ \sqrt{N}\mathbf{H}(\bar{\mathbf{Y}}_\bullet^\pi - \bar{\mathbf{Y}}_{\dots}),$$

where $\bar{\mathbf{Y}}_{\dots} = \bar{Y}_{\dots} \cdot \mathbf{1}_T$ and $\bar{Z}_{\tilde{N}} = \bar{Y}_{\dots} = 1/\tilde{N} \cdot \sum_{i=1}^{\tilde{N}} Z_{N,i}$. Based on this representation, we can split the proof of Theorem 3 in two results. There, we first show that the conditional distribution of $\sqrt{N}(\bar{\mathbf{Y}}_\bullet^\pi - \bar{\mathbf{Y}}_{\dots})$ given the data is asymptotically multivariate normal. However, it turns out that the resulting covariance matrix is different from $\boldsymbol{\Sigma}$. Our approach corrects for the 'wrong' covariance structure by studentizing with $\hat{\boldsymbol{\Sigma}}^\pi$, which is shown in a second step. Altogether, this proves the consistency of the WTPS as stated in Theorem 3 as well as the properties of the corresponding test mentioned in Remarks 3.1 and 3.2.

Note that there exist finite limits $b_i = \lim_{\min(n_i) \rightarrow \infty} \tilde{N}/n_i \in (1, \infty)$, $i \in \{1, \dots, a\}$ because of assumption (1) and $0 < \max_{i=1, \dots, a} (t_i) < \infty$.

LEMMA 8.1 *Under the assumptions of Theorem 3, the conditional permutation distribution of*

$$\sqrt{N}(\bar{\mathbf{Y}}_\bullet^\pi - \bar{\mathbf{Y}}_{\dots})$$

given the observed data \mathbf{Y} weakly converges to a multivariate normal — $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{\Gamma})$ — distribution in probability, where

$$\sigma^2 = \sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} (\sigma_{is}^2 + \mu_{is}^2) - \left(\sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is} \right)^2 \quad (8.1)$$

with $\sigma_{is}^2 = \text{var}(Y_{iks})$ and

$$\mathbf{\Gamma} = \bigoplus_{i=1}^a \kappa_i^{-1} \mathbf{I}_{t_i} - \mathbf{J}_T = \text{diag}(\kappa_1^{-1} \mathbf{I}_{t_1}, \dots, \kappa_a^{-1} \mathbf{I}_{t_a}) - \mathbf{J}_T. \quad (8.2)$$

Proof: First note that the classical Cramér-Wold device cannot be applied directly in this context due to the occurrence of uncountably many exceptional sets. Therefore we will apply a modified Cramér-Wold device, see, e.g., the proof of Theorem 4.1 in [5]. Let D be a dense and countable subset of \mathbb{R}^T . Then for every fixed $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_T)^\top \in D$ and $M_0 = 0, M_1 = n_1, \dots, M_{t_1} = t_1 n_1, M_{t_1+1} = t_1 n_1 + n_2, \dots, M_T = \tilde{N}$, we have

$$\begin{aligned} \sqrt{\tilde{N}} \boldsymbol{\lambda}^\top \bar{\mathbf{Y}}_{\bullet} &= \sum_{i=1}^a \sum_{k=M_{i-1}+1}^{M_i} \frac{\sqrt{\tilde{N}}}{n_i} \lambda_i Z_{N,k} \\ &= \sqrt{\tilde{N}} \sum_{s=1}^{\tilde{N}} c_{Ns} \frac{Z_{N,s}}{\sqrt{\tilde{N}}}, \end{aligned}$$

where $c_{Ns} = \sqrt{\tilde{N}} \sum_{i=1}^T \mathbf{1}\{M_{i-1} + 1 \leq s \leq M_i\} \lambda_i / n_i$. This implies

$$\begin{aligned} \sqrt{\tilde{N}} \boldsymbol{\lambda}^\top (\bar{\mathbf{Y}}_{\bullet}^{\pi} - \bar{\mathbf{Y}}_{\dots}) &= \sqrt{\tilde{N}} \sum_{s=1}^{\tilde{N}} c_{Ns} \frac{Z_{N,\pi(s)} - \bar{Z}_{\tilde{N}}}{\sqrt{\tilde{N}}} \\ &\stackrel{d}{=} \sqrt{\tilde{N}} \sum_{s=1}^{\tilde{N}} c_{N\pi(s)} \frac{Z_{N,s} - \bar{Z}_{\tilde{N}}}{\sqrt{\tilde{N}}}, \end{aligned} \quad (8.3)$$

since π is uniformly distributed on the set of all permutations of the numbers $\{1, \dots, \tilde{N}\}$. Let $b_i = \lim_{\min(n_i) \rightarrow \infty} \tilde{N}/n_i$ with $b_i < \infty$ because of (1) and $\max(t_i) < \infty$.

We now apply Theorem 4.1 in [5] to prove the conditional convergence in distribution. Therefore, we have to prove the following conditions:

$$\frac{1}{\sqrt{\tilde{N}}} \max_{1 \leq i \leq \tilde{N}} |Z_{N,i} - \bar{Z}_{\tilde{N}}| \xrightarrow{\text{Pr}} 0 \quad (8.4)$$

$$\frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} (Z_{N,i} - \bar{Z}_{\tilde{N}})^2 \xrightarrow{\text{Pr}} \sigma^2 \quad (8.5)$$

$$\max_{1 \leq s \leq \tilde{N}} |c_{N,s} - \bar{c}| \xrightarrow{\text{Pr}} 0 \quad (8.6)$$

$$\sum_{s=1}^{\tilde{N}} (c_{N,s} - \bar{c})^2 \xrightarrow{\text{Pr}} \sigma_\lambda^2 = \sum_{i=1}^T \lambda_i^2 b_i - \left(\sum_{i=1}^T \lambda_i \right)^2 \quad (8.7)$$

$$\sqrt{\tilde{N}}(c_{N,\pi(1)} - \bar{c}) \xrightarrow{d} W \text{ with } E(W) = 0 \text{ and } \text{var}(W) = \sigma_\lambda^2 \quad (8.8)$$

Condition (8.4) as well as (8.6) – (8.8) follow analogous to [6]: Since the random variables within each of the a groups are i.i.d. with finite variance, they fulfill (8.4). The convergence in (8.6) is obvious and since $\sqrt{\tilde{N}} \cdot \bar{c} = \sum_{i=1}^T \lambda_i$ we have

$$\begin{aligned} \sum_{s=1}^{\tilde{N}} (c_{N,s} - \bar{c})^2 &= \sum_{s=1}^{\tilde{N}} c_{N,s}^2 - (\sqrt{\tilde{N}}\bar{c})^2 = \sum_{i=1}^T \frac{\tilde{N}}{n_i} \lambda_i^2 - \left(\sum_{i=1}^T \lambda_i \right)^2 \\ &\xrightarrow{\text{Pr}} \sum_{i=1}^T \lambda_i^2 b_i - \left(\sum_{i=1}^T \lambda_i \right)^2 = \sigma_\lambda^2. \end{aligned}$$

Moreover, (8.8) holds due to

$$\Pr \left(\sqrt{\tilde{N}} c_{N,\pi(1)} = \frac{\tilde{N} \lambda_i}{n_i} \right) = \frac{n_i}{\tilde{N}} \longrightarrow \frac{1}{b_i}$$

for $i \in \{1, \dots, a\}$, i.e., for a random variable \widetilde{W} with $\Pr(\widetilde{W} = b_i \lambda_i) = 1/b_i$, $i \in \{1, \dots, a\}$, we have

$$\sqrt{\tilde{N}}(c_{N,\pi(1)} - \bar{c}) \xrightarrow{d} \widetilde{W} - \sum_{i=1}^T \lambda_i = W,$$

where W fulfills $E(W) = 0$ and $\text{var}(W) = \sigma_\lambda^2$. It remains to prove (8.5):

$$\frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} (Z_{N,i} - \bar{Z}_{\tilde{N}})^2 = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} Z_{N,i}^2 - \bar{Y}^2.$$

Consider

$$\begin{aligned}
\mathbb{E} \left(\frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} Z_{N,i}^2 \right) &= \frac{1}{\tilde{N}} \sum_{i=1}^a \sum_{s=1}^{t_i} \sum_{k=1}^{n_i} \mathbb{E} (Y_{iks}^2) \\
&= \frac{1}{\tilde{N}} \sum_{i=1}^a \sum_{s=1}^{t_i} \sum_{k=1}^{n_i} (\sigma_{is}^2 + \mu_{is}^2) \\
&= \sum_{i=1}^a \frac{n_i}{\tilde{N}} \sum_{s=1}^{t_i} (\sigma_{is}^2 + \mu_{is}^2) \\
&\rightarrow \sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} (\sigma_{is}^2 + \mu_{is}^2).
\end{aligned}$$

Furthermore:

$$\begin{aligned}
\mathbb{E}(\bar{Y}^2) &= \mathbb{E} \left\{ \left(\frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} Z_{N,i} \right)^2 \right\} \\
&= \underbrace{\text{var} \left(\frac{1}{\tilde{N}} \sum_{i=1}^a \sum_{s=1}^{t_i} \sum_{k=1}^{n_i} Y_{iks} \right)}_{\mathcal{O}(\frac{1}{\tilde{N}})} + \left\{ \mathbb{E} \left(\frac{1}{\tilde{N}} \sum_{i=1}^a \sum_{s=1}^{t_i} \sum_{k=1}^{n_i} Y_{iks} \right) \right\}^2 \quad (8.9) \\
&\rightarrow \left(\sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is} \right)^2
\end{aligned}$$

Since

$$\begin{aligned}
\text{var} \left\{ \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} (Z_{N,i} - \bar{Z}_{\tilde{N}})^2 \right\} &= \text{var} \left(\sum_{i=1}^a \frac{1}{\tilde{N}} \sum_{s=1}^{t_i} \sum_{k=1}^{n_i} Y_{iks}^2 - \bar{Y}^2 \right) \\
&= \sum_{i=1}^a \frac{1}{(\tilde{N})^2} \sum_{k=1}^{n_i} \text{var} \left\{ \sum_{s=1}^{t_i} (Y_{iks}^2 - \frac{1}{\tilde{N}} \bar{Y}^2) \right\} \\
&= \mathcal{O} \left(\frac{1}{\tilde{N}} \right)
\end{aligned}$$

because of independence and condition (2), the desired conclusion follows with Tschebyscheff's inequality.

Altogether, this implies by Theorem 4.1 in [5] convergence in distribution given the data \mathbf{Y}

$$\sqrt{\tilde{N}} \boldsymbol{\lambda}^\top (\bar{\mathbf{Y}}^\pi - \bar{\mathbf{Y}}) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \sigma_\lambda^2) \quad (8.10)$$

in probability. This convergence holds for every fixed $\lambda \in D$. Applying the subsequential principle for convergence in probability we can find a common subsequence such that (8.10) holds almost surely for all $\lambda \in D$ along this subsequence. Now continuity of the characteristic function of the limit and tightness of the conditional distribution of $\sqrt{N}(\bar{\mathbf{Y}}_{\bullet}^{\pi} - \bar{\mathbf{Y}} \dots)$ given \mathbf{Y} show that (8.10) holds almost surely for all $\lambda \in \mathbb{R}^T$ along this subsequence. Thus, an application of the classical Cramér-Wold device together with another application of the subsequence principle imply the result. \square

Now we will study the convergence of $\widehat{\Sigma}^{\pi} = N\widehat{\mathbf{V}}^{\pi} = \bigoplus_{i=1}^a N/n_i \cdot \widehat{\mathbf{V}}_i^{\pi}$.

LEMMA 8.2 *Under the assumptions of Theorem 3 we have convergence in probability*

$$\widehat{\Sigma}^{\pi} \xrightarrow{\text{Pr}} \sigma^2 \text{diag}(\kappa_1^{-1} \mathbf{I}_{t_1}, \dots, \kappa_a^{-1} \mathbf{I}_{t_a})$$

as $N \rightarrow \infty$.

Proof: It suffices to show that $(\widehat{\mathbf{V}}_i^{\pi})_{r,s} \xrightarrow{\text{Pr}} \sigma^2 \mathbf{1}\{r = s\}$ in probability for all $1 \leq r, s \leq t_i$. Therefore consider

$$\frac{n_i - 1}{n_i} (\widehat{\mathbf{V}}_i^{\pi})_{r,s} = \frac{1}{n_i} \underbrace{\sum_{k=1}^{n_i} Y_{ikr}^{\pi} Y_{iks}^{\pi}}_A - \underbrace{\bar{Y}_{i \cdot r}^{\pi} \bar{Y}_{i \cdot s}^{\pi}}_B$$

First, consider B . It holds:

$$\mathbb{E}(\bar{Y}_{i \cdot r}^{\pi} | \mathbf{Y}) \xrightarrow{\text{Pr}} \sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is}$$

for all r and all i analogous to (8.9). Furthermore, setting $d_{N,s}^{(i)} := \mathbf{1}\{(r-1)n_i + 1 \leq s \leq rn_i\}/n_i$ for $1 \leq s \leq \tilde{N}$ and using Theorem 3 from [1] we get convergence in probability of the corresponding conditional variance

$$\begin{aligned} \text{var}(\bar{Y}_{i \cdot r}^{\pi} | \mathbf{Y}) &= \text{var} \left(\sum_{s=1}^{\tilde{N}} d_{N,s}^{(i)} Z_{N,\pi(s)} | \mathbf{Z} \right) \\ &= \sum_{s=1}^{\tilde{N}} \left(d_{N,s}^{(i)} - \bar{d}_{N,\cdot}^{(i)} \right)^2 \frac{1}{\tilde{N} - 1} \sum_{s=1}^{\tilde{N}} (Z_{N,s} - \bar{Z}_{\tilde{N}})^2 \xrightarrow{\text{Pr}} 0, \end{aligned}$$

since $\sum_{s=1}^{\tilde{N}} \left(d_{N,s}^{(i)} - \bar{d}_{N,\cdot}^{(i)} \right)^2 \rightarrow 0$ as $N \rightarrow \infty$ and $1/(\tilde{N} - 1) \cdot \sum_{s=1}^{\tilde{N}} (Z_{N,s} - \bar{Z}_{\tilde{N}})^2 = \mathcal{O}_{\text{Pr}}(1)$. Altogether this implies convergence in probability by the continuous mapping theorem

$$B \xrightarrow{\text{Pr}} \left(\sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is} \right)^2.$$

For part A we distinguish two cases: First, assume $r = s$. We have

$$A = \frac{1}{n_i} \sum_{k=(r-1)n_i+1}^{rn_i} Z_{N,\pi(k)}^2.$$

Now consider the conditional expectation of A

$$\begin{aligned} \mathbb{E}(A|\mathbf{Y}) &= \frac{1}{\tilde{N}} \sum_{i=1}^a \sum_{r=1}^{t_i} \frac{1}{n_i} \sum_{k=(r-1)n_i+1}^{rn_i} Z_{N,k}^2 \\ &\xrightarrow{\text{Pr}} \sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} (\sigma_{is}^2 + \mu_{is}^2) \end{aligned}$$

as well as

$$\begin{aligned} \text{var}(A|\mathbf{Y}) &= \text{var} \left(\sum_{s=1}^{\tilde{N}} d_{N,s}^{(i)} Z_{N,\pi(s)}^2 | \mathbf{Z} \right) \\ &= \sum_{s=1}^{\tilde{N}} \left(d_{N,s}^{(i)} - \bar{d}_{N,\cdot}^{(i)} \right)^2 \frac{1}{\tilde{N}-1} \sum_{s=1}^{\tilde{N}} \left(Z_{N,s}^2 - \frac{1}{\tilde{N}-1} \sum_{r=1}^{\tilde{N}} Z_{N,r}^2 \right)^2, \end{aligned}$$

which converges to 0 in probability as above and since we have

$$\frac{1}{\tilde{N}-1} \sum_{s=1}^{\tilde{N}} \left(Z_{N,s}^2 - \frac{1}{\tilde{N}-1} \sum_{r=1}^{\tilde{N}} Z_{N,r}^2 \right)^2 = \mathcal{O}_{\text{Pr}}(1)$$

because of the existence of fourth moments.

Now, consider $r \neq s$. We have that

$$\begin{aligned} \mathbb{E}(A|\mathbf{Y}) &= \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbb{E}(Y_{ikr}^\pi Y_{iks}^\pi | \mathbf{Y}) = \mathbb{E}(Y_{111}^\pi Y_{112}^\pi | \mathbf{Y}) \\ &= \frac{1}{(\tilde{N})!} \sum_{\pi \in \mathcal{S}_{\tilde{N}}} Z_{N,\pi(1)} Z_{N,\pi(2)} \\ &= \frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i \neq j} Z_{N,i} Z_{N,j} \end{aligned}$$

Consider $\mathbb{E}(Z_{N,i} Z_{N,j})$. There are two possibilities: If $Z_{N,i}$ and $Z_{N,j}$ stem from different random vectors (i.e., from different individuals) they are independent and we can write $\mathbb{E}(Z_{N,i} Z_{N,j}) = \mathbb{E}(Z_{N,i}) \mathbb{E}(Z_{N,j})$. If they stem from the same individual, we cannot rewrite the expectation and

we denote it by $\gamma_{i,j} := E(Z_{N,i}Z_{N,j}) \in (-\infty, \infty)$. For every fixed i there are $(t_i - 1)$ possible j 's such that $Z_{N,i}$ and $Z_{N,j}$ come from the same individual. This implies:

$$\begin{aligned} E\left(\frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i \neq j} Z_{N,i}Z_{N,j}\right) &= \frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{(i,j) \in \Xi} E(Z_{N,i})E(Z_{N,j}) \quad (8.11) \\ &+ \frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i=1}^{\tilde{N}} \sum_{(i,j) \in \Lambda} \gamma_{i,j}, \end{aligned}$$

where the index sets are defined as $\Xi = \{(i, j) : i \neq j \text{ and } i, j \text{ stem from different subjects}\}$ and $\Lambda = \{(i, j) : i \neq j \text{ and } i, j \text{ stem from the same subject}\}$.

Because of the Cauchy-Schwarz inequality and Condition (2) it holds that

$$\sup_{i,j} |\gamma_{i,j}| \leq 2 \sup_i E(Z_{N,i}^2) \leq C < \infty.$$

Thus, it follows that

$$\frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i=1}^{\tilde{N}} \sum_{(i,j) \in \Lambda} \gamma_{i,j} \leq \frac{\tilde{N}(\max t_i - 1)}{\tilde{N}(\tilde{N}-1)} C \rightarrow 0$$

as $N \rightarrow \infty$. For the first summand on the right hand side in Equation (8.11), it holds that

$$\begin{aligned} \frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{(i,j) \in \Xi} E(Z_{N,i})E(Z_{N,j}) &= \frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i=1}^a \sum_{s=1}^{t_i} \sum_{j=1}^a \sum_{r=1}^{t_j} \mu_{is} \mu_{jr} - o(1) \\ &\rightarrow \left(\sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is} \right)^2. \end{aligned}$$

To complete the proof it remains to show that $\text{var}\left(\frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i \neq j} Z_{N,i}Z_{N,j}\right) \rightarrow 0$. Thus,

$$\text{var}\left(\frac{1}{\tilde{N}(\tilde{N}-1)} \sum_{i \neq j} Z_{N,i}Z_{N,j}\right) = \frac{1}{\{\tilde{N}(\tilde{N}-1)\}^2} \sum_{i_1 \neq j_1} \sum_{i_2 \neq j_2} \text{cov}(Z_{N,i_1}Z_{N,j_1}, Z_{N,i_2}Z_{N,j_2}).$$

As above, we distinguish between the cases $(i, j) \in \Xi$ and $(i, j) \in \Lambda$. If $Z_{N,i_1}Z_{N,j_1}$ and $Z_{N,i_2}Z_{N,j_2}$ stem from different individuals it holds that $\text{cov}(Z_{N,i_1}Z_{N,j_1}, Z_{N,i_2}Z_{N,j_2}) = 0$ because of independence. In all other cases it holds that

$$\text{cov}(Z_{N,i_1}Z_{N,j_1}, Z_{N,i_2}Z_{N,j_2}) \leq 2 \sup_i E(Z_{N,i}^4) = \tilde{C} < \infty$$

because of assumption (2) and the Cauchy-Schwarz inequality.

Furthermore, for every fixed i_1 and j_1 there are less than $5(\max t_i)^4$ possibilities for $Z_{N,i_2}Z_{N,j_2}$

to stem from the same individual(s) as $Z_{N,i_1}Z_{N,j_1}$, such that at least one of the sums cancels out and $\text{var}\left(\frac{1}{\{\tilde{N}(\tilde{N}-1)\}} \cdot \sum_{i \neq j} Z_{N,i}Z_{N,j}\right) \rightarrow 0$ for all i, j as $N \rightarrow \infty$.

This implies that for $r \neq s$

$$\frac{n_i - 1}{n_i} \left(\widehat{\mathbf{V}}_i^\pi\right)_{r,s} = A - B \xrightarrow{\text{Pr}} \left(\sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is}\right)^2 - \left(\sum_{i=1}^a \frac{1}{b_i} \sum_{s=1}^{t_i} \mu_{is}\right)^2 = 0$$

and for $r = s$ we have $\left(\widehat{\mathbf{V}}_i^\pi\right)_{r,s} \xrightarrow{\text{Pr}} \sigma^2$. Altogether, this proves the desired result. \square

We are now able to prove Theorem 3.

Proof: Applying the continuous mapping theorem together with Lemma 8.1 yields conditional convergence in distribution given \mathbf{Y}

$$\mathbf{H}\sqrt{N}(\overline{\mathbf{Y}}^\pi - \overline{\mathbf{Y}} \dots) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathbf{H} \mathbf{D} \mathbf{H}^\top),$$

where $\mathbf{D} := \text{diag}(\kappa_1^{-1} \mathbf{I}_{t_1}, \dots, \kappa_a^{-1} \mathbf{I}_{t_a})$. Moreover, we have convergence in probability

$$\mathbf{H} \widehat{\Sigma}^\pi \mathbf{H}^\top \xrightarrow{\text{Pr}} \sigma^2 \mathbf{H} \mathbf{D} \mathbf{H}^\top$$

by Lemma 8.2. Since $\det(\widehat{\mathbf{V}}_i^\pi) > 0$ almost surely for N large enough due to $\Sigma > 0$, the corresponding Moore-Penrose inverse converges as well in probability and hence another application of the continuous mapping theorem proves the result using Theorem 9.2.2 in [7]. \square

9 Other resampling approaches

9.1 Nonparametric bootstrap approach

Here, we consider a nonparametric bootstrap sample $\mathbf{Y}^* = (Y_{111}^*, \dots, Y_{an_a t_a}^*)$ drawn with replacement from the pooled observation vector $\mathbf{Y} = (Y_{111}, \dots, Y_{an_a t_a})$. Therefore, given the observations, the bootstrap components are all independent with identical distribution which is given by the empirical distribution of \mathbf{Y}^* . The WTS of the bootstrap sample is given by

$$Q_N^* = N(\overline{\mathbf{Y}}^*)^\top \mathbf{H}^\top (\mathbf{H} \widehat{\Sigma}^* \mathbf{H}^\top)^+ \mathbf{H} \overline{\mathbf{Y}}^*,$$

where $\overline{\mathbf{Y}}^*$ is the vector of means of the bootstrap sample and $\widehat{\Sigma}^*$ denotes their covariance matrix.

THEOREM 9.1 *The distribution of Q_N^* conditioned on the observed data \mathbf{Y} weakly converges to the central χ_f^2 distribution in probability, where $f = \text{rank}(\mathbf{H})$. In particular, we have*

$$\sup_{x \in \mathbb{R}} \left| \Pr_\mu(Q_N^* \leq x | \mathbf{Y}) - \Pr_{\mu_0}(Q_N \leq x) \right| \rightarrow 0 \quad (9.12)$$

in probability for any underlying parameter $\mu \in \mathbb{R}^T$ and $\mu_0 \in \mathcal{H}_0(\mathbf{H})$.

Proof: The result follows analogously to the proof of Theorem 3.1 in the paper. \square

Note, that a nonparametric bootstrap version based on drawing with replacement from the observation vectors as in [3] performed considerably worse than the parametric bootstrap approach described below and is therefore not reported here.

In addition, we have also studied a nonparametric bootstrap version of the ATS (although this is in general not asymptotically correct) given by

$$F_N^* = \frac{N}{\text{tr}(\mathbf{T}\widehat{\Sigma}^*)} (\overline{\mathbf{Y}}_\bullet^*)^\top \mathbf{T} \overline{\mathbf{Y}}_\bullet^*.$$

A corresponding permutation version of the ATS has not been considered since it is also asymptotically only an approximation.

9.2 Parametric bootstrap approach

We have also considered a parametric bootstrap approach as studied by, e.g., Konietzschke et al. [3]. Here, the parametric bootstrap variables are generated as

$$\mathbf{Y}_i^* \stackrel{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \widehat{\mathbf{V}}_i), \quad 1 \leq i \leq a.$$

The idea behind this approach is to obtain a more accurate finite-sample approximation by mimicking the given covariance structure of the original data. We can again compute the WTS and ATS from the parametric bootstrap vectors as

$$Q_N^* = N (\overline{\mathbf{Y}}_\bullet^*)^\top \mathbf{H}^\top (\mathbf{H} \widehat{\Sigma}^* \mathbf{H}^\top)^+ \mathbf{H} \overline{\mathbf{Y}}_\bullet^*,$$

and

$$F_N^* = \frac{N}{\text{tr}(\mathbf{T}\widehat{\Sigma}^*)} (\overline{\mathbf{Y}}_\bullet^*)^\top \mathbf{T} \overline{\mathbf{Y}}_\bullet^*,$$

where $\overline{\mathbf{Y}}_\bullet^*$ is the vector of means of the parametric bootstrap sample and $\widehat{\Sigma}^*$ denotes their empirical covariance matrix.

THEOREM 9.2 *The distribution of Q_N^* conditioned on the observed data \mathbf{Y} weakly converges to the central χ_f^2 distribution in probability, where $f = \text{rank}(\mathbf{H})$. In particular, we have*

$$\sup_{x \in \mathbb{R}} |\Pr_\mu(Q_N^* \leq x | \mathbf{Y}) - \Pr_{\mu_0}(Q_N \leq x)| \rightarrow 0 \quad (9.13)$$

in probability for any underlying parameters $\mu, \mu_0 \in \mathbb{R}^T$ with $\mathbf{H}\mu_0 = \mathbf{0}$.

Furthermore, for the ATS of the parametric bootstrap sample it also holds that

$$\sup_{x \in \mathbb{R}} |\Pr_\mu(F_N^* \leq x | \mathbf{Y}) - \Pr_{\mu_0}(F_N \leq x)| \rightarrow 0 \quad (9.14)$$

in probability for any underlying parameters $\mu, \mu_0 \in \mathbb{R}^T$ with $\mathbf{H}\mu_0 = \mathbf{0}$. Thus, the conditional distribution of F_N^ always approximates the null distribution of F_N .*

Proof: The result for the WTS follows analogously to the proof of Theorem 3.2 in [3]. For the parametric bootstrap version of the ATS the result is obtained by the multivariate Lindeberg-Feller Theorem, the Continuous Mapping Theorem and another application of Slutsky's Theorem. The details are left to the reader. \square

9.3 Type-I error rates

In the following, we present the results of the detailed simulation studies conducted as described in Section 4 of the paper. For comparison, the results of the permutation approach are also included. The results for the hypothesis of no time effect T are presented in Tables 6, 7 and 8 for the normal, log-normal and exponential distribution, respectively. The results for the hypothesis of no group \times time interaction are in Tables 9, 10 and 11, respectively. The parametric bootstrap approach is denoted by PBS, the nonparametric bootstrap by NPBS. The results are again compared to the asymptotic quantiles, i.e., the $\mathcal{F}(\hat{\nu}, \infty)$ -quantile for the ATS and the χ^2_f -quantile for the WTS. A permutation version of the ATS has not been considered for the reasons stated above. The covariance settings and the number of simulated individuals are the same as described in Section 4. It turns out that, in terms of type-I error control, the permutation approach performs best across the different scenarios. The parametric bootstrap, in comparison, tends to rather conservative or liberal decisions when applied to the ATS and WTS, respectively, whereas the non-parametric bootstrap approach for the WTS leads to similar but slightly worse results than the permutation procedure.

Table 6: Simulation results for the hypothesis of no time effect with normal distribution.
normal distribution

Cov. Setting	T		$t = 4$		$t = 8$	
	n	Method	ATS	WTS	ATS	WTS
1	$n^{(1)}$	Permutation	NA	0.050	NA	0.050
		PBS	0.041	0.052	0.034	0.059
		NPBS	0.048	0.047	0.052	0.050
		asymptotic	0.046	0.085	0.040	0.177
	$n^{(2)}$	Permutation	NA	0.048	NA	0.052
		PBS	0.041	0.050	0.034	0.060
		NPBS	0.046	0.048	0.052	0.050
		asymptotic	0.046	0.086	0.040	0.177
	$n^{(3)}$	Permutation	NA	0.051	NA	0.052
		PBS	0.045	0.048	0.040	0.050
		NPBS	0.051	0.051	0.050	0.052
		asymptotic	0.050	0.078	0.043	0.135
2	$n^{(1)}$	Permutation	NA	0.050	NA	0.051
		PBS	0.046	0.052	0.036	0.059
		NPBS	0.056	0.051	0.054	0.050
		asymptotic	0.051	0.085	0.042	0.177
	$n^{(2)}$	Permutation	NA	0.051	NA	0.052
		PBS	0.046	0.052	0.036	0.060
		NPBS	0.057	0.051	0.056	0.052
		asymptotic	0.052	0.086	0.043	0.177
	$n^{(3)}$	Permutation	NA	0.051	NA	0.052
		PBS	0.049	0.048	0.038	0.049
		NPBS	0.059	0.049	0.054	0.051
		asymptotic	0.053	0.077	0.041	0.135
3	$n^{(1)}$	Permutation	NA	0.052	NA	0.062
		PBS	0.041	0.052	0.040	0.064
		NPBS	0.052	0.052	0.069	0.061
		asymptotic	0.046	0.092	0.044	0.198
	$n^{(2)}$	Permutation	NA	0.045	NA	0.042
		PBS	0.047	0.052	0.043	0.056
		NPBS	0.056	0.043	0.075	0.042
		asymptotic	0.051	0.080	0.048	0.155
	$n^{(3)}$	Permutation	NA	0.053	NA	0.054
		PBS	0.047	0.050	0.044	0.049
		NPBS	0.058	0.051	0.073	0.052
		asymptotic	0.051	0.078	0.048	0.136

Table 7: Simulation results for the hypothesis of no time effect with log-normal distribution.
log-normal distribution

Cov. Setting		T	$t = 4$		$t = 8$	
	n	Method	ATS	WTS	ATS	WTS
1	$n^{(1)}$	Permutation	NA	0.051	NA	0.047
		PBS	0.026	0.055	0.017	0.075
		NPBS	0.051	0.050	0.047	0.048
		asymptotic	0.032	0.094	0.021	0.198
	$n^{(2)}$	Permutation	NA	0.052	NA	0.046
		PBS	0.025	0.058	0.016	0.074
		NPBS	0.048	0.051	0.054	0.046
		asymptotic	0.031	0.090	0.020	0.198
	$n^{(3)}$	Permutation	NA	0.051	NA	0.048
		PBS	0.026	0.056	0.019	0.077
		NPBS	0.052	0.050	0.051	0.050
		asymptotic	0.031	0.089	0.021	0.186
2	$n^{(1)}$	Permutation	NA	0.067	NA	0.053
		PBS	0.035	0.072	0.018	0.084
		NPBS	0.060	0.066	0.052	0.053
		asymptotic	0.040	0.110	0.022	0.207
	$n^{(2)}$	Permutation	NA	0.067	NA	0.051
		PBS	0.034	0.073	0.018	0.082
		NPBS	0.061	0.066	0.057	0.052
		asymptotic	0.040	0.107	0.022	0.203
	$n^{(3)}$	Permutation	NA	0.070	NA	0.057
		PBS	0.037	0.072	0.021	0.080
		NPBS	0.065	0.068	0.057	0.057
		asymptotic	0.042	0.107	0.024	0.197
3	$n^{(1)}$	Permutation	NA	0.057	NA	0.064
		PBS	0.027	0.059	0.021	0.082
		NPBS	0.054	0.058	0.063	0.064
		asymptotic	0.033	0.101	0.024	0.221
	$n^{(2)}$	Permutation	NA	0.053	NA	0.048
		PBS	0.031	0.060	0.028	0.075
		NPBS	0.057	0.053	0.079	0.047
		asymptotic	0.037	0.090	0.033	0.190
	$n^{(3)}$	Permutation	NA	0.057	NA	0.062
		PBS	0.031	0.059	0.027	0.079
		NPBS	0.057	0.054	0.075	0.062
		asymptotic	0.036	0.092	0.031	0.191

Table 8: Simulation results for the hypothesis of no time effect with exponential distribution.
exponential distribution

T		$t = 4$		$t = 8$		
Cov. Setting	n	Method	ATS	WTS	ATS	WTS
1	$n^{(1)}$	Permutation	NA	0.048	NA	0.051
		PBS	0.038	0.055	0.026	0.070
		NPBS	0.052	0.049	0.052	0.051
		asymptotic	0.045	0.090	0.034	0.194
	$n^{(2)}$	Permutation	NA	0.053	NA	0.048
		PBS	0.039	0.057	0.029	0.069
		NPBS	0.053	0.053	0.050	0.048
		asymptotic	0.046	0.096	0.032	0.191
	$n^{(3)}$	Permutation	NA	0.054	NA	0.050
		PBS	0.041	0.057	0.031	0.059
		NPBS	0.053	0.054	0.049	0.050
		asymptotic	0.046	0.086	0.034	0.151
2	$n^{(1)}$	Permutation	NA	0.054	NA	0.052
		PBS	0.040	0.059	0.029	0.070
		NPBS	0.062	0.055	0.057	0.052
		asymptotic	0.048	0.093	0.035	0.194
	$n^{(2)}$	Permutation	NA	0.060	NA	0.051
		PBS	0.044	0.064	0.029	0.074
		NPBS	0.063	0.059	0.056	0.052
		asymptotic	0.050	0.101	0.034	0.193
	$n^{(3)}$	Permutation	NA	0.058	NA	0.051
		PBS	0.045	0.062	0.032	0.062
		NPBS	0.060	0.058	0.052	0.051
		asymptotic	0.050	0.088	0.036	0.154
3	$n^{(1)}$	Permutation	NA	0.055	NA	0.066
		PBS	0.041	0.055	0.034	0.074
		NPBS	0.060	0.054	0.073	0.065
		asymptotic	0.049	0.098	0.042	0.218
	$n^{(2)}$	Permutation	NA	0.049	NA	0.045
		PBS	0.045	0.058	0.039	0.067
		NPBS	0.062	0.049	0.078	0.044
		asymptotic	0.050	0.090	0.046	0.173
	$n^{(3)}$	Permutation	NA	0.055	NA	0.056
		PBS	0.045	0.058	0.038	0.060
		NPBS	0.062	0.055	0.072	0.056
		asymptotic	0.050	0.087	0.042	0.153

Table 9: Simulation results for the hypothesis of no group \times time interaction with normal distribution.

normal distribution						
GT		$t = 4$		$t = 8$		
Cov. Setting	n	Method	ATS	WTS	ATS	WTS
1	$n^{(1)}$	Permutation	NA	0.046	NA	0.051
		PBS	0.039	0.051	0.025	0.077
		NPBS	0.049	0.046	0.052	0.051
		asymptotic	0.049	0.135	0.033	0.432
	$n^{(2)}$	Permutation	NA	0.052	NA	0.050
		PBS	0.042	0.056	0.026	0.075
		NPBS	0.052	0.051	0.051	0.049
		asymptotic	0.053	0.142	0.034	0.433
	$n^{(3)}$	Permutation	NA	0.049	NA	0.051
		PBS	0.041	0.049	0.032	0.046
		NPBS	0.050	0.050	0.054	0.050
		asymptotic	0.048	0.126	0.039	0.366
2	$n^{(1)}$	Permutation	NA	0.050	NA	0.052
		PBS	0.045	0.054	0.030	0.076
		NPBS	0.060	0.050	0.055	0.053
		asymptotic	0.053	0.132	0.038	0.429
	$n^{(2)}$	Permutation	NA	0.054	NA	0.050
		PBS	0.044	0.056	0.029	0.072
		NPBS	0.059	0.053	0.057	0.051
		asymptotic	0.053	0.141	0.038	0.431
	$n^{(3)}$	Permutation	NA	0.052	NA	0.050
		PBS	0.044	0.049	0.034	0.046
		NPBS	0.059	0.052	0.060	0.050
		asymptotic	0.050	0.122	0.040	0.366
3	$n^{(1)}$	Permutation	NA	0.050	NA	0.065
		PBS	0.043	0.051	0.033	0.082
		NPBS	0.061	0.049	0.075	0.069
		asymptotic	0.054	0.141	0.040	0.465
	$n^{(2)}$	Permutation	NA	0.045	NA	0.037
		PBS	0.046	0.054	0.040	0.069
		NPBS	0.057	0.047	0.078	0.037
		asymptotic	0.053	0.135	0.049	0.393
	$n^{(3)}$	Permutation	NA	0.049	NA	0.053
		PBS	0.043	0.048	0.038	0.047
		NPBS	0.064	0.050	0.077	0.051
		asymptotic	0.051	0.126	0.045	0.363

Table 10: Simulation results for the hypothesis of no group \times time interaction with log-normal distribution.

log-normal distribution						
GT			$t = 4$		$t = 8$	
Cov. Setting	n	Method	ATS	WTS	ATS	WTS
1	$n^{(1)}$	Permutation	NA	0.047	NA	0.053
		PBS	0.019	0.040	0.009	0.061
		NPBS	0.048	0.047	0.048	0.052
		asymptotic	0.024	0.121	0.012	0.426
	$n^{(2)}$	Permutation	NA	0.053	NA	0.051
		PBS	0.017	0.044	0.009	0.055
		NPBS	0.048	0.053	0.050	0.048
		asymptotic	0.022	0.128	0.013	0.431
	$n^{(3)}$	Permutation	NA	0.048	NA	0.051
		PBS	0.018	0.037	0.010	0.042
		NPBS	0.048	0.046	0.047	0.051
		asymptotic	0.024	0.118	0.012	0.406
2	$n^{(1)}$	Permutation	NA	0.051	NA	0.054
		PBS	0.019	0.044	0.010	0.062
		NPBS	0.056	0.051	0.052	0.054
		asymptotic	0.025	0.129	0.014	0.427
	$n^{(2)}$	Permutation	NA	0.054	NA	0.052
		PBS	0.019	0.044	0.011	0.056
		NPBS	0.056	0.053	0.052	0.051
		asymptotic	0.026	0.130	0.013	0.432
	$n^{(3)}$	Permutation	NA	0.050	NA	0.052
		PBS	0.018	0.038	0.010	0.042
		NPBS	0.056	0.049	0.053	0.052
		asymptotic	0.023	0.120	0.013	0.403
3	$n^{(1)}$	Permutation	NA	0.050	NA	0.062
		PBS	0.022	0.042	0.014	0.067
		NPBS	0.053	0.050	0.068	0.060
		asymptotic	0.029	0.133	0.020	0.457
	$n^{(2)}$	Permutation	NA	0.045	NA	0.036
		PBS	0.022	0.043	0.020	0.053
		NPBS	0.055	0.046	0.076	0.035
		asymptotic	0.028	0.121	0.024	0.399
	$n^{(3)}$	Permutation	NA	0.049	NA	0.053
		PBS	0.023	0.037	0.014	0.043
		NPBS	0.059	0.046	0.071	0.054
		asymptotic	0.028	0.122	0.020	0.408

Table 11: Simulation results for the hypothesis of no group \times time interaction with exponential distribution.

exponential distribution						
GT		$t = 4$		$t = 8$		
Cov. Setting	n	Method	ATS	WTS	ATS	WTS
1	$n^{(1)}$	Permutation	NA	0.054	NA	0.054
		PBS	0.036	0.057	0.018	0.076
		NPBS	0.055	0.055	0.049	0.055
		asymptotic	0.043	0.146	0.024	0.442
	$n^{(2)}$	Permutation	NA	0.054	NA	0.050
		PBS	0.030	0.057	0.019	0.072
		NPBS	0.051	0.054	0.048	0.050
		asymptotic	0.041	0.148	0.024	0.443
	$n^{(3)}$	Permutation	NA	0.047	NA	0.054
		PBS	0.029	0.043	0.023	0.052
		NPBS	0.047	0.046	0.054	0.054
		asymptotic	0.036	0.122	0.028	0.397
2	$n^{(1)}$	Permutation	NA	0.059	NA	0.057
		PBS	0.040	0.061	0.019	0.077
		NPBS	0.065	0.061	0.054	0.056
		asymptotic	0.048	0.151	0.027	0.444
	$n^{(2)}$	Permutation	NA	0.059	NA	0.052
		PBS	0.035	0.060	0.019	0.072
		NPBS	0.058	0.059	0.050	0.051
		asymptotic	0.042	0.153	0.025	0.448
	$n^{(3)}$	Permutation	NA	0.048	NA	0.055
		PBS	0.028	0.042	0.024	0.051
		NPBS	0.051	0.048	0.058	0.054
		asymptotic	0.034	0.121	0.029	0.397
3	$n^{(1)}$	Permutation	NA	0.061	NA	0.068
		PBS	0.039	0.059	0.024	0.083
		NPBS	0.067	0.060	0.068	0.069
		asymptotic	0.047	0.155	0.032	0.473
	$n^{(2)}$	Permutation	NA	0.049	NA	0.037
		PBS	0.038	0.056	0.033	0.062
		NPBS	0.056	0.050	0.078	0.036
		asymptotic	0.043	0.140	0.042	0.406
	$n^{(3)}$	Permutation	NA	0.047	NA	0.058
		PBS	0.031	0.043	0.033	0.052
		NPBS	0.055	0.047	0.080	0.057
		asymptotic	0.037	0.122	0.041	0.402

10 Additional simulation results

10.1 Quality of the approximation

Recall that we defined

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

as well as

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

for the distances between the quantile functions of the WTS (F_N^{-1}) and the χ^2 -distribution (F^{-1}) and the WTPS ($(F_N^\pi)^{-1}$), respectively. The results for all simulation settings described in the paper are presented in Tables 12 and 13. Some plots of exemplarily chosen scenarios are in Figures 1-3.

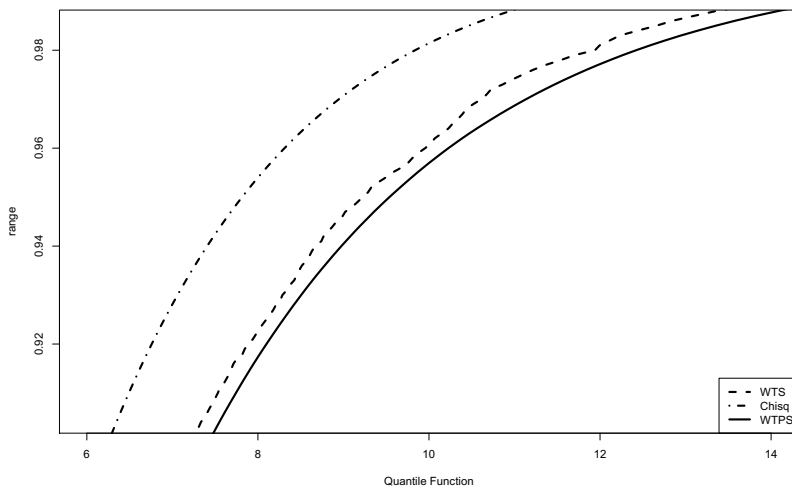


Figure 1: *Quantile functions of the WTS, WTPS and the corresponding χ^2 -distribution in the simulation setting with normally distributed data, $t = 4$, covariance matrix setting 3, $\mathbf{n}^{(2)}$ and under the null hypothesis of no time effect.*

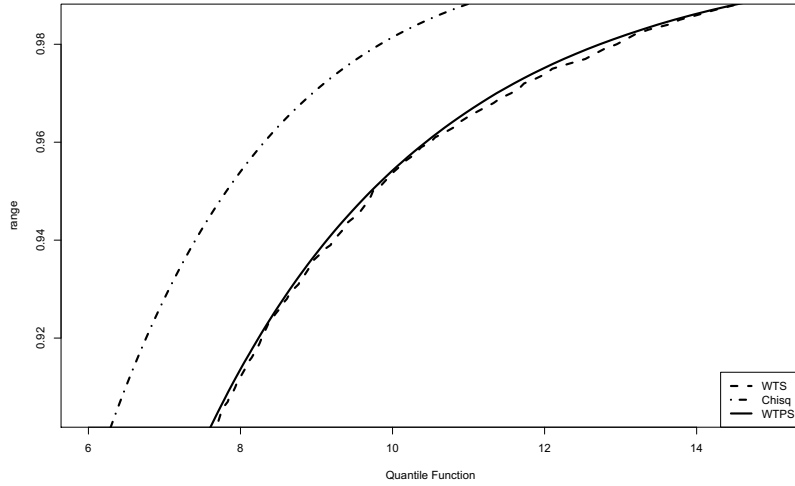


Figure 2: *Quantile functions of the WTS, WTPS and the corresponding χ^2 -distribution in the simulation setting with exponentially distributed data, $t = 4$, covariance matrix setting 1, $\mathbf{n}^{(1)}$ and under the null hypothesis of no time effect.*

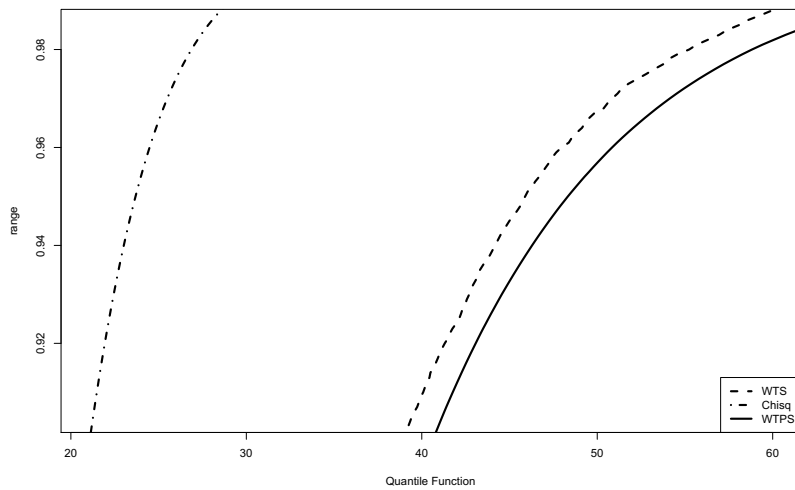


Figure 3: *Quantile functions of the WTS, WTPS and the corresponding χ^2 -distribution in the simulation setting with exponentially distributed data, $t = 8$, covariance matrix setting 3, $\mathbf{n}^{(2)}$ and under the null hypothesis of no interaction effect.*

Table 12: Simulation results for the distances between the quantile functions for the hypothesis of no time effect.

normal distribution					
T		$t = 4$		$t = 8$	
Cov. Setting		χ^2	WTPS	χ^2	WTPS
1	$\mathbf{n}^{(1)}$	3.683	0.411	10.548	0.381
	$\mathbf{n}^{(2)}$	3.299	0.310	11.393	0.654
	$\mathbf{n}^{(3)}$	2.198	0.213	7.771	0.281
2	$\mathbf{n}^{(1)}$	3.620	0.564	10.494	0.286
	$\mathbf{n}^{(2)}$	3.378	0.227	11.604	0.993
	$\mathbf{n}^{(3)}$	1.991	0.226	7.646	0.297
3	$\mathbf{n}^{(1)}$	4.451	1.186	12.515	2.086
	$\mathbf{n}^{(2)}$	2.599	0.731	9.564	1.486
	$\mathbf{n}^{(3)}$	2.264	0.105	7.571	0.344
log-normal distribution					
T		$t = 4$		$t = 8$	
Cov. Setting		χ^2	WTPS	χ^2	WTPS
1	$\mathbf{n}^{(1)}$	3.097	0.532	12.399	1.044
	$\mathbf{n}^{(2)}$	3.960	0.386	14.293	0.998
	$\mathbf{n}^{(3)}$	3.087	0.363	12.768	0.421
2	$\mathbf{n}^{(1)}$	5.645	2.258	14.165	1.105
	$\mathbf{n}^{(2)}$	6.045	2.656	15.484	2.617
	$\mathbf{n}^{(3)}$	5.062	1.891	14.239	1.815
3	$\mathbf{n}^{(1)}$	3.977	0.517	14.547	2.299
	$\mathbf{n}^{(2)}$	4.000	0.526	12.740	0.610
	$\mathbf{n}^{(3)}$	3.561	0.643	14.238	3.203
exponential distribution					
T		$t = 4$		$t = 8$	
Cov. Setting		χ^2	WTPS	χ^2	WTPS
1	$\mathbf{n}^{(1)}$	3.617	0.283	11.750	1.054
	$\mathbf{n}^{(2)}$	4.245	0.491	12.098	0.948
	$\mathbf{n}^{(3)}$	2.906	0.382	9.685	0.885
2	$\mathbf{n}^{(1)}$	4.761	1.262	11.704	0.628
	$\mathbf{n}^{(2)}$	4.961	1.366	12.201	0.724
	$\mathbf{n}^{(3)}$	3.833	0.915	10.226	0.286
3	$\mathbf{n}^{(1)}$	4.567	0.969	13.840	2.089
	$\mathbf{n}^{(2)}$	3.700	0.290	10.504	1.729
	$\mathbf{n}^{(3)}$	3.179	0.343	10.093	0.601

Table 13: Simulation results for the distances between the quantile functions for the hypothesis of no interaction.

normal distribution					
GT		$t = 4$		$t = 8$	
Cov. Setting		χ^2	WTPS	χ^2	WTPS
1	$\mathbf{n}^{(1)}$	9.151	0.420	40.954	1.135
	$\mathbf{n}^{(2)}$	8.872	0.573	41.179	1.451
	$\mathbf{n}^{(3)}$	7.789	0.711	30.617	0.905
2	$\mathbf{n}^{(1)}$	8.648	0.582	42.023	1.804
	$\mathbf{n}^{(2)}$	8.727	0.497	41.980	1.928
	$\mathbf{n}^{(3)}$	7.951	1.166	30.031	0.956
3	$\mathbf{n}^{(1)}$	10.280	1.108	48.106	7.618
	$\mathbf{n}^{(2)}$	7.700	1.463	36.252	4.470
	$\mathbf{n}^{(3)}$	7.579	0.604	31.374	1.461
log-normal distribution					
GT		$t = 4$		$t = 8$	
Cov. Setting		χ^2	WTPS	χ^2	WTPS
1	$\mathbf{n}^{(1)}$	5.292	0.610	30.474	1.058
	$\mathbf{n}^{(2)}$	5.952	0.360	30.014	1.671
	$\mathbf{n}^{(3)}$	5.767	0.467	27.019	1.024
2	$\mathbf{n}^{(1)}$	5.340	0.524	30.986	0.770
	$\mathbf{n}^{(2)}$	6.307	0.812	29.960	1.329
	$\mathbf{n}^{(3)}$	5.826	0.674	27.346	0.558
3	$\mathbf{n}^{(1)}$	6.425	0.298	34.755	2.106
	$\mathbf{n}^{(2)}$	5.124	1.408	26.657	6.691
	$\mathbf{n}^{(3)}$	5.561	0.182	27.517	1.363
exponential distribution					
GT		$t = 4$		$t = 8$	
Cov. Setting		χ^2	WTPS	χ^2	WTPS
1	$\mathbf{n}^{(1)}$	8.416	0.431	36.706	0.968
	$\mathbf{n}^{(2)}$	9.066	1.184	37.318	1.295
	$\mathbf{n}^{(3)}$	6.016	0.618	29.863	1.073
2	$\mathbf{n}^{(1)}$	8.523	0.869	36.999	1.044
	$\mathbf{n}^{(2)}$	9.206	1.510	37.160	1.436
	$\mathbf{n}^{(3)}$	6.445	0.293	30.130	0.925
3	$\mathbf{n}^{(1)}$	9.415	1.219	42.643	5.264
	$\mathbf{n}^{(2)}$	7.638	0.946	32.131	5.851
	$\mathbf{n}^{(3)}$	6.490	0.260	30.012	0.689

10.2 Large time correlations

In order to investigate the behavior of the proposed test for situations with very large time correlations, we investigated the auto-regressive setting with a correlation of $\rho = 0.9$ as well as

a Toeplitz-type covariance matrix with

$$\Sigma_i = \left(\frac{0.9}{0.9 + |\ell - j|} \right)_{\ell, j \leq t}, \quad i \in \{1, 2, 3\}$$

for all three groups. The resulting type-I error rates for the hypothesis of *no time effect* and *no group \times time interaction* are displayed in Tables 14 and 15, respectively.

Table 14: Type-I error rates in scenarios with large time correlation for the hypothesis of *no time effect*.

		$t = 4$			$t = 8$			
		ATS	WTS	WTPS	ATS	WTS	WTPS	
normal	AR(0.9)	$\mathbf{n}^{(1)}$	0.056	0.087	0.052	0.052	0.174	0.048
		$\mathbf{n}^{(2)}$	0.058	0.087	0.052	0.056	0.178	0.048
		$\mathbf{n}^{(3)}$	0.053	0.075	0.051	0.056	0.139	0.051
	Toeplitz	$\mathbf{n}^{(1)}$	0.053	0.087	0.051	0.040	0.175	0.047
		$\mathbf{n}^{(2)}$	0.053	0.088	0.051	0.044	0.177	0.048
		$\mathbf{n}^{(3)}$	0.051	0.075	0.050	0.050	0.140	0.054
log-normal	AR(0.9)	$\mathbf{n}^{(1)}$	0.042	0.094	0.054	0.036	0.210	0.065
		$\mathbf{n}^{(2)}$	0.040	0.099	0.055	0.044	0.215	0.071
		$\mathbf{n}^{(3)}$	0.037	0.092	0.058	0.042	0.203	0.084
	Toeplitz	$\mathbf{n}^{(1)}$	0.031	0.092	0.051	0.023	0.203	0.056
		$\mathbf{n}^{(2)}$	0.031	0.096	0.053	0.025	0.208	0.060
		$\mathbf{n}^{(3)}$	0.030	0.089	0.053	0.026	0.196	0.065
exp	AR(0.9)	$\mathbf{n}^{(1)}$	0.053	0.093	0.054	0.050	0.188	0.050
		$\mathbf{n}^{(2)}$	0.051	0.091	0.054	0.051	0.194	0.054
		$\mathbf{n}^{(3)}$	0.054	0.082	0.054	0.049	0.165	0.065
	Toeplitz	$\mathbf{n}^{(1)}$	0.048	0.092	0.053	0.039	0.185	0.049
		$\mathbf{n}^{(2)}$	0.044	0.091	0.053	0.037	0.191	0.050
		$\mathbf{n}^{(3)}$	0.046	0.081	0.051	0.037	0.164	0.056

10.3 Large effect of factor B

In this scenario, we simulated a 2×2 repeated measures design with $t = 4$ repeated measures, exponentially distributed errors and a Toeplitz covariance structure as described above. In order to investigate the behavior of the permutation procedure in a scenario with large effects of one factor, we set

$$\mu_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}, \quad i, j \in \{1, 2\}$$

and chose $\alpha_i = 0$, $(\alpha\beta)_{ij} = 0$ and $\beta_1 = \delta \cdot \mathbf{1}_t = -\beta_2$ with $\delta = (0, 0.5, 1, 5, 10, 100)$. Type-I error rates for the null hypothesis of *no effect of factor A* as well as *no interaction effect AB* are given in Table 16. Note that ATS and WTS are identical in these scenarios.

Table 15: Type-I error rates in scenarios with large time correlation for the hypothesis of no group \times time interaction effect.

		$t = 4$			$t = 8$			
		ATS	WTS	WTPS	ATS	WTS	WTPS	
normal	AR(0.9)	$\mathbf{n}^{(1)}$	0.059	0.149	0.053	0.058	0.436	0.052
		$\mathbf{n}^{(2)}$	0.065	0.145	0.052	0.056	0.430	0.054
		$\mathbf{n}^{(3)}$	0.057	0.131	0.054	0.056	0.370	0.049
	Toeplitz	$\mathbf{n}^{(1)}$	0.055	0.149	0.052	0.044	0.436	0.052
		$\mathbf{n}^{(2)}$	0.057	0.145	0.052	0.039	0.431	0.052
		$\mathbf{n}^{(3)}$	0.050	0.130	0.054	0.044	0.369	0.047
lognormal	AR(0.9)	$\mathbf{n}^{(1)}$	0.034	0.126	0.042	0.038	0.433	0.039
		$\mathbf{n}^{(2)}$	0.036	0.125	0.043	0.038	0.435	0.038
		$\mathbf{n}^{(3)}$	0.034	0.124	0.047	0.037	0.407	0.047
	Toeplitz	$\mathbf{n}^{(1)}$	0.027	0.126	0.047	0.019	0.431	0.045
		$\mathbf{n}^{(2)}$	0.026	0.125	0.048	0.019	0.430	0.044
		$\mathbf{n}^{(3)}$	0.026	0.123	0.051	0.017	0.406	0.048
exp	AR(0.9)	$\mathbf{n}^{(1)}$	0.048	0.143	0.051	0.052	0.452	0.049
		$\mathbf{n}^{(2)}$	0.049	0.143	0.051	0.055	0.448	0.051
		$\mathbf{n}^{(3)}$	0.051	0.132	0.048	0.047	0.389	0.049
	Toeplitz	$\mathbf{n}^{(1)}$	0.042	0.142	0.053	0.034	0.450	0.051
		$\mathbf{n}^{(2)}$	0.040	0.142	0.050	0.033	0.445	0.054
		$\mathbf{n}^{(3)}$	0.043	0.134	0.050	0.030	0.389	0.047

Table 16: Type-I errors in a simulation setting with large effect of factor B .

δ	ATS = WTS	WTPS
	'no effect of factor A'	
0	0.056	0.050
0.5	0.056	0.050
1	0.056	0.049
5	0.056	0.049
10	0.056	0.049
100	0.056	0.048
	'no interaction effect AB'	
0	0.057	0.052
0.5	0.057	0.050
1	0.057	0.050
5	0.057	0.050
10	0.057	0.050
100	0.057	0.049

11 Power

We have also conducted several simulations to analyze the power of our method. 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 analyses. We considered a two sample repeated measures design, where we have simulated data as

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

with $\boldsymbol{\mu}_i = \mathbf{E}(\mathbf{Y}_{i1})$, $i \in \{1, 2\}$, and $\mathbf{V}_i \equiv \mathbf{I}_t$. The i.i.d. random vectors $\boldsymbol{\epsilon}_{ik} = (\epsilon_{ik1}, \dots, \epsilon_{ikt})^\top$, $i \in \{1, 2\}$, 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 or log-normal random variables. For the power simulation we have considered a trend alternative, i.e., we set $\boldsymbol{\mu}_2 = \mathbf{0}$ in the second group and $\boldsymbol{\mu}_1 = \delta \mathbf{c} = \delta(c_1, \dots, c_t)^\top$ in the first group, where $c_s = s/t$, $s \in \{1, \dots, t\}$ and $\delta \in \{0, 0.5, 1, 1.5, 2, 3\}$.

We considered a balanced design with 15 individuals per group, hypothesis matrix $\mathbf{H} = \mathbf{P}_t(\mathbf{I}_t - \mathbf{I}_t)$ and again simulated both $t = 4$ and $t = 8$ repeated measures. Figures 4 and 5 display the power comparison for the WTPS, the ATS, the approximation described by Lecoutre [4] as well as Hotelling's T^2 [2] for normal distribution and $t = 4$ and $t = 8$ repeated measures, respectively. In Figures 6 and 7, the results for the log-normal distribution are displayed. From these figures it appears that the ATS has slightly higher power for normally distributed data. 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. We also note that the approximation by Huynh-Feldt and Lecoutre performs worst for the log-normal distribution.

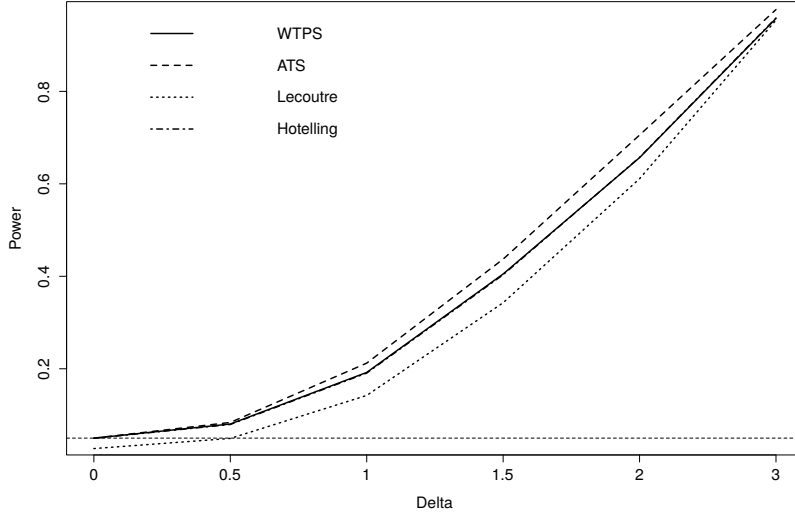


Figure 4: Power ($\alpha = 0.05$) simulation results of the WTPS, ATS, Lecoutre and Hotelling for normal distribution and $t = 4$ repeated measures under a trend alternative $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top = (\mathbf{0}^\top, \boldsymbol{\delta c}^\top)^\top$ with $\boldsymbol{\delta} = \{0, 0.5, 1, 1.5, 2, 3\}$ and $c_s = s/t, s \in \{1, \dots, t\}$.

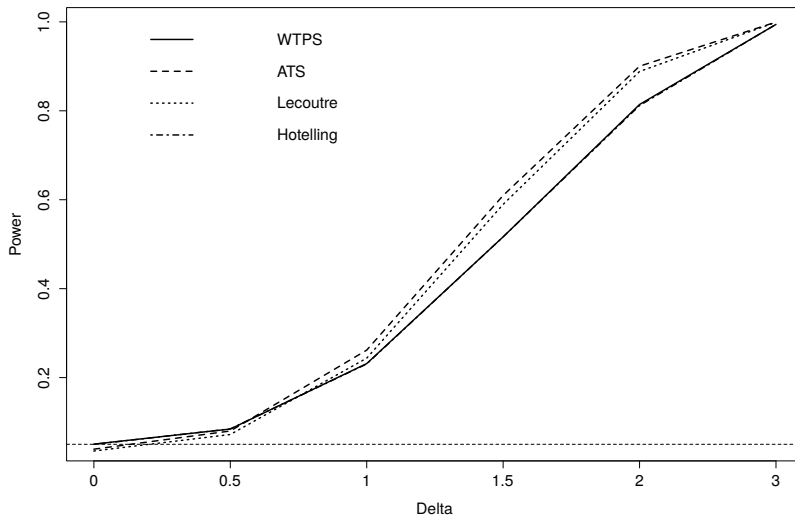


Figure 5: Power ($\alpha = 0.05$) simulation results of the WTPS, ATS, Lecoutre and Hotelling for normal distribution and $t = 8$ repeated measures under a trend alternative $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top = (\mathbf{0}^\top, \boldsymbol{\delta c}^\top)^\top$ with $\boldsymbol{\delta} = \{0, 0.5, 1, 1.5, 2, 3\}$ and $c_s = s/t, s \in \{1, \dots, t\}$.

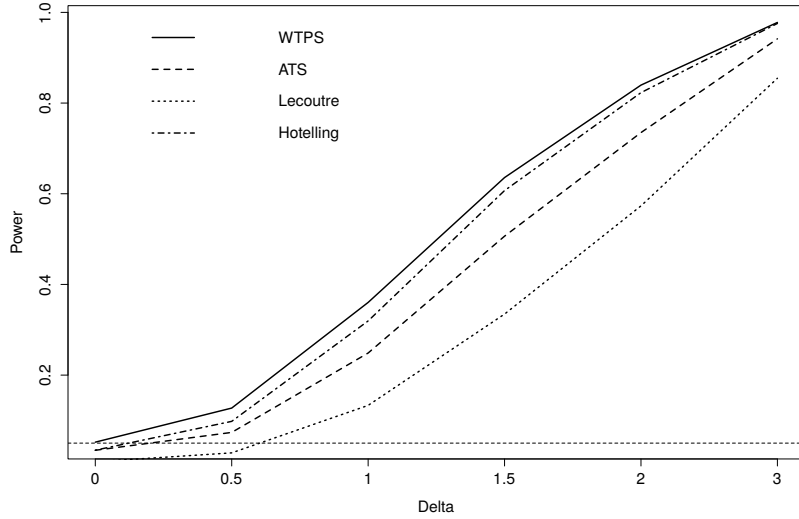


Figure 6: *Power* ($\alpha = 0.05$) simulation results of the WTPS, ATS, Lecoutre and Hotelling for log-normal distribution and $t = 4$ repeated measures under a trend alternative $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top = (\mathbf{0}^\top, \delta \mathbf{c}^\top)^\top$ with $\delta = \{0, 0.5, 1, 1.5, 2, 3\}$ and $c_s = s/t, s \in \{1, \dots, t\}$.

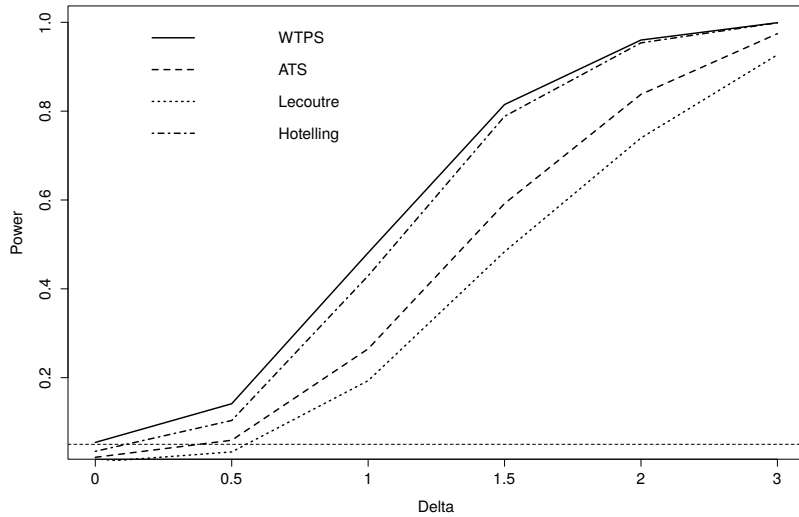


Figure 7: *Power* ($\alpha = 0.05$) simulation results of the WTPS, ATS, Lecoutre and Hotelling for log-normal distribution and $t = 8$ repeated measures under a trend alternative $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top = (\mathbf{0}^\top, \delta \mathbf{c}^\top)^\top$ with $\delta = \{0, 0.5, 1, 1.5, 2, 3\}$ and $c_s = s/t, s \in \{1, \dots, t\}$.

12 Analysis of the data example: Comparing the different approaches

We again consider the data example from Section 5 on the oxygen consumption of leukocytes. First of all, we notice that the empirical covariance matrices of the two groups appear to be quite different. The empirical covariance matrix in the Placebo-group (rounded to three digits) is given as

$$\begin{pmatrix} 0.025 & -0.022 & -0.004 & 0.009 & 0.015 & 0.025 \\ -0.022 & 0.092 & -0.005 & -0.001 & -0.024 & -0.035 \\ -0.004 & -0.005 & 0.081 & -0.013 & -0.010 & -0.004 \\ 0.009 & -0.001 & -0.013 & 0.037 & 0.044 & 0.038 \\ 0.015 & -0.024 & -0.010 & 0.044 & 0.069 & 0.063 \\ 0.025 & -0.035 & -0.004 & 0.038 & 0.063 & 0.115 \end{pmatrix}$$

whereas in the Verum-group we have

$$\begin{pmatrix} 0.043 & 0.012 & 0.046 & 0.033 & 0.014 & 0.055 \\ 0.012 & 0.113 & 0.008 & 0.009 & 0.060 & 0.032 \\ 0.046 & 0.008 & 0.065 & 0.041 & 0.005 & 0.066 \\ 0.033 & 0.009 & 0.041 & 0.047 & 0.016 & 0.059 \\ 0.014 & 0.060 & 0.005 & 0.016 & 0.058 & 0.047 \\ 0.055 & 0.032 & 0.066 & 0.059 & 0.047 & 0.116 \end{pmatrix}$$

. Thus, the assumption of homoscedasticity is not fulfilled in this data example.

For completeness, we also include some correlation matrices. Separately for the two groups, we get

$$\begin{pmatrix} 1.000 & -0.468 & -0.093 & 0.290 & 0.368 & 0.462 \\ -0.468 & 1.000 & -0.053 & -0.019 & -0.300 & -0.342 \\ -0.093 & -0.053 & 1.000 & -0.230 & -0.137 & -0.041 \\ 0.290 & -0.019 & -0.230 & 1.000 & 0.864 & 0.578 \\ 0.368 & -0.300 & -0.137 & 0.864 & 1.000 & 0.701 \\ 0.462 & -0.342 & -0.041 & 0.578 & 0.701 & 1.000 \end{pmatrix}$$

in the Placebo-group and for the Verum-group we have

$$\begin{pmatrix} 1.000 & 0.177 & 0.875 & 0.727 & 0.273 & 0.784 \\ 0.177 & 1.000 & 0.089 & 0.121 & 0.735 & 0.278 \\ 0.875 & 0.089 & 1.000 & 0.729 & 0.083 & 0.759 \\ 0.727 & 0.121 & 0.729 & 1.000 & 0.301 & 0.793 \\ 0.273 & 0.735 & 0.083 & 0.301 & 1.000 & 0.567 \\ 0.784 & 0.278 & 0.759 & 0.793 & 0.567 & 1.000 \end{pmatrix}$$

Considering the time correlations across the different groups, we get:

$$\begin{pmatrix} 1.000 & -0.018 & 0.376 & 0.557 & 0.326 & 0.632 \\ -0.018 & 1.000 & 0.367 & 0.143 & 0.324 & 0.172 \\ 0.376 & 0.367 & 1.000 & 0.303 & 0.168 & 0.477 \\ 0.557 & 0.143 & 0.303 & 1.000 & 0.593 & 0.699 \\ 0.326 & 0.324 & 0.168 & 0.593 & 1.000 & 0.670 \\ 0.632 & 0.172 & 0.477 & 0.699 & 0.670 & 1.000 \end{pmatrix}$$

The results of the analyses using the different methods are presented in the following table. The asymptotic results are again obtained by considering the corresponding $\mathcal{F}(\hat{\nu}, \infty)$ -quantile for the ATS and the χ^2_j -quantile for the WTS.

Table 17: *p-values of the analysis of the O_2 consumption data.*

	ATS			WTS			
	asymptotic	PBS	NPBS	asymptotic	Permutation	PBS	NPBS
A	0.001	0.002	0.003	0.001	0.003	0.002	0.003
B	<0.001	0.001	0.001	<0.001	<0.001	0.001	0.001
T	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
AB	0.110	0.110	0.128	0.110	0.133	0.110	0.128
AT	0.009	0.012	0.009	<0.001	<0.001	<0.001	0.001
BT	0.094	0.0108	0.105	0.115	0.151	0.161	0.155
ABT	0.117	0.138	0.132	0.116	0.164	0.170	0.157

For this data set, the results are similar for all resampling methods and the asymptotic approaches considered.

References

- [1] Hájek, J. , Šidák, Z. and Sen, P. K. (1999). *Theory of Rank Tests*. Academic Press, San Diego.
- [2] Hotelling, H. (1931). A generalization of Student's ratio. *Annals of Mathematical Statistics*, **2**, 360–378.
- [3] Konietzschke, F., Bathke, A. C., Harrar, S. W. and Pauly, M. (2015). Parametric and Non-parametric Bootstrap Methods for General MANOVA. *Journal of Multivariate Analysis* **140**, 291–301.
- [4] Lecoutre, B. (1991). A Correction for the $\tilde{\epsilon}$ Approximative Test in Repeated Measures Designs With Two or More Independent Groups. *Journal of Educational Statistics*, **16**, 371–372.

- [5] Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics*, **5**, 41–52.
- [6] Pauly, M., Brunner, E. and Konietschke, F. (2015a). Asymptotic Permutation Tests in General Factorial Designs. *Journal of the Royal Statistical Society - Series B*, **77**, 461–473.
- [7] Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications*. Wiley, New York.

Article 2

Friedrich, S., Konietschke, F. and Pauly, M. (2017). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, **113**, 38–52, DOI: 10.1016/j.csda.2016.06.016.

Reprinted from Computational Statistics and Data Analysis, Vol. 113, S. Friedrich, F. Konietschke and M. Pauly, A wild bootstrap approach for nonparametric repeated measurements, pages 38–52, Copyright (2017), with permission from Elsevier.



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A wild bootstrap approach for nonparametric repeated measurements

Sarah Friedrich^a, Frank Konietzschke^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

ARTICLE INFO

Article history:

Received 29 January 2016

Received in revised form 22 June 2016

Accepted 28 June 2016

Available online 6 July 2016

Keywords:

Longitudinal data

Quadratic forms

Rank-based methods

Repeated measures

Wild bootstrap

Wild cluster bootstrap

ABSTRACT

Repeated measures and split plot plans are often the preferred design of choice when planning experiments in life and social sciences. They are typically analyzed by mean-based methods from MANOVA or linear mixed models, requiring certain assumptions on the underlying parametric distribution. However, if count, ordinal or score data are present, these techniques show their limits since means are no adequate measure of deviations between groups. Here, nonparametric rank-based methods are preferred for making statistical inference. The common nonparametric procedures such as the Wald- or ANOVA-type tests, however, have drawbacks since they usually require large sample sizes for accurate test decisions. The aim is to enhance the small sample properties of these test statistics by means of a specific nonparametric bootstrap procedure while preserving their general applicability for all kinds of data in factorial repeated measures and split plot designs. In particular, it is shown that a specific wild bootstrap procedure inherits the large sample properties of the Wald- and ANOVA-type statistics while considerably improving their small sample behavior. The new method is motivated by and applied to a practical data example in a repeated measures design with score data.

© 2016 Elsevier B.V. All rights reserved.

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 Konietzschke 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), Konietzschke et al. (2015), Pauly et al. (2015) or Friedrich et al. (2015), among others.

If count, ordinal, ordered categorical 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 categorical 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 categorical 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 \mathbf{I}_t the t -dimensional unity matrix and by \mathbf{J}_t the $t \times t$ matrix of 1’s, i.e. $\mathbf{J}_t = \mathbf{1}_t \mathbf{1}_t'$, where $\mathbf{1}_t = (1, \dots, 1)'$ is the t -dimensional column vector of 1’s. Furthermore, let $\mathbf{P}_t = \mathbf{I}_t - \frac{1}{t} \mathbf{J}_t$ 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

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

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

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

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

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

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

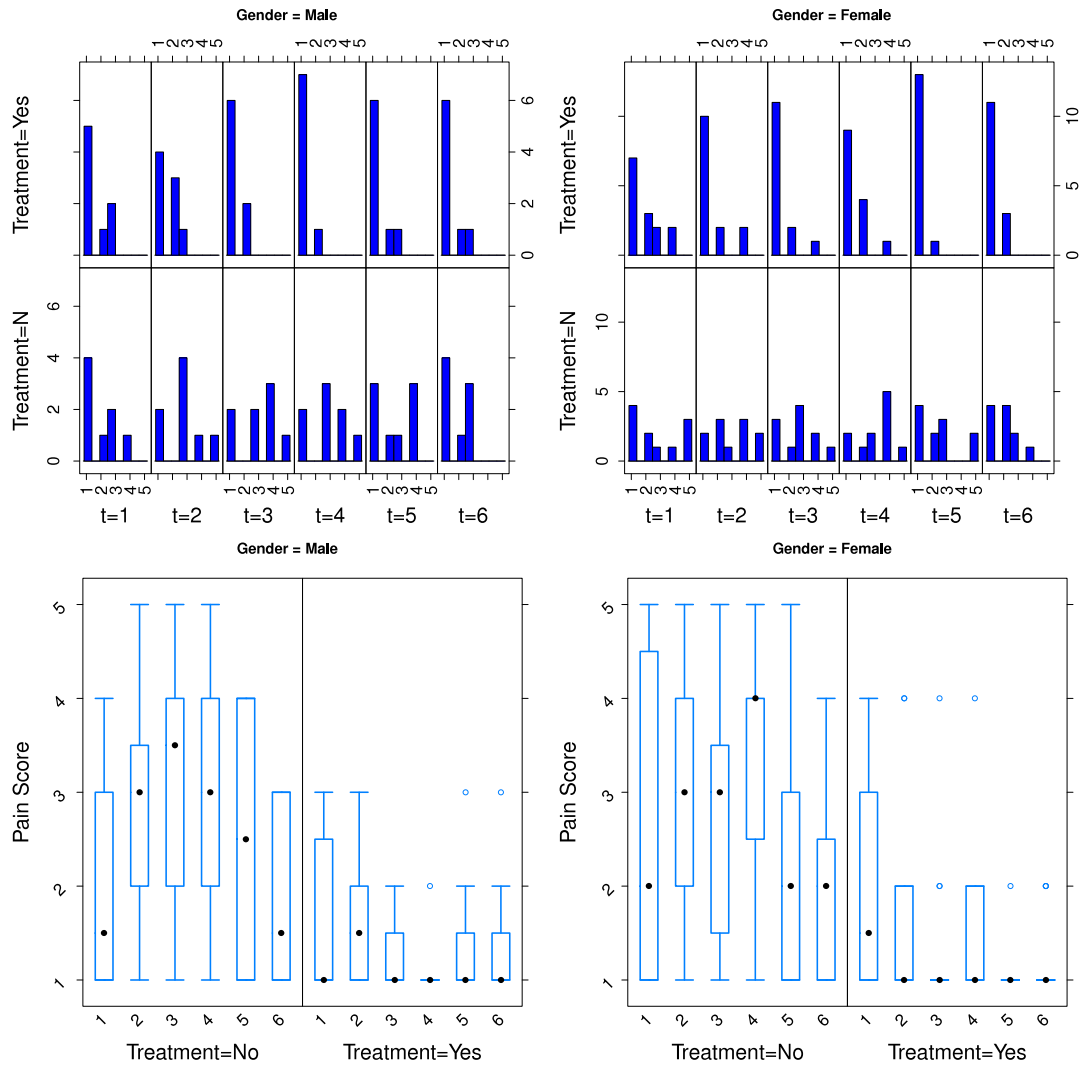


Fig. 1. Frequencies of the pain scores observed in the shoulder pain trial (Lumley, 1996).

In this set-up, null hypotheses are formulated by $H_0^F : \mathbf{CF} = \mathbf{0}$, where $\mathbf{F} = (F_{11}, \dots, F_{at})'$ denotes the vector of the distribution functions F_{is} , $i = 1, \dots, a$, $s = 1, \dots, t$ and \mathbf{C} is a suitable hypothesis matrix. Rank-statistics for testing these hypotheses are derived by considering estimates of the relative marginal effects $\mathbf{p} = (p_{11}, \dots, p_{at})'$, where $p_{is} = \int H_N dF_{is}$. Here, $H_N(x) = \frac{1}{tN} \sum_{i=1}^a \sum_{s=1}^t n_i F_{is}(x)$ 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

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

where R_{iks} denotes the (mid-)rank of X_{iks} among all tN observations and $\bar{R}_{i,s} = \frac{1}{n_i} \sum_{k=1}^{n_i} R_{iks}$. 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 $\hat{\mathbf{p}} = (\hat{p}_{11}, \dots, \hat{p}_{at})'$. Assuming the usual sample size condition

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

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

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

is the weighted block diagonal matrix of the covariance matrices $\mathbf{V}_i = \text{Cov}(\mathbf{Y}_{ik})$ of the random vectors $\mathbf{Y}_{ik} = (H(X_{ik1}), \dots, H(X_{ikt}))'$ 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 $H_0^F : \mathbf{C}\mathbf{F} = \mathbf{0}$ will be introduced. First, the Wald-type statistic (WTS) of Akritas and Arnold (1994) and Brunner and Puri (2001)

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

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

$$\widehat{\mathbf{V}}_i = \frac{1}{(tN)^2(n_i - 1)} \sum_{k=1}^{n_i} (\mathbf{R}_{ik} - \bar{\mathbf{R}}_i)(\mathbf{R}_{ik} - \bar{\mathbf{R}}_i)', \quad i = 1, \dots, a,$$

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

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

$$Q_N \xrightarrow{d} Q \sim \chi_{\text{rank}(\mathbf{C})}^2. \quad (2.5)$$

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:

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

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 $H_0^F : \mathbf{C}\mathbf{F} = \mathbf{0}$ and assumption (2.2), the statistic A_N has, asymptotically as $N \rightarrow \infty$, the same distribution as a weighted sum of χ_1^2 -distributed random variables, i.e.,

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

where $\xi_{is} \stackrel{i.i.d.}{\sim} \chi_1^2$ and the weights λ_{is} are the eigenvalues of $\mathbf{T}\boldsymbol{\Sigma}$, 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

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

can be approximated by a central $F(\hat{f}, \infty)$ -distribution with $\hat{f} = \frac{(\text{tr}(\mathbf{T}\widehat{\boldsymbol{\Sigma}}))^2}{\text{tr}(\mathbf{T}\widehat{\boldsymbol{\Sigma}}\mathbf{T}\widehat{\boldsymbol{\Sigma}})}$ 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 (Konietzschke et al., 2015) in a semiparametric framework. For nonparametric bivariate data, Konietzschke 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 (\mathbf{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 $\mathbf{Z}_{ik} = (\mathbf{R}_{ik} - \bar{\mathbf{R}}_i)$, $i = 1, \dots, a$; $k = 1, \dots, n_i$ denote the centered rank vectors and let ϵ_{ik} denote independent and identically distributed random signs; thus fulfilling $E(\epsilon_{11}) = 0$ and $\text{Var}(\epsilon_{11}) = 1$. 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 $\text{Var}(\epsilon_{11}) = 1$. Now, consider the resampling vectors

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

which depict a conditional distribution of the centered rank vectors \mathbf{Z}_{ik} around zero. The shape of the distribution depends on \mathbf{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 \mathbf{Z}_{ik}^* is a symmetrization of the fixed vectors \mathbf{Z}_{ik} . Now, let

$$\hat{\mathbf{p}}_i^\epsilon = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{\epsilon_{ik}}{tN} (\mathbf{R}_{ik} - \bar{\mathbf{R}}_i) = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{tN} \mathbf{Z}_{ik}^*$$

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

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

of the vectors \mathbf{Z}_{ik}^* , respectively. For convenience, the vectors $\hat{\mathbf{p}}_i^\epsilon$ are collected in the vector $\hat{\mathbf{p}}^\epsilon = (\hat{\mathbf{p}}_1^\epsilon, \dots, \hat{\mathbf{p}}_a^\epsilon)'$. 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 \mathbf{R}_{ik} . In the next theorem, the conditional multivariate distribution of $\sqrt{N}\hat{\mathbf{p}}^\epsilon$ will be examined.

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

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

$$Q_N^\epsilon = N(\hat{\mathbf{p}}^\epsilon)' \mathbf{C}' (\mathbf{C} \widehat{\boldsymbol{\Sigma}}^\epsilon \mathbf{C}')^{-1} \mathbf{C} \hat{\mathbf{p}}^\epsilon, \quad (3.8)$$

and the ANOVA-type statistic (WATS)

$$F_N^\epsilon = \frac{N}{\text{tr}(\mathbf{T}\hat{\Sigma}^\epsilon)} (\hat{\mathbf{p}}^\epsilon)' \mathbf{T} \hat{\mathbf{p}}^\epsilon. \quad (3.9)$$

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

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

where $P_{H_0}(Q_N \leq x)$ 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 $\mathbf{p} \in \mathbb{R}^{at}$, i.e. we have

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

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 $\sqrt{N}\hat{\mathbf{C}}\hat{\mathbf{p}}$ under the hypothesis $H_0^F: \mathbf{C}\mathbf{F} = \mathbf{0}$. 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

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

were generated from independent observations $Y_{ik} \sim U[0, 1]$ and $Z_{iks} \sim U[0, 1]$, $i = 1, 2$, $k = 1, \dots, n_i$ and $s = 1, \dots, t$. 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	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:

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

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

Setting 1 (AR): $\boldsymbol{\Sigma}_i = (\rho^{|l-j|})_{l,j \leq t}$, $\rho = 0.6$ for $i = 1, 2$.

Setting 2 (CS): $\boldsymbol{\Sigma}_i = \mathbf{I}_t + \mathbf{J}_t$ for $i = 1, 2$.

The independent and identically distributed random vectors $\tilde{\mathbf{X}}_{ik} = (\tilde{X}_{ik1}, \dots, \tilde{X}_{ikt})$ 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 $\mathbf{n}^{(1)} = (10, 10)$ and $\mathbf{n}^{(2)} = (10, 20)$.

	Cov. setting	\mathbf{n}	Method	$t = 4$		$t = 8$	
				ATS	WTS	ATS	WTS
A	AR	$\mathbf{n}^{(1)}$	Classic	0.064	0.066	0.066	0.067
			Wild bootstrap	0.050	0.050	0.052	0.052
		$\mathbf{n}^{(2)}$	Classic	0.063	0.065	0.067	0.068
			Wild bootstrap	0.052	0.052	0.054	0.054
	CS	$\mathbf{n}^{(1)}$	Classic	0.064	0.066	0.067	0.067
			Wild bootstrap	0.050	0.050	0.052	0.052
	$\mathbf{n}^{(2)}$	Classic	0.064	0.066	0.067	0.067	
		Wild bootstrap	0.052	0.052	0.054	0.054	
T	AR	$\mathbf{n}^{(1)}$	Classic	0.055	0.114	0.050	0.314
			Wild bootstrap	0.050	0.049	0.053	0.049
		$\mathbf{n}^{(2)}$	Classic	0.056	0.109	0.050	0.273
			Wild bootstrap	0.053	0.053	0.053	0.058
	CS	$\mathbf{n}^{(1)}$	Classic	0.052	0.115	0.037	0.314
			Wild bootstrap	0.049	0.050	0.047	0.048
	$\mathbf{n}^{(2)}$	Classic	0.052	0.108	0.037	0.270	
		Wild bootstrap	0.051	0.053	0.048	0.056	
A:T	AR	$\mathbf{n}^{(1)}$	Classic	0.055	0.115	0.049	0.315
			Wild bootstrap	0.051	0.050	0.052	0.049
		$\mathbf{n}^{(2)}$	Classic	0.057	0.109	0.049	0.273
			Wild bootstrap	0.053	0.053	0.053	0.058
	CS	$\mathbf{n}^{(1)}$	Classic	0.052	0.114	0.037	0.315
			Wild bootstrap	0.049	0.050	0.048	0.047
	$\mathbf{n}^{(2)}$	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 $\mathbf{n}^{(1)} = (10, 10)$ and $\mathbf{n}^{(2)} = (10, 20)$.

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

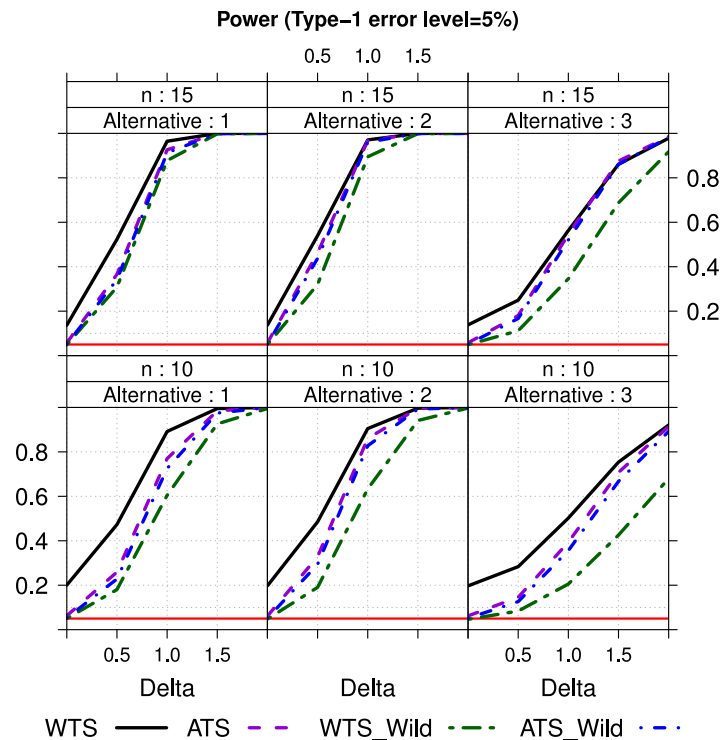


Fig. 2. Power simulation results (type-1 error level $\alpha = 5\%$) of the four investigated methods to detect the Alternatives 1–3 defined above with varying sample sizes $n = 10$ and $n = 15$, respectively.

structure $\mathbf{V}_{ij} = (0.6)^{|i-j|}$, $t = 4$ 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 \times gender \times 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

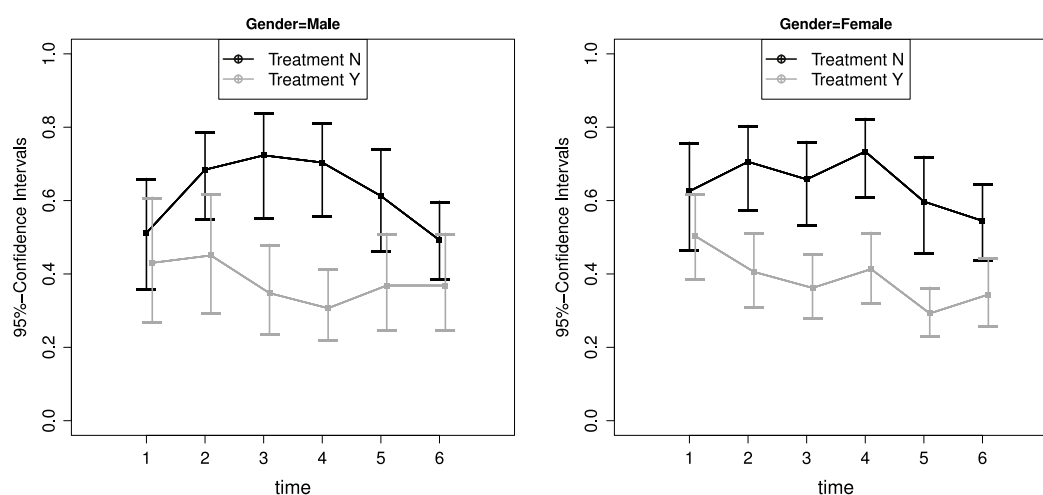


Fig. 3. Joint analysis of the data example: Treatment specific plots of the relative effects with 95%-confidence intervals—Gender: Male (left) and Female (right).

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

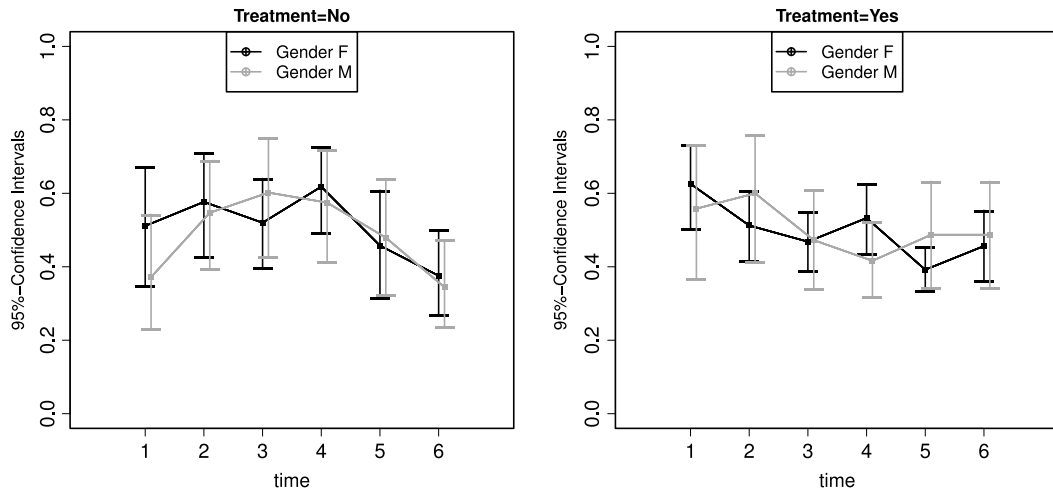


Fig. 4. Separate analysis of the data example per treatment: Plots of the relative effects with 95%-confidence intervals—Treatment: No (left) and Yes (right).

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 $H_0^p : \mathbf{Cp} = \mathbf{0}$ 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

$$p_{is} = \int H_N dF_{is}, \quad i = 1, \dots, a, \quad s = 1, \dots, t, \quad (\text{A.12})$$

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

$$\hat{F}_{is}(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{2} [\mathbb{1}(x > X_{ik_s}) + \mathbb{1}(x \geq X_{ik_s})], \quad (\text{A.13})$$

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

$$\hat{p}_{is} = \int \hat{H} d\hat{F}_{is} = \frac{1}{n_i} \sum_{k=1}^{n_i} \hat{H}(X_{ik_s}). \quad (\text{A.14})$$

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

$$\sqrt{N}\mathbf{C}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} N(\mathbf{0}, \mathbf{C}\Sigma\mathbf{C}') \quad (\text{A.15})$$

under H_0^f . Here, $\Sigma = \bigoplus_{i=1}^a \frac{1}{\kappa_i} \mathbf{V}_i$ and $\mathbf{V}_i = \text{Cov}(\mathbf{Y}_{i1})$. Note that the covariance matrices \mathbf{V}_i may not be equal, even if a homoscedastic model is assumed for \mathbf{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

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

has asymptotically a central $\chi_{\text{rank}(\mathbf{C})}^2$ distribution under H_0^f . Finally, the result follows by replacing Σ with $\hat{\Sigma}$ by applying the multivariate Slutsky Theorem and noting that the involved Moore–Penrose inverse is continuous since $\Sigma > \mathbf{0}$ by assumption. \square

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

Proof of Theorem 3.1. Due to conditional independence of the random variables $\hat{\mathbf{p}}_i^\epsilon$, $i = 1, \dots, a$, 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

$$\max_{1 \leq i \leq a} \frac{\sqrt{N} \|\hat{\mathbf{Y}}_{ik} - \hat{\mathbf{Y}}_i\|}{n_i} \xrightarrow{P} 0, \quad N \rightarrow \infty, \quad (\text{A.16})$$

as well as

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

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

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

Thus, we can conclude convergence in distribution

$$\frac{\sqrt{N}}{n_i} \sum_{k=1}^{n_i} \epsilon_{ik} (\widehat{\mathbf{Y}}_{ik} - \bar{\mathbf{Y}}_{i\cdot}) \xrightarrow{d} N\left(0, \frac{1}{\kappa_i} \mathbf{V}_i\right) \quad (i = 1, \dots, a)$$

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

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

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

First, it holds that

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

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

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

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

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. \square

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 analogously to $\widehat{\mathbf{p}}^\epsilon$, if we consider $\widehat{Z}_{iks} = \widehat{G}(X_{iks})$ instead of \widehat{Y}_{iks} , i.e. let $\widehat{\mathbf{q}}^\epsilon = \frac{1}{n_i} \sum_{k=1}^{n_i} \epsilon_{ik} (\widehat{\mathbf{Z}}_{ik} - \widehat{\mathbf{Z}}_{i\cdot})$ for $\widehat{\mathbf{Z}}_{i\cdot} = n_i^{-1} \sum_{k=1}^{n_i} \widehat{\mathbf{Z}}_{ik}$. Given the data \mathbf{X} , we have conditional convergence in distribution

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

in probability, where $\widetilde{\boldsymbol{\Sigma}} = \oplus_{\kappa_i} \frac{1}{\kappa_i} \widetilde{\mathbf{V}}_i$ and $\widetilde{\mathbf{V}}_i = \text{Cov}(G(X_{iks}))$, 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

- Akritis, M.G., 2011. Nonparametric models for anova and ancova designs. In: International Encyclopedia of Statistical Science. Springer, pp. 964–968.
- Akritis, 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.
- Akritis, 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. 45.1. 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-iiid 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., Touloupoulou, 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.

Supplement To
A Wild Bootstrap Approach for Nonparametric Repeated
Measurements

Sarah Friedrich^a, Frank Konietschke^b and Markus Pauly^{a,*}

June 22, 2016

Abstract

In this supplementary material to the authors' paper 'A Wild Bootstrap Approach for Nonparametric Repeated Measurements' we present additional simulation results for both ordinal and continuous data with different α levels ($\alpha = 1\%$ and $\alpha = 10\%$). The results obtained are similar to the ones found at 5% level.

^a Institute of Statistics, Ulm University, Helmholtzstr. 20, 89081 Ulm, Germany

^b The University of Texas at Dallas, 800 W. Campbell Road, Richardson, Texas 75080-3021,
USA

* Corresponding author: markus.pauly@uni-ulm.de.

8 Additional simulation results

The simulation setting is the same as in Section 4. We analyzed both ordinal and continuous (normal and lognormal) data for the nominal level of $\alpha = 1\%$ and $\alpha = 10\%$, respectively. The results are presented in Tables 8 – 13, demonstrating a similar behavior as already observed for the 5%-level: The wild bootstrap versions of the two test statistics improve their behavior, showing a more accurate type-1 error rate control. For $\alpha = 1\%$ the difference between both wild bootstrap tests are marginal, whereas for $\alpha = 10\%$ the wild bootstrap ATS seems to be the method of choice since the bootstrap WTS sometimes results in rejection frequencies that are slightly larger than the nominal 10%.

8.1 Ordinal Data

Table 8: Simulation results ($\alpha = 1\%$) 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	n	Method	$t = 4$		$t = 8$	
			ATS	WTS	ATS	WTS
A	$n_1 = n_2 = 10$	classic	0.020	0.020	0.021	0.021
		Wild Bootstrap	0.012	0.012	0.013	0.013
	$n_1 = 10, n_2 = 20$	classic	0.020	0.020	0.021	0.021
		Wild Bootstrap	0.013	0.013	0.015	0.015
T	$n_1 = n_2 = 10$	classic	0.012	0.048	0.006	0.188
		Wild Bootstrap	0.011	0.011	0.010	0.012
	$n_1 = 10, n_2 = 20$	classic	0.012	0.043	0.006	0.151
		Wild Bootstrap	0.012	0.013	0.011	0.014
A:T	$n_1 = n_2 = 10$	classic	0.012	0.046	0.006	0.187
		Wild Bootstrap	0.011	0.011	0.010	0.011
	$n_1 = 10, n_2 = 20$	classic	0.013	0.043	0.006	0.150
		Wild Bootstrap	0.013	0.013	0.011	0.014

Table 9: Simulation results ($\alpha = 10\%$) 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	\mathbf{n}	Method	$t = 4$		$t = 8$	
			ATS	WTS	ATS	WTS
A	$n_1 = n_2 = 10$	classic	0.116	0.116	0.115	0.115
		Wild Bootstrap	0.100	0.100	0.099	0.099
	$n_1 = 10, n_2 = 20$	classic	0.118	0.118	0.116	0.116
		Wild Bootstrap	0.103	0.103	0.101	0.101
T	$n_1 = n_2 = 10$	classic	0.107	0.183	0.085	0.399
		Wild Bootstrap	0.102	0.102	0.097	0.099
	$n_1 = 10, n_2 = 20$	classic	0.104	0.174	0.086	0.365
		Wild Bootstrap	0.100	0.105	0.098	0.112
A:T	$n_1 = n_2 = 10$	classic	0.105	0.180	0.084	0.398
		Wild Bootstrap	0.100	0.100	0.097	0.098
	$n_1 = 10, n_2 = 20$	classic	0.105	0.176	0.085	0.361
		Wild Bootstrap	0.101	0.106	0.098	0.111

8.2 Continuous Data

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

Cov. Setting	\mathbf{n}	Method	$t = 4$		$t = 8$		
			ATS	WTS	ATS	WTS	
A	$\mathbf{n}^{(1)}$	classic	0.018	0.019	0.020	0.020	
		Wild Bootstrap	0.012	0.012	0.012	0.012	
	$\mathbf{n}^{(2)}$	classic	0.018	0.019	0.019	0.020	
		Wild Bootstrap	0.013	0.013	0.013	0.013	
	CS	$\mathbf{n}^{(1)}$	classic	0.018	0.019	0.020	0.021
		Wild Bootstrap	0.011	0.011	0.012	0.012	
$\mathbf{n}^{(2)}$	classic	0.018	0.019	0.021	0.021		
	Wild Bootstrap	0.013	0.013	0.013	0.013		
T	$\mathbf{n}^{(1)}$	classic	0.015	0.045	0.013	0.189	
		Wild Bootstrap	0.012	0.010	0.013	0.011	
	$\mathbf{n}^{(2)}$	classic	0.015	0.040	0.013	0.149	
		Wild Bootstrap	0.014	0.012	0.014	0.014	
	CS	$\mathbf{n}^{(1)}$	classic	0.011	0.045	0.006	0.187
		Wild Bootstrap	0.011	0.011	0.010	0.010	
$\mathbf{n}^{(2)}$	classic	0.012	0.040	0.006	0.144		
	Wild Bootstrap	0.012	0.012	0.011	0.013		
A:T	$\mathbf{n}^{(1)}$	classic	0.015	0.045	0.013	0.187	
		Wild Bootstrap	0.013	0.011	0.014	0.011	
	$\mathbf{n}^{(2)}$	classic	0.015	0.040	0.013	0.148	
		Wild Bootstrap	0.013	0.013	0.013	0.013	
	CS	$\mathbf{n}^{(1)}$	classic	0.011	0.045	0.006	0.186
		Wild Bootstrap	0.011	0.011	0.010	0.010	
$\mathbf{n}^{(2)}$	classic	0.012	0.041	0.006	0.143		
	Wild Bootstrap	0.012	0.013	0.011	0.013		

Table 11: Type-1 error simulation results for normally distributed data with sample sizes $\mathbf{n}^{(1)} = (10, 10)$ and $\mathbf{n}^{(2)} = (10, 20)$, $\alpha = 10\%$.

Cov. Setting	\mathbf{n}	Method	$t = 4$		$t = 8$		
			ATS	WTS	ATS	WTS	
A	$\mathbf{n}^{(1)}$	classic	0.114	0.116	0.117	0.118	
		Wild Bootstrap	0.100	0.100	0.102	0.102	
	$\mathbf{n}^{(2)}$	classic	0.112	0.113	0.117	0.118	
		Wild Bootstrap	0.099	0.099	0.103	0.103	
	CS	$\mathbf{n}^{(1)}$	classic	0.115	0.117	0.118	0.118
		Wild Bootstrap	0.101	0.101	0.101	0.101	
	$\mathbf{n}^{(2)}$	classic	0.112	0.114	0.118	0.118	
	Wild Bootstrap	0.100	0.100	0.104	0.104		
T	$\mathbf{n}^{(1)}$	classic	0.103	0.178	0.095	0.400	
		Wild Bootstrap	0.100	0.099	0.102	0.098	
	$\mathbf{n}^{(2)}$	classic	0.105	0.173	0.095	0.361	
		Wild Bootstrap	0.102	0.103	0.102	0.111	
	CS	$\mathbf{n}^{(1)}$	classic	0.102	0.177	0.083	0.403
		Wild Bootstrap	0.099	0.100	0.097	0.097	
	$\mathbf{n}^{(2)}$	classic	0.103	0.172	0.084	0.359	
	Wild Bootstrap	0.100	0.103	0.098	0.110		
A:T	$\mathbf{n}^{(1)}$	classic	0.102	0.179	0.093	0.402	
		Wild Bootstrap	0.099	0.100	0.100	0.098	
	$\mathbf{n}^{(2)}$	classic	0.105	0.173	0.093	0.361	
		Wild Bootstrap	0.102	0.103	0.100	0.109	
	CS	$\mathbf{n}^{(1)}$	classic	0.102	0.179	0.084	0.402
		Wild Bootstrap	0.098	0.099	0.097	0.097	
	$\mathbf{n}^{(2)}$	classic	0.103	0.172	0.082	0.356	
	Wild Bootstrap	0.100	0.103	0.097	0.108		

Table 12: Type-1 error simulation results for log-normally distributed data with sample sizes $\mathbf{n}^{(1)} = (10, 10)$ and $\mathbf{n}^{(2)} = (10, 20)$, $\alpha = 1\%$.

Cov. Setting	\mathbf{n}	Method	$t = 4$		$t = 8$	
			ATS	WTS	ATS	WTS
A	AR	$\mathbf{n}^{(1)}$ classic	0.020	0.021	0.020	0.020
		$\mathbf{n}^{(1)}$ Wild Bootstrap	0.012	0.012	0.012	0.012
		$\mathbf{n}^{(2)}$ classic	0.019	0.020	0.020	0.020
		$\mathbf{n}^{(2)}$ Wild Bootstrap	0.013	0.013	0.013	0.013
	CS	$\mathbf{n}^{(1)}$ classic	0.020	0.021	0.021	0.021
		$\mathbf{n}^{(1)}$ Wild Bootstrap	0.012	0.012	0.013	0.013
		$\mathbf{n}^{(2)}$ classic	0.019	0.020	0.021	0.021
		$\mathbf{n}^{(2)}$ Wild Bootstrap	0.013	0.013	0.014	0.014
T	AR	$\mathbf{n}^{(1)}$ classic	0.017	0.050	0.015	0.195
		$\mathbf{n}^{(1)}$ Wild Bootstrap	0.013	0.012	0.014	0.012
		$\mathbf{n}^{(2)}$ classic	0.017	0.047	0.016	0.157
		$\mathbf{n}^{(2)}$ Wild Bootstrap	0.015	0.015	0.015	0.015
	CS	$\mathbf{n}^{(1)}$ classic	0.011	0.052	0.005	0.204
		$\mathbf{n}^{(1)}$ Wild Bootstrap	0.010	0.013	0.008	0.014
		$\mathbf{n}^{(2)}$ classic	0.011	0.046	0.005	0.157
		$\mathbf{n}^{(2)}$ Wild Bootstrap	0.011	0.015	0.010	0.017
A:T	AR	$\mathbf{n}^{(1)}$ classic	0.017	0.046	0.016	0.188
		$\mathbf{n}^{(1)}$ Wild Bootstrap	0.014	0.011	0.014	0.010
		$\mathbf{n}^{(2)}$ classic	0.017	0.043	0.016	0.151
		$\mathbf{n}^{(2)}$ Wild Bootstrap	0.014	0.013	0.015	0.014
	CS	$\mathbf{n}^{(1)}$ classic	0.011	0.045	0.005	0.182
		$\mathbf{n}^{(1)}$ Wild Bootstrap	0.011	0.010	0.010	0.009
		$\mathbf{n}^{(2)}$ classic	0.012	0.043	0.006	0.140
		$\mathbf{n}^{(2)}$ Wild Bootstrap	0.011	0.013	0.011	0.011

Table 13: Type-1 error simulation results for log-normally distributed data with sample sizes $\mathbf{n}^{(1)} = (10, 10)$ and $\mathbf{n}^{(2)} = (10, 20)$, $\alpha = 10\%$.

Cov. Setting	\mathbf{n}	Method	$t = 4$		$t = 8$	
			ATS	WTS	ATS	WTS
A	AR	$\mathbf{n}^{(1)}$ classic	0.116	0.117	0.117	0.117
		Wild Bootstrap	0.100	0.100	0.101	0.101
		$\mathbf{n}^{(2)}$ classic	0.113	0.115	0.118	0.118
		Wild Bootstrap	0.100	0.100	0.103	0.103
	CS	$\mathbf{n}^{(1)}$ classic	0.115	0.117	0.117	0.118
		Wild Bootstrap	0.101	0.101	0.101	0.101
		$\mathbf{n}^{(2)}$ classic	0.114	0.115	0.118	0.119
		Wild Bootstrap	0.100	0.100	0.104	0.104
T	AR	$\mathbf{n}^{(1)}$ classic	0.107	0.186	0.100	0.408
		Wild Bootstrap	0.102	0.106	0.103	0.105
		$\mathbf{n}^{(2)}$ classic	0.108	0.181	0.100	0.370
		Wild Bootstrap	0.106	0.111	0.104	0.117
	CS	$\mathbf{n}^{(1)}$ classic	0.101	0.184	0.080	0.419
		Wild Bootstrap	0.098	0.106	0.093	0.113
		$\mathbf{n}^{(2)}$ classic	0.102	0.179	0.081	0.372
		Wild Bootstrap	0.099	0.110	0.095	0.124
A:T	AR	$\mathbf{n}^{(1)}$ classic	0.105	0.179	0.098	0.402
		Wild Bootstrap	0.101	0.100	0.101	0.097
		$\mathbf{n}^{(2)}$ classic	0.106	0.175	0.098	0.364
		Wild Bootstrap	0.103	0.105	0.102	0.112
	CS	$\mathbf{n}^{(1)}$ classic	0.103	0.179	0.083	0.402
		Wild Bootstrap	0.099	0.099	0.097	0.094
		$\mathbf{n}^{(2)}$ classic	0.103	0.175	0.081	0.358
		Wild Bootstrap	0.100	0.105	0.095	0.108

Article 3

Friedrich, S. and Pauly, M. (2017). MATS: Inference for potentially singular and heteroscedastic MANOVA. *arXiv preprint arXiv:1704.03731*.

MATS: Inference for potentially Singular and Heteroscedastic MANOVA

Sarah Friedrich, Markus Pauly

Institute of Statistics, Ulm University, Germany

Abstract

In many experiments in the life sciences, several endpoints are recorded per subject. The analysis of such multivariate data is usually based on MANOVA models assuming multivariate normality and covariance homogeneity. These assumptions, however, are often not met in practice. Furthermore, test statistics should be invariant under scale transformations of the data, since the endpoints may be measured on different scales. In the context of high-dimensional data, Srivastava and Kubokawa (2013) proposed such a test statistic for a specific one-way model, which, however, relies on the assumption of a common non-singular covariance matrix. We modify and extend this test statistic to factorial MANOVA designs, incorporating general heteroscedastic models. In particular, our only distributional assumption is the existence of the group-wise covariance matrices, which may even be singular. We base inference on quantiles of resampling distributions, and derive confidence regions and ellipsoids based on these quantiles. In a simulation study, we extensively analyze the behavior of these procedures. Finally, the methods are applied to a data set containing information on the 2016 presidential elections in the USA with unequal and singular empirical covariance matrices.

Keywords: Multivariate Data; Parametric Bootstrap; Confidence Regions; Singular Covariance Matrices

Email addresses: sarah.friedrich@uni-ulm.de (Sarah Friedrich), sarah.friedrich@uni-ulm.de (Sarah Friedrich)

Preprint submitted to Elsevier

December 6, 2017

1. Motivation and Introduction

In many experiments in, e.g., biology, ecology and psychology several endpoints, potentially measured on different scales, are recorded per subject. As an example, we consider a data set on the 2016 presidential elections in the USA containing demographic data on counties from US census. For our exemplary analysis, we aim to investigate whether the states differ with respect to some demographic factors. In addition to unequal empirical covariance matrices between groups, analysis is further complicated by their singularity.

The analysis of such multivariate data is typically based on classical MANOVA models assuming multivariate normality and/or homogeneity of the covariance matrices, see, e.g., [1, 14, 15, 20, 25, 34, 44]. These assumptions, however, are often not met in practice (as in the motivating example) and it is well known that the methods perform poorly in case of heterogeneous data [23, 40]. Furthermore, the test statistic should be invariant under scale transformations of the components, since the endpoints may be measured on different scales. Thus, test statistics of multivariate ANOVA-type (ATS) as, e.g., proposed in [5] and studied in [16], are only applicable if all endpoints are measured on the same scale, i.e., for repeated measures designs. Assuming non-singular covariance matrices and certain moment assumptions the scale invariance is typically accomplished by utilizing test statistics of Wald-type (WTS). However, inference procedures based thereon require (extremely) large sample sizes for being accurate, see [23, 37, 41]. In particular, even the novel approaches of [23] and [37] showed a more or less liberal behavior in case of skewed distributions. Moreover, their procedures cannot be used to analyze the motivating data example with possibly singular covariance matrices. Therefore, we follow a different approach by modifying the above mentioned ANOVA-type statistic (MATS). It is motivated from the modified Dempster statistic proposed in [39] for high-dimensional one-way MANOVA. This statistic is also invariant under the change of units of measurements. However, until now, it has only been developed for a homoscedastic one-way setting assuming non-singularity and a specific distributional structure that is motivated from multivariate normality.

It is the aim of the present paper to modify and extend the [39] approach to factorial MANOVA designs, incorporating general heteroscedastic models. In particular, we only postulate existence of the group-wise covariance matrices, which may even be singular. The small sample behavior of our test statistic is enhanced by applying bootstrap techniques as in [23]. Thereby, the novel MATS procedure enables us to relax the usual MANOVA assumptions in several ways: While incorporating general heteroscedastic designs and allowing for potentially singular covariance matrices we postulate their existence as solely distributional assumption, i.e., only finite second moments are required. Moreover, we gain a procedure that is more robust against deviations from symmetry and homoscedasticity than the usual WTS approaches.

So far, only few approaches have been investigated which do not assume normality or equal covariance matrices (or both). Examples in the nonparametric framework include the permutation based nonparametric combination method [32, 33] and the rank-based tests presented in [6] and [7] for Split Plot Designs and in [2] and [27] for MANOVA designs. However, these methods are either not applicable for general MANOVA models or based on null hypotheses formulated in terms of distribution functions. In contrast we wish to derive inference procedures (tests and confidence regions) for contrasts of mean vectors. Here, beside all previously mentioned procedures, only methods for specific designs have been developed, see [11] for two-sample problems, [42, 43] for robust but homoscedastic one-way MANOVA or [18] for a particular two-way MANOVA. To our knowledge, mean-based MANOVA procedures allowing for possibly singular covariance matrices have not been developed so far.

The paper is organized as follows: In Section 2 we describe the statistical model and hypotheses. Furthermore, we propose a new test statistic, which is applicable to singular covariance matrices and is invariant under scale transformations of the data. In Section 3, we present three different resampling approaches, which are used for the derivation of statistical tests as well as confidence regions and simultaneous confidence intervals for contrasts in Section 4. The different approaches are compared in a large simulation study (Section 5), where we analyze different factorial designs with a wide variety of distributions and covariance settings. The motivating data example is analyzed in Section 6 and we conclude with some final remarks and discussion in Section 7. All proofs are deferred to the supplementary material, where we also provide further simulation results and the analysis of an additional data example.

2. Statistical Model, Hypotheses and Test Statistics

Throughout the paper, we will use the following notation. We denote by \mathbf{I}_d the d -dimensional unit matrix and by \mathbf{J}_d the $d \times d$ matrix of 1's, i.e., $\mathbf{J}_d = \mathbf{1}_d \mathbf{1}_d^\top$, where $\mathbf{1}_d = (1, \dots, 1)^\top$ is the d -dimensional column vector of 1's. Furthermore, let $\mathbf{P}_d = \mathbf{I}_d - d^{-1} \mathbf{J}_d$ denote the d -dimensional centering matrix. By \oplus and \otimes we denote the direct sum and the Kronecker product, respectively.

In order to cover different factorial designs of interest, we establish the general model

$$\mathbf{X}_{ik} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ik}$$

for treatment group $i = 1, \dots, a$ and individual $k = 1, \dots, n_i$, on which we measure d -variate observations. Here $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{id})^\top \in \mathbb{R}^d$ for $i = 1, \dots, a$. A factorial structure can be incorporated by splitting up indices, see, e.g., [23]. For fixed $1 \leq i \leq a$, the error terms $\boldsymbol{\epsilon}_{ik}$ are assumed to be independent and identically distributed d -dimensional random vectors, for which the following conditions hold:

- (1) $E(\boldsymbol{\epsilon}_{i1}) = \mathbf{0}$, $i = 1, \dots, a$,
- (2) $0 < \sigma_{is}^2 = \text{Var}(X_{iks}) < \infty$, $i = 1, \dots, a$, $s = 1, \dots, d$,
- (3) $\text{Cov}(\boldsymbol{\epsilon}_{i1}) = \mathbf{V}_i \geq \mathbf{0}$, $i = 1, \dots, a$.

Thus, we only assume the existence of second moments. For convenience, we aggregate the individual vectors into $\mathbf{X} = (\mathbf{X}_{11}^\top, \dots, \mathbf{X}_{an_a}^\top)^\top$ as well as $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_a^\top)^\top$. Denote by $N = \sum_{i=1}^a n_i$ the total sample size. In order to derive asymptotic results in this framework, we will throughout assume the usual sample size condition:

$$n_i/N \rightarrow \kappa_i > 0, i = 1, \dots, a$$

as $N \rightarrow \infty$.

An estimator for $\boldsymbol{\mu}$ is given by the vector of pooled group means $\bar{\mathbf{X}}_i = n_i^{-1} \sum_{k=1}^{n_i} \mathbf{X}_{ik}$, $i = 1, \dots, a$, which we denote by $\bar{\mathbf{X}}_\bullet = (\bar{\mathbf{X}}_1^\top, \dots, \bar{\mathbf{X}}_a^\top)^\top$. The covariance matrix of $\sqrt{N} \bar{\mathbf{X}}_\bullet$ is given by

$$\boldsymbol{\Sigma}_N = \text{Cov}(\sqrt{N} \bar{\mathbf{X}}_\bullet) = \text{diag} \left(\frac{N}{n_i} \mathbf{V}_i : 1 \leq i \leq a \right),$$

where the group-specific covariance matrices \mathbf{V}_i are estimated by the empirical covariance matrices

$$\widehat{\mathbf{V}}_i = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (\mathbf{X}_{ik} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ik} - \bar{\mathbf{X}}_i)^\top$$

resulting in

$$\widehat{\boldsymbol{\Sigma}}_N = \text{diag} \left(\frac{N}{n_i} \widehat{\mathbf{V}}_i : 1 \leq i \leq a \right).$$

In this semi-parametric framework, hypotheses are formulated in terms of the mean vector as $H_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, where \mathbf{H} is a suitable contrast matrix, i.e., $\mathbf{H}\mathbf{1}_{ad} = \mathbf{0}$. Note that we can use the unique projection matrix $\mathbf{T} = \mathbf{H}^\top (\mathbf{H}\mathbf{H}^\top)^+ \mathbf{H}$, where $(\mathbf{H}\mathbf{H}^\top)^+$ denotes the Moore-Penrose inverse of $\mathbf{H}\mathbf{H}^\top$. It is $\mathbf{T} = \mathbf{T}^2$, $\mathbf{T} = \mathbf{T}^\top$ and $\mathbf{T}\boldsymbol{\mu} = \mathbf{0} \Leftrightarrow \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, see, e.g., [8].

A commonly used test statistic for multivariate data is the Wald-type statistic (WTS)

$$T_N = N \bar{\mathbf{X}}_\bullet^\top \mathbf{T} (\mathbf{T} \widehat{\boldsymbol{\Sigma}}_N \mathbf{T})^+ \mathbf{T} \bar{\mathbf{X}}_\bullet, \quad (2.1)$$

which requires the additional assumption (3') $V_i > 0$, $i = 1, \dots, a$. It is easy to show that the WTS has under $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, asymptotically, as $N \rightarrow \infty$, a χ_f^2 -distribution with $f = \text{rank}(\mathbf{T})$ degrees of freedom if (1) – (3') holds. However, large sample sizes are necessary to maintain a pre-assigned level α using quantiles of the limiting χ^2 -distribution. [23] proposed different resampling procedures in order to improve the small sample behavior of the WTS for multivariate data. Therein, a parametric bootstrap approach turned out to be the best in case that the underlying distributions are not too skewed and/or too heteroscedastic. In the latter cases all considered procedures were more or less liberal. Moreover, assuming only (3) instead of (3') for the WTS would in general not lead to an asymptotic χ_f^2 -limit distribution. The reason for this are possible rank jumps between $\widehat{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}_N\mathbf{T}$, $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}$ and \mathbf{T} . To accept this, suppose that $\text{rank}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}) = 2$, while $\text{rank}(\mathbf{T}) = 4$ (this corresponds to Scenario S5 in the simulation studies below). If additionally $\lim_{N \rightarrow \infty} \text{rank}(\widehat{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}_N\mathbf{T}) = \text{rank}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}) = 2$, we have that the WTS follows, asymptotically, a χ_g^2 -distribution under the null hypothesis, where $g = \text{rank}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}) = 2$. The Wald-type test, however, compares T_N to the quantile of a χ_4^2 -distribution. Thus, for a chosen significance level of $\alpha = 0.05$ this results in a true asymptotic ($N \rightarrow \infty$) type-I error of

$$\Pr(T_N > \chi_{4;0.95}^2) = 1 - \Pr(T_N \leq \chi_{4;0.95}^2) \approx 0.0087,$$

i.e., a strictly conservative behavior of the test. Here $\chi_{f;1-\alpha}^2$ denotes the $(1 - \alpha)$ -quantile of the χ_f^2 -distribution. Similarly, for $\alpha = 0.1$ and $\alpha = 0.01$ we obtain asymptotically inflated type-I error rates of 0.02 and 0.0013 (both again conservative), respectively. Moreover, the situation is even more complicated since $\lim_{N \rightarrow \infty} \text{rank}(\widehat{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}_N\mathbf{T}) = \text{rank}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T})$ is neither verifiable in practice nor holds in general.

We tackle this problem in the current paper, where we not only relax the assumption (3') on the unknown covariance matrices but also gain a procedure that is more robust to deviations from symmetry and homoscedasticity. To this end, we consider a different test statistic, namely a multivariate version of the ANOVA-type statistic (ATS) proposed by [5] for repeated measures designs, which we obtain by erasing the Moore-Penrose term from (2.1):

$$\tilde{Q}_N = N\bar{\mathbf{X}}_{\bullet}^{\top} \mathbf{T} \bar{\mathbf{X}}_{\bullet}.$$

In the special two-sample case where we wish to test the null hypothesis $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$, this is equivalent to the test statistic proposed by [15].

The drawback of the ATS for multivariate data is that it is not invariant under scale transformations of the components, e.g., under change of units ($cm \mapsto m$ or $g \mapsto kg$) in one or more components. We demonstrate this problem in a real data analysis given in the supplementary material, where we exemplify that a simple unit change can completely alter the test decision of the ATS. Thus, we consider a slightly modified version of the ATS, which we denote as MATS:

$$Q_N = N\bar{\mathbf{X}}_{\bullet}^{\top} \mathbf{T} (\mathbf{T} \widehat{\mathbf{D}}_N \mathbf{T})^+ \bar{\mathbf{X}}_{\bullet}. \quad (2.2)$$

Here, $\widehat{\mathbf{D}}_N = \text{diag}(N/n_i \cdot \widehat{\sigma}_{is}^2)$, $i = 1, \dots, a$, $s = 1, \dots, d$, where $\widehat{\sigma}_{is}^2$ is the empirical variance of component s in group i . A related test statistic has been proposed by [39] in the context of high-dimensional ($d \rightarrow \infty$) data for a special non-singular one-way MANOVA design. Here, we investigate in the classical multivariate case (with fixed d) how its finite sample performance can be enhanced considerably. We start by analyzing its asymptotic limit behavior.

THEOREM 2.1. *Under Conditions (1), (2) and (3) and under $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, the test statistic Q_N in (2.2) has asymptotically, as $N \rightarrow \infty$, the same distribution as a weighted sum of χ_1^2 distributed random variables, where the weights λ_{is} are the eigenvalues of $\mathbf{V} = \mathbf{T}(\mathbf{T}\mathbf{D}\mathbf{T})^+ \mathbf{T}\boldsymbol{\Sigma}$ for $\mathbf{D} = \text{diag}(\kappa_i^{-1} \sigma_{is}^2)$ and $\boldsymbol{\Sigma} = \text{diag}(\kappa_i^{-1} \mathbf{V}_i)$, i.e.,*

$$Q_N = N\bar{\mathbf{X}}_{\bullet}^{\top} \mathbf{T} (\mathbf{T} \widehat{\mathbf{D}}_N \mathbf{T})^+ \bar{\mathbf{X}}_{\bullet} \xrightarrow{\mathcal{D}} Z = \sum_{i=1}^a \sum_{s=1}^d \lambda_{is} Z_{is},$$

with $Z_{is} \stackrel{i.i.d.}{\sim} \chi_1^2$ and " $\xrightarrow{\mathcal{D}}$ " denoting convergence in distribution.

Thus, we obtain an asymptotic level α benchmark test $\varphi_N = \mathbb{1}\{Q_N > c_{1-\alpha}\}$ for $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, where $c_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of Z . However, the distribution of Z depends on the unknown variances $\sigma_{is}^2, i = 1, \dots, a, s = 1, \dots, d$ so that φ_N is infeasible for most practical situations. For this reason we consider different bootstrap approaches in order to approximate the unknown limiting distribution and to derive adequate and asymptotically correct inference procedures based on Q_N in (2.2). This will be explained in detail in the next section. Apart from statistical test decisions discussed in Section 4.1, a central part of statistical analyses is the construction of confidence intervals, which allows for deeper insight into the variability and the magnitude of effects. This univariate concept can be generalized to multivariate endpoints by constructing multivariate confidence regions and simultaneous confidence intervals for contrasts $\mathbf{h}^\top \boldsymbol{\mu}$ for any contrast vector $\mathbf{h} \in \mathbb{R}^{ad}$ of interest. Details on the derivation of such confidence regions for $\mathbf{h}^\top \boldsymbol{\mu}$ are given in Section 4.2 below.

3. Bootstrap Procedures

The first bootstrap procedure we consider is a parametric bootstrap approach as proposed by [23] for the WTS. The second one is a wild bootstrap approach, which has already been successfully applied in the context of repeated measures or clustered data, see [9, 10] or [17]. The third procedure is a group-wise, nonparametric bootstrap approach. All of these bootstrap approaches are based on the test statistic Q_N in (2.2). Note that the procedures derived in the following can also be used for multiple testing problems, either by applying the closed testing principle [30, 38] or in the context of simultaneous contrast tests [19, 21].

3.1. A Parametric Bootstrap Approach

This asymptotic model based bootstrap approach has successfully been used in univariate one- and two-way factorial designs ([24, 46]), and has recently been applied to Wald-type statistics for general MANOVA by [23] and [37], additionally assuming finite fourth moments. The approach is as follows: Given the data, we generate a parametric bootstrap sample as

$$\mathbf{X}_{i1}^*, \dots, \mathbf{X}_{in_i}^* \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \widehat{\mathbf{V}}_i), \quad i = 1, \dots, a.$$

The idea behind this method is to obtain an accurate finite sample approximation by mimicking the covariance structure given in the observed data. This is achieved by calculating Q_N^* from the bootstrap variables $\mathbf{X}_{i1}^*, \dots, \mathbf{X}_{in_i}^*$, i.e.,

$$Q_N^* = N(\overline{\mathbf{X}}_\bullet^*)^\top \mathbf{T}(\mathbf{T}\widehat{\mathbf{D}}_N^* \mathbf{T})^+ \mathbf{T}\overline{\mathbf{X}}_\bullet^*. \quad (3.1)$$

We then obtain a parametric bootstrap test by comparing the original test statistic Q_N with the conditional $(1 - \alpha)$ -quantile $c_{1-\alpha}^*$ of its bootstrap version Q_N^* .

THEOREM 3.1. *The conditional distribution of Q_N^* weakly approximates the null distribution of Q_N in probability for any parameter $\boldsymbol{\mu} \in \mathbb{R}^{ad}$ and $\boldsymbol{\mu}_0$ with $\mathbf{T}\boldsymbol{\mu}_0 = \mathbf{0}$, i.e.,*

$$\sup_{x \in \mathbb{R}} |\Pr_{\boldsymbol{\mu}}(Q_N^* \leq x | \mathbf{X}) - \Pr_{\boldsymbol{\mu}_0}(Q_N \leq x)| \xrightarrow{\Pr} 0,$$

where $\Pr_{\boldsymbol{\mu}}(Q_N \leq x)$ and $\Pr_{\boldsymbol{\mu}}(Q_N^* \leq x | \mathbf{X})$ denote the unconditional and conditional distribution of Q_N and Q_N^* , respectively, if $\boldsymbol{\mu}$ is the true underlying mean vector.

3.2. A Wild Bootstrap Approach

Another resampling approach, which is based on multiplying the fixed data with random weights, is the so-called wild bootstrap procedure. To this end, let W_{ik} denote i.i.d. random variables, independent of \mathbf{X} , with $E(W_{ik}) = 0$, $\text{Var}(W_{ik}) = 1$ and $\sup_{i,k} E(W_{ik}^4) < \infty$. We obtain a bootstrap sample as

$$\mathbf{X}_{ik}^* = W_{ik}(\mathbf{X}_{ik} - \bar{\mathbf{X}}_i), i = 1, \dots, a, k = 1, \dots, n_i.$$

Note that there are different choices for the random weights W_{ik} , e.g., Rademacher distributed random variables [13] or weights satisfying different moment conditions, see, e.g., [3, 28, 29, 45]. The choice of the weights typically depends on the situation. In our simulation studies, we have investigated the performance of different weights such as Rademacher distributed as well as $\mathcal{N}(0, 1)$ distributed weights (see [26] for this specific choice). The results were rather similar and we therefore only present the results of our simulation study for standard normally distributed weights in Section 5 below.

Based on the bootstrap variables \mathbf{X}_{ik}^* , we can calculate Q_N^* in the same way as described for Q_N^* in (3.1) above. A wild bootstrap test is finally obtained by comparing Q_N to the conditional $(1 - \alpha)$ -quantile of its wild bootstrap version Q_N^* .

THEOREM 3.2. *The conditional distribution of Q_N^* weakly approximates the null distribution of Q_N in probability for any parameter $\boldsymbol{\mu} \in \mathbb{R}^{ad}$ and $\boldsymbol{\mu}_0$ with $\mathbf{T}\boldsymbol{\mu}_0 = \mathbf{0}$, i.e.,*

$$\sup_{x \in \mathbb{R}} |\Pr(Q_N^* \leq x | \mathbf{X}) - \Pr(Q_N \leq x)| \xrightarrow{\Pr} 0.$$

3.3. A nonparametric bootstrap approach

The third approach we consider is the nonparametric bootstrap. Here, for each group $i = 1, \dots, a$, we randomly draw n_i independent selections \mathbf{X}_{ik}^\dagger with replacement from the i -th sample $\{\mathbf{X}_{i1}, \dots, \mathbf{X}_{in_i}\}$. The bootstrap test statistic Q_N^\dagger is then calculated in the same way as described above, i.e., by recalculating Q_N with $\mathbf{X}_{ik}^\dagger, i = 1, \dots, a, k = 1, \dots, n_i$. Finally, a nonparametric bootstrap test is obtained by comparing the original test statistic Q_N to the empirical $(1 - \alpha)$ -quantile of Q_N^\dagger . The asymptotic validity of this method is guaranteed by

THEOREM 3.3. *The conditional distribution of Q_N^\dagger weakly approximates the null distribution of Q_N in probability for any parameter $\boldsymbol{\mu} \in \mathbb{R}^{ad}$ and $\boldsymbol{\mu}_0$ with $\mathbf{T}\boldsymbol{\mu}_0 = \mathbf{0}$, i.e.,*

$$\sup_{x \in \mathbb{R}} |\Pr(Q_N^\dagger \leq x | \mathbf{X}) - \Pr(Q_N \leq x)| \xrightarrow{\Pr} 0.$$

4. Statistical Applications

We now want to base statistical inference on the modified test statistic in (2.2) using the bootstrap approaches described above. A thorough statistical analysis should ideally consist of two parts: First, statistical tests give insight into significant effects of the different factors as well as possible interactions. We therefore consider important properties of statistical tests based on the bootstrap approaches in Section 4.1. Second, it is helpful to construct confidence regions for the unknown parameters of interest in order to gain a more detailed insight into the nature of the estimates. The derivation of such confidence regions is discussed in Section 4.2.

4.1. Statistical Tests

In this section, we analyze the statistical properties of the bootstrap procedures described above. For ease of notation, we will only state the results for the parametric bootstrap procedure, i.e., consider the test statistic Q_N^* based on X_{ik}^* throughout. Note, however, that the results are also valid for the wild and the nonparametric bootstrap procedure, i.e., the test statistics Q_N^* and Q_N^\dagger .

As mentioned above, a bootstrap test $\varphi^* = \mathbb{1}\{Q_N > c_{1-\alpha}^*\}$ is obtained by comparing the original test statistic Q_N to the $(1 - \alpha)$ -quantile $c_{1-\alpha}^*$ of its bootstrap version. In particular, p -values are numerically computed as follows:

- (1) Given the data X , calculate the MATS Q_N for the null hypothesis of interest.
- (2) Bootstrap the data with either of the bootstrap approaches described above and calculate the corresponding test statistic $Q_N^{*,1}$.
- (3) Repeat step (2) a large number of times, e.g., $B = 10,000$ times, and obtain values $Q_N^{*,1}, \dots, Q_N^{*,B}$.
- (4) Calculate the p -value based on the empirical distribution of $Q_N^{*,1}, \dots, Q_N^{*,B}$ as

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^B \mathbb{1}\{Q_N \leq Q_N^{*,b}\}.$$

Theorems 3.1 – 3.3 imply that the corresponding tests asymptotically keep the pre-assigned level α under the null hypothesis and are consistent for any fixed alternative $T\boldsymbol{\mu} \neq \mathbf{0}$, i.e., $E_\mu(\varphi^*) \rightarrow \alpha \cdot \mathbb{1}\{T\boldsymbol{\mu} = \mathbf{0}\} + \mathbb{1}\{T\boldsymbol{\mu} \neq \mathbf{0}\}$. Moreover, for local alternatives $H_1 : T\boldsymbol{\mu} = \sqrt{N}^{-1}T\boldsymbol{\nu}$, $\boldsymbol{\nu} \in \mathbb{R}^{ad}$, the bootstrap tests have the same asymptotic power as $\varphi_N = \mathbb{1}\{Q_N > c_{1-\alpha}\}$, where $c_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of Z given in Theorem 2.1. In particular, the asymptotic relative efficiency of the bootstrap tests compared to φ_N is 1 in this situation.

4.2. Confidence regions and confidence intervals for contrasts

In order to conduct a thorough statistical analysis, interpretation of the results should not be based on p -values alone. In addition, it is helpful to construct confidence regions for the unknown parameter. The concept of a confidence region is the same as that of a confidence interval in the univariate setting: We want to construct a multivariate region, which is likely to contain the true, but unknown parameter of interest. The aim of this section is to derive multivariate confidence regions and simultaneous confidence intervals for contrasts $\mathbf{h}^\top \boldsymbol{\mu}$ for any contrast vector \mathbf{h} of interest. Such contrasts include, e.g., the difference in means $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ in two-sample problems, Dunnett's many-to-one comparisons, Tukey's all-pairwise comparisons, and many more, see, e.g., [21] for specific examples. In this section, we will base the derivation of confidence regions on the bootstrap approximations given in Section 3, i.e., we will use one of the bootstrap quantiles. Again, we only formulate the results for $c_{1-\alpha}^*$.

For the derivation of a confidence region, first note that the results from Section 4.1 imply that the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\mu}_0$ for a vector of contrasts $\mathbf{H}\boldsymbol{\mu}_0$, $\mathbf{H} = (\mathbf{h}_1 | \dots | \mathbf{h}_q)^\top \in \mathbb{R}^{q \times ad}$, $\boldsymbol{\mu}_0 \in \mathbb{R}^{ad}$, is rejected at asymptotic level α , if $N(\mathbf{H}\bar{\mathbf{X}}_\bullet - \mathbf{H}\boldsymbol{\mu}_0)^\top (\mathbf{H}\widehat{\mathbf{D}}_N \mathbf{H}^\top)^+ (\mathbf{H}\bar{\mathbf{X}}_\bullet - \mathbf{H}\boldsymbol{\mu}_0)$ is larger than the bootstrap quantile $c_{1-\alpha}^*$. Thus, a confidence region for the vector of contrasts $\mathbf{H}\boldsymbol{\mu}$ is determined by the set of all $\mathbf{H}\boldsymbol{\mu}$ such that

$$N(\mathbf{H}\bar{\mathbf{X}}_\bullet - \mathbf{H}\boldsymbol{\mu})^\top (\mathbf{H}\widehat{\mathbf{D}}_N \mathbf{H}^\top)^+ (\mathbf{H}\bar{\mathbf{X}}_\bullet - \mathbf{H}\boldsymbol{\mu}) \leq c_{1-\alpha}^*.$$

A confidence ellipsoid is now obtained based on the eigenvalues $\widehat{\lambda}_s$ and eigenvectors $\widehat{\mathbf{e}}_s$ of $\mathbf{H}\widehat{\mathbf{D}}_N \mathbf{H}^\top$. As in [22], the direction and lengths of its axes are determined by going $\sqrt{\widehat{\lambda}_s \cdot c_{1-\alpha}^* / N}$ units along the eigenvectors $\widehat{\mathbf{e}}_s$ of $\mathbf{H}\widehat{\mathbf{D}}_N \mathbf{H}^\top$. In other words, the axes of the ellipsoid are given by

$$\mathbf{H}\bar{\mathbf{X}}_\bullet \pm \sqrt{\widehat{\lambda}_s \cdot c_{1-\alpha}^* / N} \cdot \widehat{\mathbf{e}}_s, \quad s = 1, \dots, d. \quad (4.1)$$

Note that this approach is similar to the construction of confidence intervals in the univariate case, where we exploit the one-to-one relationship between CIs and tests. While we can calculate (4.1) for arbitrary dimensions d , we cannot display the joint confidence region graphically for $d \geq 4$. In the two-sample case with $d = 2$ endpoints, however, the ellipse can be plotted: Beginning at the center $\mathbf{H}\bar{\mathbf{X}}_\bullet$, the axes of the ellipsoid are given by $\pm \sqrt{\lambda_s \cdot c_{1-\alpha}^*/N} \cdot \widehat{\boldsymbol{\epsilon}}_s$, $s = 1, 2$. That is, the confidence ellipse extends $\sqrt{\lambda_s \cdot c_{1-\alpha}^*/N}$ units along the estimated eigenvector $\widehat{\boldsymbol{\epsilon}}_s$ for $s = 1, 2$. Therefore, we get a graphical representation of the relation between the group-mean differences $\mu_{11} - \mu_{21}$ and $\mu_{12} - \mu_{22}$ of the first and second component, see Section 10 and Figure 6 in the supplementary material for an example.

Concerning the derivation of multiple contrast tests and simultaneous confidence intervals for contrasts, we consider the family of hypotheses

$$\Omega = \{H_0 : \mathbf{h}_\ell^\top \boldsymbol{\mu} = \mathbf{0} \text{ with } \mathbf{h}_\ell \neq \mathbf{0}, \ell = 1, \dots, q\}.$$

As shown in Sections 2 and 3 a test statistic for testing the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ is given by Q_N in (2.2). Consequently, working with a single contrast \mathbf{h}_ℓ as contrast matrix leads to the test statistic

$$Q_N^\ell = N(\mathbf{h}_\ell^\top \bar{\mathbf{X}}_\bullet)^\top (\mathbf{h}_\ell^\top \widehat{\mathbf{D}}_N \mathbf{h}_\ell)^{-1} (\mathbf{h}_\ell^\top \bar{\mathbf{X}}_\bullet) = N \frac{\left(\sum_{i=1}^a \sum_{s=1}^d h_{\ell, is} \bar{X}_{i,s} \right)^2}{\sum_{i=1}^a \sum_{s=1}^d h_{\ell, is}^2 \widehat{\sigma}_{is}^2}$$

for the null hypotheses $H_0^\ell : \mathbf{h}_\ell^\top \boldsymbol{\mu} = \mathbf{0}$, $\ell = 1, \dots, q$. Here, $\mathbf{h}_\ell = (h_{\ell, 11}, \dots, h_{\ell, ad})^\top$. To obtain a single critical value with one of the bootstrap methods we may, e.g., consider the usual maximum or sum statistics. We exemplify the idea for the latter. Thus, let

$$S_N \equiv N(\mathbf{H}\bar{\mathbf{X}}_\bullet)^\top \text{diag} \left((\mathbf{h}_\ell^\top \widehat{\mathbf{D}}_N \mathbf{h}_\ell)^{-1} : \ell = 1, \dots, q \right) \mathbf{H}\bar{\mathbf{X}}_\bullet = \sum_{\ell=1}^q Q_N^\ell$$

and denote by $q_{1-\alpha}^*$ the conditional $(1 - \alpha)$ -quantile of its corresponding bootstrap version S_N^* . From the proofs of Theorem 3.1 – 3.3 given in the supplement it follows for any of the three bootstrap methods described in Section 3 that $\widehat{\sigma}_{is}^*$ are consistent estimates for σ_{is} ($i = 1, \dots, a$, $s = 1, \dots, d$) and that $\sqrt{N}\mathbf{H}\bar{\mathbf{X}}_\bullet^*$ asymptotically mimics the distribution of $\sqrt{N}\mathbf{H}(\bar{\mathbf{X}}_\bullet - \boldsymbol{\mu})$. Thus, the continuous mapping theorem implies $\Pr(S_N \leq q_{1-\alpha}^*) \rightarrow 1 - \alpha$ as $N \rightarrow \infty$ and therefore

$$\Pr \left(\bigcap_{\ell=1}^q \{Q_N^\ell \leq q_{1-\alpha}^*\} \right) \leq \Pr \left(\sum_{\ell=1}^q Q_N^\ell \leq q_{1-\alpha}^* \right) \rightarrow 1 - \alpha, N \rightarrow \infty.$$

This implies, that simultaneous $100(1 - \alpha)\%$ confidence intervals for contrasts $\mathbf{h}_\ell^\top \boldsymbol{\mu}$, $\ell = 1, \dots, q$, are given by

$$\mathbf{h}_\ell^\top \bar{\mathbf{X}}_\bullet \pm \sqrt{q_{1-\alpha}^* \cdot \mathbf{h}_\ell^\top \widehat{\mathbf{D}}_N \mathbf{h}_\ell / N}.$$

In the supplement we additionally explain that the bootstrap idea also works for the usual maximum statistic.

5. Simulations

The procedures described in Section 3 are valid for large sample sizes. In order to investigate their behavior for small samples, we have conducted various simulations. In the simulation studies, the behavior of the proposed approaches was compared to a parametric bootstrap approach for the WTS as in [23] since this turned out to perform better than other resampling versions of the WTS and Wilk's Λ . For comparison, we also included the asymptotic χ^2 approximation of the WTS. All simulations were conducted using R Version 3.3.1 [35] each with $n_{\text{sim}} = 5,000$ simulation and $n_{\text{boot}} = 5,000$ bootstrap runs. We investigated a one- and a two-factorial design.

5.1. One-way layout

For the one-way layout, data was generated as in [23]. We considered $a = 2$ treatment groups and $d \in \{4, 8\}$ endpoints as well as the following covariance settings:

$$\begin{aligned} \text{Setting 1:} \quad & \mathbf{V}_1 = \mathbf{I}_d + 0.5(\mathbf{J}_d - \mathbf{I}_d) = \mathbf{V}_2, \\ \text{Setting 2:} \quad & \mathbf{V}_1 = \left((0.6)^{|r-s|} \right)_{r,s=1}^d = \mathbf{V}_2, \\ \text{Setting 3:} \quad & \mathbf{V}_1 = \mathbf{I}_d + 0.5(\mathbf{J}_d - \mathbf{I}_d) \text{ and } \mathbf{V}_2 = \mathbf{I}_d \cdot 3 + 0.5(\mathbf{J}_d - \mathbf{I}_d), \\ \text{Setting 4:} \quad & \mathbf{V}_1 = \left((0.6)^{|r-s|} \right)_{r,s=1}^d \text{ and } \mathbf{V}_2 = \left((0.6)^{|r-s|} \right)_{r,s=1}^d + \mathbf{I}_d \cdot 2. \end{aligned}$$

Setting 1 represents a compound symmetry structure, while setting 2 is an autoregressive covariance structure. Both settings 1 and 2 represent homoscedastic scenarios while settings 3 and 4 display two scenarios with unequal covariance structures. Data was generated by

$$\mathbf{X}_{ik} = \boldsymbol{\mu}_i + \mathbf{V}_i^{1/2} \boldsymbol{\epsilon}_{ik}, \quad i = 1, \dots, a; \quad k = 1, \dots, n_i,$$

where $\mathbf{V}_i^{1/2}$ denotes the square root of the matrix \mathbf{V}_i , i.e., $\mathbf{V}_i = \mathbf{V}_i^{1/2} \cdot \mathbf{V}_i^{1/2}$. The mean vectors $\boldsymbol{\mu}_i$ were set to $\mathbf{0}$ in both groups. The i.i.d. random errors $\boldsymbol{\epsilon}_{ik} = (\epsilon_{ik1}, \dots, \epsilon_{ikd})^\top$ with mean $\mathbf{E}(\boldsymbol{\epsilon}_{ik}) = \mathbf{0}_d$ and $\text{Cov}(\boldsymbol{\epsilon}_{ik}) = \mathbf{I}_{d \times d}$ were generated by simulating independent standardized components

$$\epsilon_{iks} = \frac{Y_{iks} - \mathbf{E}(Y_{iks})}{\sqrt{\text{Var}(Y_{iks})}}$$

for various distributions of Y_{iks} . In particular, we simulated normal, χ_3^2 , lognormal, t_3 and double-exponential distributed random variables. We investigated balanced as well as unbalanced designs with sample size vectors $\mathbf{n}^{(1)} = (10, 10)^\top$, $\mathbf{n}^{(2)} = (20, 20)^\top$, $\mathbf{n}^{(3)} = (10, 20)^\top$ and $\mathbf{n}^{(4)} = (20, 10)^\top$, respectively. A major criterion concerning the accuracy of the procedures is their behavior in situations where increasing variances (settings 3 and 4 above) are combined with increasing sample sizes ($\mathbf{n}^{(3)}$, positive pairing) or decreasing sample sizes ($\mathbf{n}^{(4)}$, negative pairing). In this setting, we tested the null hypothesis $H_0^\mu : \{(\mathbf{P}_a \otimes \mathbf{I}_d) \boldsymbol{\mu} = \mathbf{0}\} = \{\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2\}$, i.e., no treatment effect. The resulting type-I error rates (nominal level $\alpha = 5\%$) for $d = 4$ and $d = 8$ endpoints are displayed in Table 1 (normal distribution) and Table 2 (χ_3^2 -distribution), respectively. Further simulation results for lognormal, t_3 and double-exponential distributed errors for the parametric bootstrap of WTS and MATS can be found in Tables 7 – 9 in the supplementary material.

As already noticed by [23], the WTS with the χ^2 approximation is far too liberal, reaching type-I error rates of more than 50% in some scenarios (e.g., for $d = 8$ with negative pairing, i.e., covariance setting S3 and $\mathbf{n} = (20, 10)^\top$). Even in the scenarios with only $d = 4$ dimensions and $\mathbf{n} = (20, 20)^\top$, the error rates are around 9% instead of 5%. The parametric bootstrap of the WTS greatly improves this behavior for all situations. However, it still shows a rather liberal behavior with type-I error rates of around 10% in some situations, e.g., $d = 8$ dimensions with S3 or S4 and $\mathbf{n} = (20, 10)^\top$ in Tables 1 and 2.

The wild bootstrap of the MATS shows a rather liberal behavior across all scenarios and can therefore not be recommended in practice. In contrast, both the parametric and the nonparametric bootstrap of the MATS show a very accurate type-I error rate control. The nonparametric bootstrap is often slightly more conservative than the parametric bootstrap and thus works better in situations with negative pairing, especially for the χ_3^2 -distribution, i.e., for S3 and S4 with $\mathbf{n} = (20, 10)^\top$ and $d = 4$ or $d = 8$ dimensions in Table 2. In most other scenarios, however, the parametric bootstrap yields slightly better results. The improvement of the parametric bootstrap MATS over WTS (PBS) and nonparametric bootstrap MATS is most pronounced for large d , i.e., in situations where d is close to $\min(n_1, n_2)$.

However, in situations with negative pairing and skewed distributions (see Table 2 as well as Table 7 in the supplementary material), the parametric bootstrap MATS shows a slightly liberal behavior. For t_3 and double-exponentially distributed errors and negative pairing, in contrast, the parametric bootstrap MATS is slightly conservative, see Tables 8 and 9 in the supplementary material, respectively.

Surprisingly, the resampling approaches based on the MATS improve with growing d in most settings, i.e., when the number of endpoints is closer to the sample size. The WTS approach, in contrast, gets worse in these scenarios. This might be an interesting approach for future research in high-dimensional settings such as in [31].

As a result, we find that the MATS with the parametric bootstrap approximation is the best procedure in most scenarios. Especially, it is less conservative than the nonparametric bootstrap approximation and less liberal than the WTS equipped with the parametric bootstrap approach over all simulation settings. Only in situations with negative pairing and skewed distributions, the new procedure shows a slightly liberal behavior.

Table 1: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) in the one-way layout for the normal distribution.

d	Cov	n	WTS (χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS (NPBS)
$d = 4$	S1	(10, 10)	15.2	4.5	6.9	5.2	4.4
		(10, 20)	14.5	5.9	6.9	5.1	4.5
		(20, 10)	14	5.6	7.2	5.3	4.8
		(20, 20)	9.5	5.3	6	5.1	5.1
	S2	(10, 10)	15.2	4.5	7	5	4.5
		(10, 20)	14.5	5.8	6.9	5	4.5
		(20, 10)	14	5.6	7.3	5.5	4.9
		(20, 20)	9.5	5.4	6.3	5.2	5
	S3	(10, 10)	18.3	5.5	7.3	4.8	3.6
		(10, 20)	10.9	4.7	6.4	4.8	4.4
		(20, 10)	21.4	6.6	7.8	4.8	3.4
		(20, 20)	11.2	5.7	6.3	5.1	4.6
	S4	(10, 10)	18.3	5.6	7.5	4.8	3.9
		(10, 20)	11	5.2	6.1	4.6	4.3
		(20, 10)	21	6.7	7.9	4.7	3.2
		(20, 20)	10.9	5.7	6.2	5.0	4.7
$d = 8$	S1	(10, 10)	38.6	4.7	7.7	5.1	4.3
		(10, 20)	31	6.2	6.9	5	4.2
		(20, 10)	32.1	6.1	6.6	4.6	4
		(20, 20)	17.0	4.9	5.8	4.8	4.8
	S2	(10, 10)	38.6	4.5	7.9	4.3	3.4
		(10, 20)	31	6.3	7.4	4.3	3.6
		(20, 10)	32.1	6.1	7	4.1	3.4
		(20, 20)	17.0	4.7	6.2	4.8	4.5
	S3	(10, 10)	50.1	6.6	7.9	4.2	2.8
		(10, 20)	21.8	4.1	6.3	4.4	4.1
		(20, 10)	55	10.3	8.5	3.6	2.2
		(20, 20)	21.9	5.4	6.1	4.0	3.6
	S4	(10, 10)	48.9	6.3	7.8	3.6	2.4
		(10, 20)	21.9	4.2	6.3	3.8	3.4
		(20, 10)	54.1	10.4	8.4	3.5	2
		(20, 20)	21.8	5.2	6.0	3.9	3.6

Table 2: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) in the one-way layout for the χ^2_3 -distribution.

d	Cov	n	WTS (χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS (NPBS)
$d = 4$	S1	(10, 10)	15.3	4	7.1	4.8	3.4
		(10, 20)	13.9	5.5	7.3	5.6	4.6
		(20, 10)	14.6	5.7	7.7	5.9	4.6
		(20, 20)	8.9	4.7	6.3	5.5	5
	S2	(10, 10)	15.3	4.1	7.1	4.5	3.2
		(10, 20)	13.9	5.6	7.5	5.5	4.5
		(20, 10)	14.6	5.8	7.7	5.5	4.5
		(20, 20)	8.9	4.7	6.3	5.3	4.7
	S3	(10, 10)	20.6	7.1	9.5	6.1	3.8
		(10, 20)	11.2	4.8	6.9	4.8	3.7
		(20, 10)	26.2	10.9	12.3	8.9	5.6
		(20, 20)	12.8	6.6	7.6	6	4.7
	S4	(10, 10)	21.2	7.2	9.6	6.2	3.8
		(10, 20)	11.1	5	6.9	4.7	3.4
		(20, 10)	26.5	10.7	12.7	8.9	5.6
		(20, 20)	12.9	6.7	7.7	6.2	4.7
$d = 8$	S1	(10, 10)	39.3	3.8	7.7	4.9	3.4
		(10, 20)	32.3	5.5	7.6	5.9	4.7
		(20, 10)	33.4	6.3	7.2	5.1	4.2
		(20, 20)	16.9	4.5	5.9	4.9	4.6
	S2	(10, 10)	39.3	3.8	8.1	4.3	2.7
		(10, 20)	32.3	5.5	8.6	5.2	4
		(20, 10)	33.4	6.3	7.6	4.9	3.9
		(20, 20)	16.9	4.5	6.2	4.5	4.0
	S3	(10, 10)	53.1	6.8	10.2	5.5	3.1
		(10, 20)	23.4	4.8	6.8	4.6	3.5
		(20, 10)	59.9	13.9	13.7	8.1	4.6
		(20, 20)	24.8	6.9	7.9	5.7	4.1
	S4	(10, 10)	52.5	6.3	11	5.3	2.6
		(10, 20)	24.3	4.5	7.1	4.1	2.8
		(20, 10)	59	13.6	14.8	8.4	4.5
		(20, 20)	24.3	6.9	7.7	5.7	4.0

5.1.1. Singular Covariance Matrix

In order to analyze the behavior of the discussed methods in designs involving singular covariance matrices, we considered the one-way layout described above with $a = 2$ groups and $d \in \{4, 8\}$ observations involving the following covariance settings (displayed for $d = 4$):

$$\text{Setting 5: } \mathbf{V}_1 = \begin{pmatrix} 1 & 1/2 & 1 & 1 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1 & 1/2 & 1 & 1 \\ 1 & 1/2 & 1 & 1 \end{pmatrix}, \mathbf{V}_2 = \mathbf{V}_1 + 0.5 \cdot \mathbf{J}_d$$

$$\text{Setting 6: } \mathbf{V}_1 = \begin{pmatrix} 1 & 0.6 & 0.36 & 0.18 \\ 0.6 & 1 & 0.6 & 0.3 \\ 0.36 & 0.6 & 1 & 0.5 \\ 0.18 & 0.3 & 0.5 & 0.25 \end{pmatrix}, \mathbf{V}_2 = \mathbf{V}_1 + 0.5 \cdot \mathbf{J}_d$$

$$\text{Setting 7: } \mathbf{V}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0.5 \end{pmatrix}, \mathbf{V}_2 = \mathbf{V}_1 + 0.5 \cdot \mathbf{J}_d$$

Setting 6 is based on an $AR(0.6)$ covariance matrix (see setting 2 above), where the last row and column have been replaced by half the row/column before, respectively. Setting 7 is based on $\tilde{\mathbf{V}}_1 = \text{diag}(\sqrt{2^s})$, $s = 0, \dots, d-1$, where the last row and column have been replaced by half the row/column before. We have considered the same sample size vectors as above.

The results are displayed in Tables 3 and 4. The parametric bootstrap of the MATS again yields the best results in almost all scenarios. The wild bootstrap, in contrast, is again rather liberal. For the χ^2 approximation of the WTS, the results are in concordance with the theoretical reflections mentioned in Section 2: Covariance setting S5 corresponds to the case, where the rank of \mathbf{T} and $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}$ differs and as calculated above, the χ^2 -approximation becomes very conservative here. In setting S6 and S7, in contrast, there is no rank jump despite the singular covariance matrices and the χ^2 -approximation shows its usual liberal behavior. Since the rank of $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}$ is not known in practice, the WTS should not be used for data with possibly singular covariance matrices. It turns out, however, that the parametric bootstrap of the WTS is relatively robust against singular covariance matrices. Its behavior is comparable to the scenarios above with non-singular covariance matrices. It is, however, rather liberal for $\mathbf{n} = (20, 10)^\top$, especially with the χ_3^2 -distribution, see Table 4. This behavior is improved by the parametric bootstrap MATS, e.g., for $d = 8$ and S7, the WTS (PBS) leads to a type-I error of 9%, whereas the MATS (PBS) is at 5.1%. The nonparametric bootstrap, in contrast, sometimes leads to strictly conservative test decisions. This is especially apparent for $d = 8$ and covariance setting S7 in Tables 3 and 4.

Table 3: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) in the one-way layout with singular covariance matrices for the normal distribution.

d	Cov	n	WTS (χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS (NPBS)
$d = 4$	S5	(10, 10)	3.1	5.1	5.9	4.8	4.4
		(10, 20)	2.6	5.1	5.7	4.9	4.6
		(20, 10)	2.7	5.3	6.6	5.3	4.7
		(20, 20)	1.7	4.8	5.6	5.2	5.0
	S6	(10, 10)	16.7	5.3	7.1	5	4.6
		(10, 20)	12.4	5.1	6	4.7	4.3
		(20, 10)	17.1	6.1	7.1	5.3	4.6
		(20, 20)	10.2	5.5	6.0	5.2	5.1
	S7	(10, 10)	16.6	5.2	7.3	4.7	4
		(10, 20)	12.3	5.8	6.6	4.6	4.2
		(20, 10)	16.3	5.7	6.9	4.5	4
		(20, 20)	9.4	4.8	5.9	4.8	4.8
$d = 8$	S5	(10, 10)	2.8	4.5	6.2	5	4.7
		(10, 20)	2.3	4.9	5.5	4.9	4.7
		(20, 10)	2.6	4.4	5.6	4.7	4.3
		(20, 20)	1.5	4.6	5.5	4.9	4.8
	S6	(10, 10)	39.5	4.4	8.2	5	4.2
		(10, 20)	28.8	5.4	6.6	4.6	4.2
		(20, 10)	35.7	7.0	7.5	4.8	4.0
		(20, 20)	17.3	4.7	6.1	4.8	4.5
	S7	(10, 10)	38.8	4.2	7.4	4	2.9
		(10, 20)	27.4	5.2	6.6	3.6	3
		(20, 10)	36.3	6.3	7.4	3.8	3.2
		(20, 20)	17.3	5.1	5.9	4.2	4.0

Table 4: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) in the one-way layout with singular covariance matrices for the χ^2_3 -distribution.

d	Cov	n	WTS (χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS (NPBS)
$d = 4$	S5	(10, 10)	2.7	4.2	6.7	5.4	4.5
		(10, 20)	1.9	4.8	5.9	4.9	4.5
		(20, 10)	3.5	6.5	7.3	5.9	5.2
		(20, 20)	1.7	4.9	6.2	5.7	5.5
	S6	(10, 10)	19.7	7.1	7.4	5.2	4.1
		(10, 20)	14.6	7.1	6.4	5	4.3
		(20, 10)	20.1	8.5	8.1	6.4	5.4
		(20, 20)	11.4	6.3	6.5	5.7	5.3
	S7	(10, 10)	19.4	7	7.1	4.1	3.1
		(10, 20)	14.5	6.7	6.4	4.2	3.4
		(20, 10)	20.3	8.7	8.3	6.1	4.5
		(20, 20)	11.7	6.4	6.1	5.1	4.5
$d = 8$	S5	(10, 10)	2.4	4.7	6.1	5.1	4.6
		(10, 20)	2.6	5.3	6.1	5.3	5
		(20, 10)	3	5.6	6	5.1	4.6
		(20, 20)	1.2	4.5	5.9	5.3	5.1
	S6	(10, 10)	43.1	5.4	8.2	5.1	3.9
		(10, 20)	30.7	6.6	7.3	5.2	4.2
		(20, 10)	39.2	8.7	8.3	5.6	4.5
		(20, 20)	19.3	5.6	6.8	5.1	4.7
	S7	(10, 10)	42.4	5.5	7.5	3.3	1.7
		(10, 20)	31.1	6.3	7.1	4	2.4
		(20, 10)	39.5	9	9.2	5.1	3.4
		(20, 20)	18.7	5.3	5.1	3.2	2.6

5.2. Two-way layout

We have investigated the behavior of the methods in a setting with two crossed factors A and B , which is again adapted from [23]. In particular, we simulated a 2×2 designs with covariance matrices similar to the one-way layout above. A detailed description of the simulation settings as well as the results for the main and interaction effects are deferred to the supplementary material. Here we only summarize our findings: Since the total sample size N is larger in this scenario, the asymptotic results come into play and therefore all methods lead to more accurate results than in the one-way layout. Nevertheless we find a similar behavior as in the one-way layout: Again, the MATS and the WTS with the parametric bootstrap approach control the type-I error very accurately, whereas the nonparametric bootstrap approach leads to slightly more conservative results. Both the WTS with χ^2 approximation and the wild bootstrap MATS can not be recommended due to their liberal behavior. In situations with negative pairing (covariance setting 10 and 11 with sample size vector $\mathbf{n}^{(3)}$), the parametric bootstrap MATS improves the slightly liberal behavior of the WTS, see e.g., Table 10 for the normal distribution, where the WTS (PBS) leads to a type-I error of 6.1%, while the MATS (PBS) is at 4.9%.

5.3. Power

We have investigated the empirical power of the proposed methods to detect a fixed alternative in the simulation scenarios above. Data was simulated as described in Section 5.1 but now with $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (\delta, \dots, \delta)^\top$ for varying shifts $\delta \in \{0, 0.5, 1, 1.5, 2, 3\}$. Due to the liberality of the classical Wald-type test and the wild bootstrapped MATS, we have only considered the WTS with parametric bootstrap as well as the parametric and nonparametric bootstrap of the MATS. The results for selected scenarios are displayed in Figures 1 – 2. The plots show that both resampling versions of the MATS have a higher power for detecting the fixed alternative than the WTS. The parametric bootstrap of the MATS has a slightly higher power than the nonparametric bootstrap, a behavior that is more pronounced for the χ^2 -distribution (Figure 1). Moreover, the power analysis shows a clear advantage of applying the parametric bootstrap approach to the MATS over its application to the WTS. For example, in the scenario with normally distributed data, $d = 8$ dimensions, covariance setting S4 and $\mathbf{n} = (10, 20)^\top$ observations (Figure 2), the parametric bootstrap MATS has twice as much power as its WTS version in case of $\delta = 0.5$ (34.4% as compared to 16.7%). Similar differences can also be observed in some of the other settings.

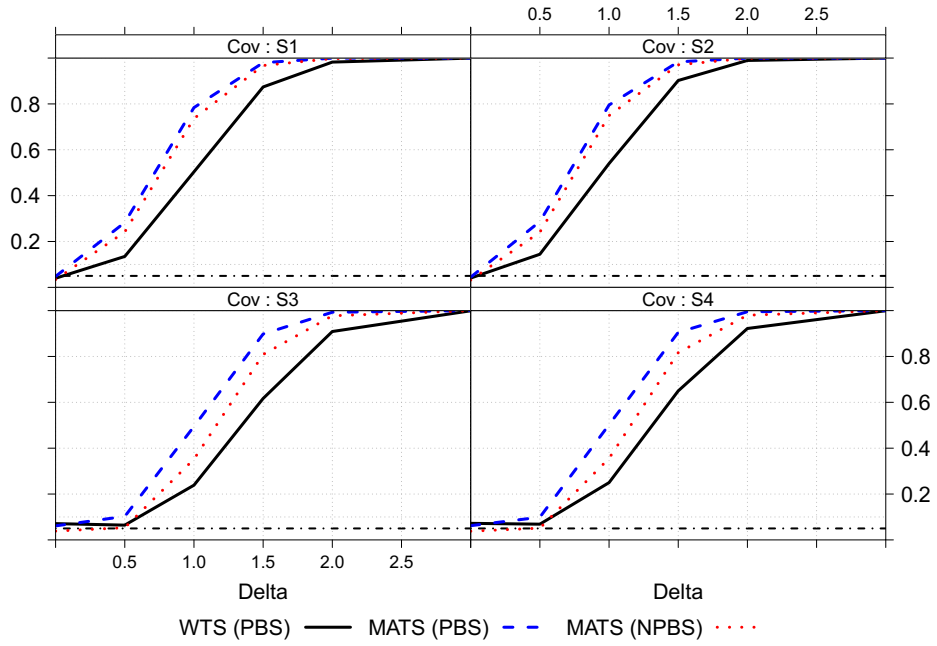


Figure 1: Empirical power results for the WTS with parametric bootstrap as well as the MATS with parametric (PBS) and nonparametric (NPBS) bootstrap for χ^2_3 -distributed data with $d = 4$ dimensions and sample sizes $\mathbf{n} = (10, 10)^T$.

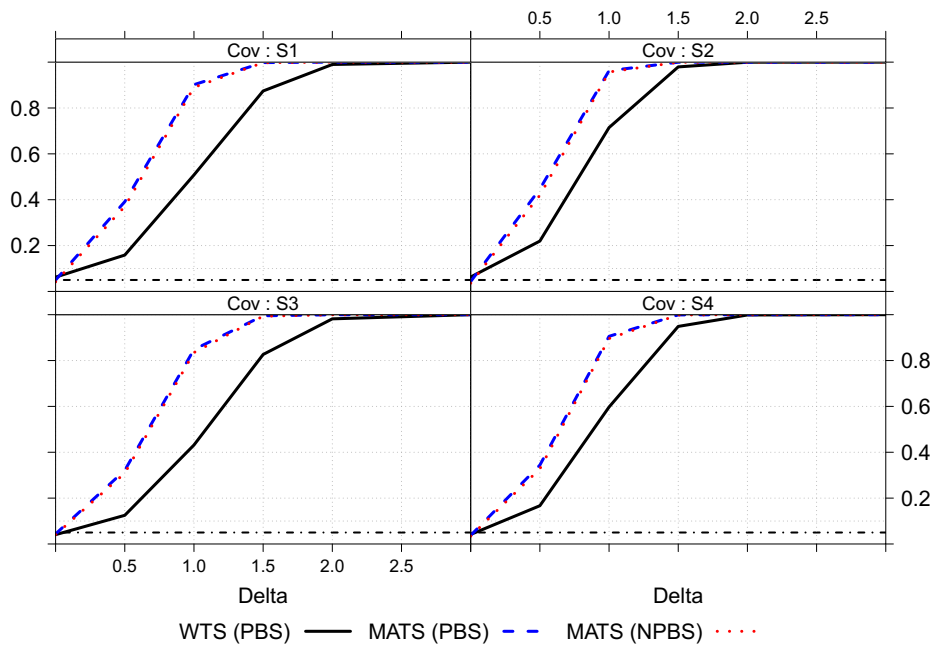


Figure 2: Empirical power results for the WTS with parametric bootstrap as well as the MATS with parametric (PBS) and nonparametric (NPBS) bootstrap for normally distributed data with $d = 8$ dimensions and sample sizes $\mathbf{n} = (10, 20)^T$.

6. Application: Analysis of the Data Example

As a data example, we consider 7 demographic factors of US citizens in 43 states. Our aim is to investigate whether these factors differ between the states. The full data set ‘county_facts.csv’ is available from kaggle (<https://www.kaggle.com/joelwilson/2012-2016-presidential-elections>). In order to have sufficient sample sizes for the analysis and to avoid a high-dimensional setting, we exclude all states with less than 15 counties. In particular, we removed Connecticut, Delaware, Hawaii, Massachusetts, New Hampshire, Rhode Island and Vermont. We consider the following demographic factors: the population estimate for 2014 (PST045214), the percentage of female citizens in 2014 (SEX255214) as well as the percentage of white (RHI125214), black or African American (RHI225214), American Indian and Alaska native (RHI325214), Asian (RHI425214) and native Hawaiian and other pacific islanders (RHI525214) citizens in 2014. This results in a one-way layout with $a = 43$ levels of the factor ‘state’ and $d = 7$ dimensions. The sample sizes and mean values for the different states can be found in Table 5. Figure 3 exemplarily displays boxplots for the percentage of white citizens across the different states.

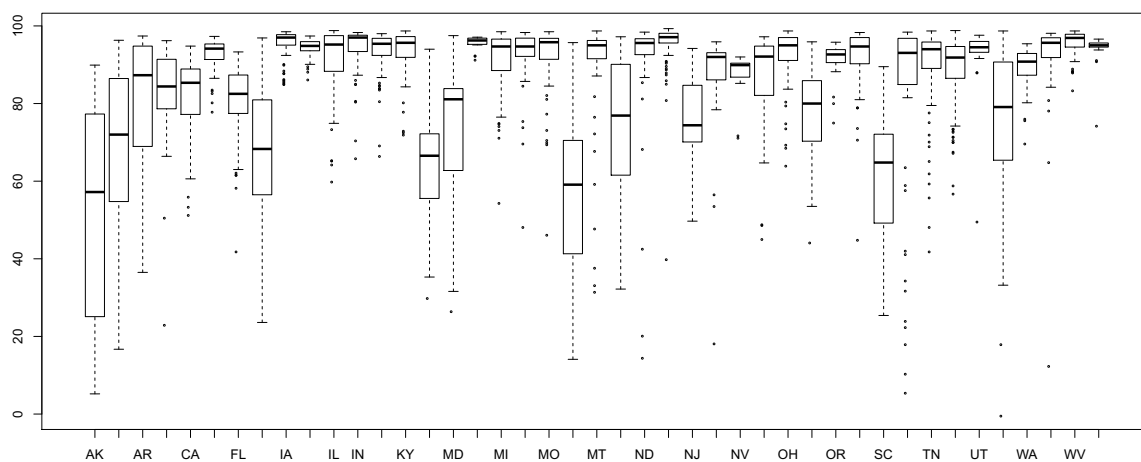


Figure 3: Boxplots of the percentage of white citizens across the different states.

We now want to analyze, whether there is a significant difference in the multivariate means for the different states. The null hypothesis of interest thus is $H_0 : \{(P_{43} \otimes I_7)\mu = \mathbf{0}\}$. Since the empirical covariance matrix is computationally singular in this example (reciprocal condition number $1.7e-16$), we cannot apply the Wald-type test. Thus, we consider the parametric bootstrap approach of the MATS which yielded the best results in the simulation study. Computation of the MATS results in a value of $Q_N = 393.927$ and the parametric bootstrap routine with 1,000 bootstrap runs gives a p -value of $p < 0.0001$, implying that there is indeed a significant difference between the states with respect to the 7 demographic measurements.

A confidence region for this effect can be constructed as described in Section 4. The analysis of this example, including the calculation of the confidence region, can be conducted using the R package `MANOVA.RM`.

7. Conclusions and Discussion

We have investigated a test statistic for multivariate data (MATS) which is based on a modified Dempster statistic. Contrary to classical MANOVA models, we incorporate general heteroscedastic designs and allow for singular covariance matrices while postulating their existence as solely distributional assumption. Moreover, our proposed MATS statistic is invariant under linear transformations of the response variables.

In order to improve the small sample behavior of the test statistic, we have investigated different bootstrap approaches, namely a parametric bootstrap, a wild bootstrap and a nonparametric bootstrap procedure. We have rigorously proven that they lead to asymptotically exact and consistent tests and even analyzed their local power behavior.

In a large simulation study, the parametric bootstrap turned out to perform best in most scenarios, even with skewed data and heteroscedastic variances. Although the type-I error control is still not ideal in the latter case, the method performed advantageous over the parametric bootstrap of the WTS proposed in [23] and has the additional advantage of being applicable to situations with singular covariance matrices. In situations with skewed distributions, the parametric bootstrap of the MATS yielded more robust results than the WTS. The wild bootstrap approach, in contrast, turned out to be very liberal in all scenarios, while the nonparametric bootstrap was mostly slightly more conservative than the parametric bootstrap. Power simulations showed a clear advantage of the parametric bootstrap MATS compared to the WTS (PBS) as well as the nonparametric bootstrap. All in all, we therefore recommend the parametric bootstrap based on the MATS for practical applications in a multivariate setting.

Furthermore, we have constructed confidence regions and simultaneous confidence intervals for contrasts $\mathbf{h}^\top \boldsymbol{\mu}$ based on the bootstrap quantiles. These confidence regions provide an additional benefit for the analysis of multivariate data since they allow for more detailed insight into the nature of the estimates.

In order to facilitate application of the proposed methods, the parametric bootstrap test and the calculation of confidence regions are implemented in the R package `MANOVA.RM`.

Following the idea of [39] we plan to extend our concepts to the high-dimensional setting, i.e., where the sample size N may be less than the dimension d . This approach looks promising, since we have seen in the simulation study that the MATS with the parametric bootstrap approach exhibited an improved type-I error control with increasing d . However, the extension to high-dimensional data requires different techniques and will be part of future research.

Acknowledgment

The authors would like to thank Dr. Jan Paul and Prof. Dr. Volker Rasche for providing the cardiology data example used in the supplement. This work was supported by the German Research Foundation projects DFG PA 2409/3-1 and PA 2409/4-1.

References

References

- [1] M. S. Bartlett. A note on tests of significance in multivariate analysis. *Mathematical Proceedings of the Cambridge Philosophical Society*, 35(02):180–185, 1939. Cambridge University Press.
- [2] A. C. Bathke, S. W. Harrar, and L. V. Madden. How to compare small multivariate samples using nonparametric tests. *Computational Statistics & Data Analysis*, 52(11):4951–4965, 2008.
- [3] J. Beyersmann, S. D. Termini, and M. Pauly. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 40(3):387–402, 2013.
- [4] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.
- [5] E. Brunner. Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity. *Mathematical Statistics with Applications in Biometry*, 2001.
- [6] E. Brunner, F. Konietschke, M. Pauly, and M. L. Puri. Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.
- [7] E. Brunner, U. Munzel, and M. L. Puri. Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis*, 70(2):286–317, 1999.
- [8] E. Brunner and M. L. Puri. Nonparametric methods in factorial designs. *Statistical papers*, 42(1):1–52, 2001.
- [9] A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.
- [10] A. C. Cameron and D. L. Miller. A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.
- [11] E. Chung and J. P. Romano. Multivariate and multiple permutation tests. *Journal of Econometrics*, 193(1):76–91, 2016.
- [12] S. Csörgö. On the law of large numbers for the bootstrap mean. *Statistics & probability letters*, 14(1):1–7, 1992.

- [13] R. Davidson and E. Flachaire. The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169, 2008.
- [14] A. P. Dempster. A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 29(4):995–1010, 1958.
- [15] A. P. Dempster. A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16(1):41–50, 1960.
- [16] S. Friedrich, E. Brunner, and M. Pauly. Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153:255–265, 2017.
- [17] S. Friedrich, F. Konietzschke, and M. Pauly. A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, 2016.
- [18] S. W. Harrar and A. C. Bathke. A modified two-factor multivariate analysis of variance: asymptotics and small sample approximations. *Annals of the Institute of Statistical Mathematics*, 64(1):135–165, 2012.
- [19] M. Hasler and L. A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 50(5):793–800, 2008.
- [20] H. Hotelling. A generalized t -test and measure of multivariate dispersion. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.
- [21] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.
- [22] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. 6th edition, Prentice Hall, 2007.
- [23] F. Konietzschke, A. Bathke, S. Harrar, and M. Pauly. Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, 140:291–301, 2015.
- [24] K. Krishnamoorthy and F. Lu. A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation*, 80(8):873–887, 2010.
- [25] D. Lawley. A generalization of fisher’s z test. *Biometrika*, 30(1-2):180–187, 1938.
- [26] D. Lin. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16(8):901–910, 1997.
- [27] C. Liu, A. C. Bathke, and S. W. Harrar. A nonparametric version of wilks’ lambda - asymptotic results and small sample approximations. *Statistics & Probability Letters*, 81(10):1502–1506, 2011.
- [28] R. Y. Liu. Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708, 1988.
- [29] E. Mammen. *When does bootstrap work? Asymptotic results and simulations*. Springer Science & Business Media, 1993.
- [30] R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.
- [31] M. Pauly, D. Ellenberger, and E. Brunner. Analysis of high-dimensional one group repeated measures designs. *Statistics*, 49:1243–1261, 2015.
- [32] F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.
- [33] F. Pesarin and L. Salmaso. A review and some new results on permutation testing for multivariate problems. *Statistics and Computing*, 22(2):639–646, 2012.
- [34] K. Pillai. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, 26(1):117–121, 1955.
- [35] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
- [36] C. Rao and S. Mitra. *Generalized inverse of matrices and its applications*. Wiley New York, 1971.
- [37] Ł. Smaga. Bootstrap methods for multivariate hypothesis testing. *Communications in Statistics-Simulation and Computation*, 2016. Just accepted.
- [38] E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5):641–656, 2008.
- [39] M. S. Srivastava and T. Kubokawa. Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis*, 115:204–216, 2013.
- [40] G. Vallejo and M. Ato. Robust tests for multivariate factorial designs under heteroscedasticity. *Behavior research methods*, 44(2):471–489, 2012.
- [41] G. Vallejo, M. Fernández, and P. E. Livacic-Rojas. Analysis of unbalanced factorial designs with heteroscedastic data. *Journal of Statistical Computation and Simulation*, 80(1):75–88, 2010.
- [42] S. Van Aelst and G. Willems. Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association*, 106(494):706–718, 2011.
- [43] S. Van Aelst and G. Willems. Fast and robust bootstrap for multivariate inference: the R package FRB. *Journal of Statistical Software*, 53(3):1–32, 2013.
- [44] S. S. Wilks. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, 17(3):257–281, 1946.
- [45] C.-F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.
- [46] L.-W. Xu, F.-Q. Yang, S. Qin, et al. A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, 115:172–180, 2013.

Supplementary Material to MATS: Inference for potentially Singular and Heteroscedastic MANOVA

Sarah Friedrich, Markus Pauly

Institute of Statistics, Ulm University, Germany

Sarah Friedrich, Markus Pauly

Institute of Statistics, Ulm University, Germany

Abstract

In this supplementary material to the authors' paper "MATS: Inference for potentially singular and heteroscedastic MANOVA" we provide the proofs of all theorems as well as some additional simulation results for different distributions and a two-way layout. Furthermore, we provide an additional data example from cardiology, where we also explain the problem of the ATS in more detail.

Keywords: Multivariate Data; Parametric Bootstrap; Confidence Regions; Singular Covariance Matrices

Email addresses: sarah.friedrich@uni-ulm.de (Sarah Friedrich), sarah.friedrich@uni-ulm.de (Sarah Friedrich)

Preprint submitted to Elsevier

December 6, 2017

8. Proofs

Proof of Theorem 2.1

The result follows directly from the representation theorem for quadratic forms [36] and the continuous mapping theorem by noting that $\sqrt{N}(\bar{\mathbf{X}}_{\bullet} - \boldsymbol{\mu})$ has, asymptotically, as $N \rightarrow \infty$, a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = \lim_{N \rightarrow \infty} \boldsymbol{\Sigma}_N = \text{diag}(\kappa_i^{-1} \mathbf{V}_i)$. Moreover, $\widehat{\mathbf{D}}_N$ is consistent for $\mathbf{D} = \text{diag}(\kappa_i^{-1} \sigma_{is}^2)$, where the latter is of full rank by assumption. Thus, $T\widehat{\mathbf{D}}_N T$ converges in probability to $T\mathbf{D}T$ and since there is finally no rank jump in this convergence we eventually obtain

$$(T\widehat{\mathbf{D}}_N T)^+ \xrightarrow{\text{Pr}} (T\mathbf{D}T)^+, \quad (8.1)$$

where $\xrightarrow{\text{Pr}}$ denotes convergence in probability. \square

Proof of Theorem 3.1

Let

$$\mathbf{Y}_{ik} := \mathbf{X}_{ik} \cdot \mathbb{1}\{\|\mathbf{X}_{ik}\| \leq \delta\}$$

for $\delta > 0$. Then, \mathbf{Y}_{ik} has finite moments of any order, especially fourth moments exist. Analogously, given $\mathbf{Y} = (\mathbf{Y}_{ik})_{i,k}$, let $\mathbf{Y}_{ik}^* \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \widehat{\mathbf{V}}_i(\delta))$, where $\widehat{\mathbf{V}}_i(\delta) = (n_i - 1)^{-1} \sum_{k=1}^{n_i} (\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_i)^\top$ and define $\mathcal{Q}_{N,\delta}^* = N(\bar{\mathbf{Y}}_{\bullet}^*)^\top T(\widehat{\mathbf{D}}_{\delta}^* T)^+ T \bar{\mathbf{Y}}_{\bullet}^*$, where $\widehat{\mathbf{D}}_{\delta}^* = \text{diag}(N/n_i \cdot (\widehat{\sigma}_{is}^*(\delta))^2)$ and $(\widehat{\sigma}_{is}^*(\delta))^2$ is the empirical variance of \mathbf{Y}_{iks}^* .

First, we apply the multivariate Lindeberg-Feller CLT to show that, given \mathbf{X} , $\sqrt{N} \bar{\mathbf{Y}}_{\bullet}^*$ converges in distribution to a normal distributed random variable. Therefore, consider $\widetilde{\mathbf{Y}}_i^* := \sqrt{N}/n_i \mathbf{Y}_{ik}^*$ and $\mathbf{V}_i(\delta) = \text{Cov}(\mathbf{Y}_{ik})$. The Lindeberg-Feller CLT now yields convergence in distribution given the data \mathbf{X} to a normal distributed random variable

$$\sqrt{N} \bar{\mathbf{Y}}_{\bullet}^* \xrightarrow{\mathcal{D}|\mathbf{X}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\delta}), \quad (8.2)$$

if we proof the following conditions:

$$\sum_{i=1}^a \sum_{k=1}^{n_i} \mathbb{E}(\widetilde{\mathbf{Y}}_i^* | \mathbf{X}) = \mathbf{0} \quad (8.3)$$

$$\bigoplus_{i=1}^a \sum_{k=1}^{n_i} \text{Cov}(\widetilde{\mathbf{Y}}_i^* | \mathbf{X}) \xrightarrow{\text{Pr}} \boldsymbol{\Sigma}_{\delta} := \bigoplus_{i=1}^a \frac{1}{\kappa_i} \mathbf{V}_i(\delta) \quad (8.4)$$

$$\sum_{i=1}^a \sum_{k=1}^{n_i} \mathbb{E}(\|\widetilde{\mathbf{Y}}_i^*\|^2 \cdot \mathbb{1}\{\|\widetilde{\mathbf{Y}}_i^*\| > \epsilon\} | \mathbf{X}) \xrightarrow{\text{Pr}} 0 \quad \forall \epsilon > 0. \quad (8.5)$$

Condition (8.3) follows since, given the data, $\mathbf{Y}_{ik}^* \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \widehat{\mathbf{V}}_i(\delta))$. For condition (8.4), note that

$$\bigoplus_{i=1}^a \sum_{k=1}^{n_i} \text{Cov}(\widetilde{\mathbf{Y}}_i^* | \mathbf{X}) = \bigoplus_{i=1}^a \sum_{k=1}^{n_i} \text{Cov}\left(\frac{\sqrt{N}}{n_i} \mathbf{Y}_{ik}^* | \mathbf{X}\right) = \bigoplus_{i=1}^a \sum_{k=1}^{n_i} \frac{N}{n_i^2} \widehat{\mathbf{V}}_i(\delta) = \bigoplus_{i=1}^a \frac{N}{n_i} \widehat{\mathbf{V}}_i(\delta).$$

Thus, since $\widehat{\mathbf{V}}_i(\delta)$ is a consistent estimator of $\mathbf{V}_i(\delta)$, (8.4) follows. For (8.5) note that

$$\mathbb{1}\{\|\widetilde{\mathbf{Y}}_i^*\| > \epsilon\} = 1 \Leftrightarrow \delta \geq \|\mathbf{Y}_{ik}^*\| > \frac{n_i}{\sqrt{N}} \epsilon = \frac{n_i}{N} \sqrt{N} \epsilon.$$

Since $n_i/N \rightarrow \kappa_i > 0$, the right hand side converges to infinity as $N \rightarrow \infty$. Therefore, for arbitrary fixed $\delta > 0$ and $\epsilon > 0$ we finally have $\mathbb{1}\{\|\widetilde{\mathbf{Y}}_i^*\| > \epsilon\} = 0$ for N large enough and Equation (8.5) follows. Altogether this proves (8.2).

Since $\widehat{\mathbf{D}}_\delta^* \xrightarrow{\text{Pr}} \mathbf{D}_\delta = \text{diag}(\kappa_i^{-1} \text{Var}(Y_{iks})), i = 1, \dots, a, s = 1, \dots, d$ due to existence of finite fourth moments of the truncated random variables \mathbf{Y} , it now follows from continuous mapping and the representation theorem for quadratic forms that

$$\mathcal{Q}_{N,\delta}^* = N(\overline{\mathbf{Y}}_\bullet^*)^\top \mathbf{T}(\mathbf{T}\widehat{\mathbf{D}}_\delta^* \mathbf{T})^+ \mathbf{T}\overline{\mathbf{Y}}_\bullet^* \xrightarrow{\mathcal{D}|\mathbf{X}} \tilde{Z}, N \rightarrow \infty,$$

in probability, see, e.g., [23] and the references cited therein for similar arguments. Here, $\tilde{Z} = \sum_{i=1}^a \sum_{s=1}^d \tilde{\lambda}_{is} \tilde{Z}_{is}$ with $\tilde{Z}_{is} \sim \chi_1^2$ and $\tilde{\lambda}_{is}$ are the eigenvalues of $\mathbf{T}(\mathbf{T}\mathbf{D}_\delta \mathbf{T})^+ \mathbf{T}\boldsymbol{\Sigma}_\delta$. Furthermore, since

$$\text{Cov}(\mathbf{Y}_{ik}) = \text{Cov}(\mathbf{X}_{ik} \cdot \mathbb{1}\{\|\mathbf{X}_{ik}\| \leq \delta\}) \rightarrow \text{Cov}(\mathbf{X}_{ik}), \delta \rightarrow \infty,$$

by dominated convergence and analogously $\mathbf{D}_\delta \rightarrow \mathbf{D}, \delta \rightarrow \infty$, it follows that

$$\tilde{Z} \xrightarrow{\mathcal{D}|\mathbf{X}} Z, \delta \rightarrow \infty$$

in probability, where Z is the limit variable of \mathcal{Q}_N given in Theorem 2.1. Thus, it remains to show that

$$\lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} \Pr(|\mathcal{Q}_{N,\delta}^* - \mathcal{Q}_N^*| > \epsilon | \mathbf{X}) \stackrel{\text{a.s.}}{=} 0 \text{ for all } \epsilon > 0.$$

Let $\tilde{\mathbf{Y}} := \sqrt{N} \mathbf{T} \overline{\mathbf{Y}}_\bullet^*, \tilde{\mathbf{X}} := \sqrt{N} \mathbf{T} \overline{\mathbf{X}}_\bullet^*, \mathbf{M}_\delta := (\mathbf{T} \mathbf{D}_\delta^* \mathbf{T})^+$ and $\mathbf{M} := (\mathbf{T} \mathbf{D}^* \mathbf{T})^+$. Then, $\mathcal{Q}_{N,\delta}^* = \tilde{\mathbf{Y}}^\top \mathbf{M}_\delta \tilde{\mathbf{Y}}$ and $\mathcal{Q}_N^* = \tilde{\mathbf{X}}^\top \mathbf{M} \tilde{\mathbf{X}}$ and therefore

$$\mathcal{Q}_{N,\delta}^* - \mathcal{Q}_N^* = \underbrace{\tilde{\mathbf{Y}}^\top \mathbf{M}_\delta (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}})}_{(A)} + \underbrace{(\tilde{\mathbf{Y}}^\top \mathbf{M}_\delta - \tilde{\mathbf{X}}^\top \mathbf{M}) \tilde{\mathbf{X}}}_{(B)}.$$

First, consider part (A) and let $\boldsymbol{\xi}_{ik} := \mathbf{X}_{ik} - \mathbf{Y}_{ik} = \mathbf{X}_{ik} \mathbb{1}\{\|\mathbf{X}_{ik}\| > \delta\}$. Another application of the multivariate Lindeberg-Feller CLT shows that

$$\sqrt{N}(\overline{\mathbf{Y}}_\bullet^* - \overline{\mathbf{X}}_\bullet^*) \xrightarrow{\mathcal{D}|\mathbf{X}} \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_\delta)$$

in probability, where $\tilde{\boldsymbol{\Sigma}}_\delta := \bigoplus_{i=1}^a \kappa_i^{-1} \text{Cov}(\mathbf{X}_{ik} \mathbb{1}\{\|\mathbf{X}_{ik}\| > \delta\}) = \bigoplus_{i=1}^a \kappa_i^{-1} \text{Cov}(\boldsymbol{\xi}_{ik})$.

Thus,

$$\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \xrightarrow{\mathcal{D}|\mathbf{X}} \mathcal{N}(\mathbf{0}, \mathbf{T} \tilde{\boldsymbol{\Sigma}}_\delta)$$

in probability and the representation theorem again yields $\tilde{\mathbf{Y}}^\top \mathbf{M}_\delta (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}|\mathbf{X}} B_\delta = \sum_{i=1}^a \sum_{s=1}^d \eta_{is}^{(\delta)} B_{is}^2$ in probability, where $B_{is}^2 \sim \chi_1^2$ and $\eta_{is}^{(\delta)}$ are the eigenvalues of $(\mathbf{T} \tilde{\boldsymbol{\Sigma}}_\delta)^{1/2} \mathbf{M}_\delta (\mathbf{T} \tilde{\boldsymbol{\Sigma}}_\delta)^{1/2}$.

By dominated convergence it follows that $\tilde{\boldsymbol{\Sigma}}_\delta \rightarrow \mathbf{0}$ for $\delta \rightarrow \infty$. Since $\boldsymbol{\Sigma}_\delta \rightarrow \boldsymbol{\Sigma}$ and $\mathbf{D}_\delta \rightarrow \mathbf{D}$ we finally obtain $B_\delta \rightarrow 0$ as $\delta \rightarrow \infty$. Altogether, this proves

$$\lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} (A) = 0.$$

It remains to consider part (B) which we expand as

$$(\tilde{\mathbf{Y}}^\top \mathbf{M}_\delta - \tilde{\mathbf{X}}^\top \mathbf{M}) \tilde{\mathbf{X}} = [\tilde{\mathbf{Y}}^\top \mathbf{M}_\delta - \tilde{\mathbf{X}}^\top \mathbf{M}_\delta + \tilde{\mathbf{X}}^\top (\mathbf{M}_\delta - \mathbf{M})] \tilde{\mathbf{X}}.$$

Using similar arguments as above, it follows that given the data

$$(\tilde{\mathbf{Y}}^\top - \tilde{\mathbf{X}}^\top) \mathbf{M}_\delta \tilde{\mathbf{X}}$$

converges to 0 in probability for $N \rightarrow \infty$ and, subsequently, $\delta \rightarrow \infty$.

Finally, $(\hat{\sigma}_{is}^*(\delta))^2 - (\hat{\sigma}_{is}^*)^2$ converges to zero (where $(\hat{\sigma}_{is}^*)^2$ is the empirical variance of X_{iks}^*) by dominated convergence and consistency of the variance estimators and it follows that $\mathbf{M}_\delta - \mathbf{M}$ converges to 0. This concludes the proof. \square

Proof of Theorem 3.2

Analogous to the proof of Theorem 3.1, we define

$$\mathbf{Y}_{ik} := \mathbf{X}_{ik} \cdot \mathbb{1}\{\|\mathbf{X}_{ik}\| \leq \delta\}$$

for $\delta > 0$ as well as, given $\mathbf{Y}, \mathbf{Y}_{ik}^* = W_{ik}(\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_i)$ and $\mathbf{Q}_{N,\delta}^* = N(\bar{\mathbf{Y}}_i^*)^\top \mathbf{T}(\mathbf{T}\widehat{\mathbf{D}}_\delta^* \mathbf{T})^+ \mathbf{T}\bar{\mathbf{Y}}_i^*$.

The first part of the proof follows analogous to the proof of Theorem 3.1 above. It remains to show that $(\hat{\sigma}_{is}^*(\delta))^2 - (\hat{\sigma}_{is}^*)^2$ converges to zero. Therefore, consider

$$(*) := (\hat{\sigma}_{is}^*(\delta))^2 - (\hat{\sigma}_{is}^*)^2 = \frac{1}{n_i} \sum_{k=1}^{n_i} W_{ik}^2 \xi_{iks}^2 - (\bar{\mathbf{Y}}_{i,s}^*)^2 + (\bar{\mathbf{X}}_{i,s}^*)^2,$$

where again $\xi_{iks} := X_{iks} - Y_{iks}$. For the first summand on the right hand side it holds

$$\mathbb{E}\left(\frac{1}{n_i} \sum_{k=1}^{n_i} W_{ik}^2 \xi_{iks}^2 \mid \mathbf{X}\right) = \frac{1}{n_i} \sum_{k=1}^{n_i} \xi_{iks}^2 \xrightarrow[N \rightarrow \infty]{\text{a.s.}} \mathbb{E}(X_{i1s}^2 \mathbb{1}\{\|\mathbf{X}_{ik}\| > \delta\}).$$

Now letting $\delta \rightarrow \infty$, it follows from dominated convergence that

$$\mathbb{E}(X_{i1s}^2 \mathbb{1}\{\|\mathbf{X}_{ik}\| > \delta\}) \xrightarrow{\delta \rightarrow \infty} 0.$$

Concerning $(\bar{\mathbf{X}}_{i,s}^*)^2 - (\bar{\mathbf{Y}}_{i,s}^*)^2$, we first consider $\bar{\mathbf{X}}_{i,s}^* - \bar{\mathbf{Y}}_{i,s}^* = n_i^{-1} \sum_{k=1}^{n_i} W_{ik} \xi_{iks}$. It holds that $\mathbb{E}(\bar{\mathbf{X}}_{i,s}^* - \bar{\mathbf{Y}}_{i,s}^* \mid \mathbf{X}) = 0$ as well as $\text{Var}(\bar{\mathbf{X}}_{i,s}^* - \bar{\mathbf{Y}}_{i,s}^* \mid \mathbf{X}) \xrightarrow[N \rightarrow \infty]{} \text{Var}(\bar{\mathbf{X}}_{i,s} - \bar{\mathbf{Y}}_{i,s}) \xrightarrow{\delta \rightarrow \infty} 0$ and therefore

$$\lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} \Pr(|\bar{\mathbf{X}}_{i,s}^* - \bar{\mathbf{Y}}_{i,s}^*| > \epsilon \mid \mathbf{X}) = 0.$$

The continuous mapping theorem now implies

$$\lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} \Pr((\bar{\mathbf{X}}_{i,s}^* - \bar{\mathbf{Y}}_{i,s}^*)^2 > \epsilon \mid \mathbf{X}) = 0$$

and furthermore

$$(\bar{\mathbf{X}}_{i,s}^* - \bar{\mathbf{Y}}_{i,s}^*)^2 \mathbb{1}\{\bar{\mathbf{X}}_{i,s}^* > \bar{\mathbf{Y}}_{i,s}^*\} \xrightarrow{\text{Pr}} 0.$$

Therefore,

$$0 \leq ((\bar{\mathbf{X}}_{i,s}^*)^2 - (\bar{\mathbf{Y}}_{i,s}^*)^2) \mathbb{1}\{\bar{\mathbf{X}}_{i,s}^* > \bar{\mathbf{Y}}_{i,s}^*\} \rightarrow 0$$

and analogously

$$0 \leq ((\bar{\mathbf{X}}_{i,s}^*)^2 - (\bar{\mathbf{Y}}_{i,s}^*)^2) \mathbb{1}\{\bar{\mathbf{X}}_{i,s}^* < \bar{\mathbf{Y}}_{i,s}^*\} \rightarrow 0$$

and therefore

$$\lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} \Pr(|(\bar{\mathbf{X}}_{i,s}^*)^2 - (\bar{\mathbf{Y}}_{i,s}^*)^2| > \epsilon \mid \mathbf{X}) = 0.$$

Altogether, this implies by the general Markov inequality that

$$\lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} \Pr((*) > \epsilon \mid \mathbf{X}) \leq \lim_{\delta \rightarrow \infty} \limsup_{N \rightarrow \infty} \frac{1}{\epsilon} \mathbb{E}((*) \mid \mathbf{X}) \stackrel{\text{a.s.}}{=} 0,$$

which concludes the proof. \square

Proof of Theorem 3.3

The result for the nonparametric bootstrap can be proved by conditionally following the lines of the proof of Theorem 2.1. First note, that conditional independence of the bootstrap sample $\mathbf{X}_{ik}^\dagger, i = 1, \dots, a, k = 1, \dots, n_i$ together

with the multivariate CLT for the bootstrap given in [4] implies that the conditional distribution of $\sqrt{N}(\overline{\mathbf{X}}_{\bullet}^{\dagger} - \overline{\mathbf{X}}_{\bullet})$ asymptotically, as $N \rightarrow \infty$, coincides with a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = \text{diag}(\kappa_i^{-1} \mathbf{V}_i)$ (almost surely). Moreover, the law of large numbers for the bootstrap (see, e.g., [12]) implies that $\widehat{\mathbf{D}}_N^{\dagger}$ converges almost surely to $\mathbf{D} = \text{diag}(\kappa_i^{-1} \sigma_{is}^2)$. We can thus conclude as in the proof of Theorem 2.1: $\mathbf{T} \widehat{\mathbf{D}}_N^{\dagger} \mathbf{T} \rightarrow \mathbf{T} \mathbf{D} \mathbf{T}$ holds almost surely and since there is finally no rank jump in this convergence we also obtain $(\widehat{\mathbf{T} \mathbf{D}}_N^{\dagger} \mathbf{T})^+ \xrightarrow{\text{Pr}} (\mathbf{T} \mathbf{D} \mathbf{T})^+$ almost surely. Putting these ingredients together with the continuous mapping theorem finally proves the convergence. \square

Proofs of the results in Section 4.1

Theorems 3.1 – 3.3 directly imply that the corresponding bootstrap tests $\varphi^* = \mathbb{1}\{Q_N > c_{1-\alpha}^*\}$ asymptotically keep the pre-assigned level α , since $c_{1-\alpha}^*$ is the $(1 - \alpha)$ -quantile of the (conditional) bootstrap distribution, which, given the data, converges weakly to the null distribution of Q_N in probability.

For local alternatives $H_1 : \mathbf{T}\boldsymbol{\mu} = \sqrt{N}^{-1} \mathbf{T}\boldsymbol{\nu}$, $\boldsymbol{\nu} \in \mathbb{R}^{ad}$, it holds that

$$\sqrt{N} \mathbf{T} \overline{\mathbf{X}}_{\bullet} = \sqrt{N} \mathbf{T} (\overline{\mathbf{X}}_{\bullet} - \boldsymbol{\mu}) + \mathbf{T} \boldsymbol{\nu}$$

has, asymptotically, as $N \rightarrow \infty$, a multivariate normal distribution with mean $\mathbf{T}\boldsymbol{\nu}$ and covariance matrix $\mathbf{T}\mathbf{\Sigma}\mathbf{T}$ and thus Q_N converges to $\zeta(\mathbf{T} \mathbf{D} \mathbf{T})^+ \zeta$, where $\zeta \sim \mathcal{N}(\mathbf{T}\boldsymbol{\nu}, \mathbf{T}\mathbf{\Sigma}\mathbf{T})$, by using (8.1) again. Thus, Theorems 3.1 – 3.3 imply that the bootstrap tests have the same asymptotic power as $\varphi_N = \mathbb{1}\{Q_N > c_{1-\alpha}\}$ and the asymptotic relative efficiency of the bootstrap tests φ^* compared to φ_N is 1 in this situation. \square

Proof of the results in Section 4.2

In order to derive a simultaneous contrast test formulated in the maximum statistic, we need to analyze the asymptotic joint distribution of the vector of test statistics $\mathbf{Q} = (Q_N^1, \dots, Q_N^q)^{\top}$. First note that \mathbf{Q} can be re-written as

$$\sqrt{N}(\mathbf{H}\overline{\mathbf{X}}_{\bullet} - \mathbf{H}\boldsymbol{\mu})^{\top} \sqrt{N} \text{diag}(\mathbf{h}_{\ell}^{\top} \overline{\mathbf{X}}_{\bullet}) \text{diag}(\mathbf{h}_{\ell}^{\top} \widehat{\mathbf{D}}_N \mathbf{h}_{\ell})^{-1}. \quad (8.6)$$

The last diagonal matrix converges to $\text{diag}(\mathbf{h}_{\ell}^{\top} \mathbf{D} \mathbf{h}_{\ell})^{-1}$, while the first part can be viewed as $\psi(\sqrt{N}(\mathbf{H}\overline{\mathbf{X}}_{\bullet} - \mathbf{H}\boldsymbol{\mu}))$ for a continuous function ψ . Due to the results above, $\sqrt{N}(\mathbf{H}\overline{\mathbf{X}}_{\bullet} - \mathbf{H}\boldsymbol{\mu})$ converges to a multivariate normally distributed random vector Ξ with mean $\mathbf{0}$ and covariance matrix $\mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\top}$. Thus, (8.6) converges in distribution to $\psi(\Xi) \cdot \text{diag}(\mathbf{h}_{\ell}^{\top} \mathbf{D} \mathbf{h}_{\ell})^{-1}$ due to the continuous mapping theorem and Slutsky. The same distributional convergence also holds for the bootstrapped test statistic $\mathbf{Q}^* = (Q_N^{1,*}, \dots, Q_N^{q,*})^{\top}$ given the data in probability due to Theorems 3.1 – 3.3 (they imply that $\widehat{\sigma}_{is}^*$ are consistent estimates for σ_{is} ($i = 1, \dots, a$, $s = 1, \dots, d$) and that $\sqrt{N} \mathbf{H} \overline{\mathbf{X}}_{\bullet}^*$ converges to Ξ in distribution given the data in probability). Since max is continuous, it thus follows that

$$\sup_x |\Pr\{\max(Q_N^1, \dots, Q_N^q) \leq x\} - \Pr\{\max(Q_N^{1,*}, \dots, Q_N^{q,*}) \leq x | \mathbf{X}\}| \xrightarrow{\text{Pr}} 0,$$

which concludes the proof. \square

In future research it will be investigated which method performs preferably for the derivation of simultaneous confidence intervals.

9. Further simulation results

9.1. One-way layout

In this section, we present some additional simulation results for different distributions. The simulation scenarios are the same as in the paper, but we have excluded the WTS with χ^2 -approximation, the wild and the nonparametric bootstrap of the MATS here.

The results are displayed in Tables 7 – 9 for the lognormal, t_3 and double-exponential distribution, respectively. The parametric bootstrap of the MATS keeps the pre-assigned α -level very well for the t_3 and the double-exponential distribution. Note that the validity of the parametric bootstrap of the WTS has not yet been proven for the t_3 distribution, since fourth moments do not exist in this case. With lognormally distributed data and negative pairing (setting 3 and 4 with $\mathbf{n} = (20, 10)^\top$), all procedures show a liberal behavior.

9.2. Two-way layout

To analyze the behavior of our methods in a setting with two crossed factors A and B , we considered the following simulation design, which is again adapted from [23]. We simulated a 2×2 design with sample sizes $\mathbf{n}^{(1)} = (n_{11}^{(1)}, n_{12}^{(1)}, n_{21}^{(1)}, n_{22}^{(1)})^\top = (7, 10, 13, 16)^\top$, $\mathbf{n}^{(2)} = (10, 10, 10, 10)^\top$, $\mathbf{n}^{(3)} = (16, 13, 10, 7)^\top$, $\mathbf{n}^{(4)} = (20, 20, 20, 20)^\top$. The covariance settings were chosen similar to the one-way layout above as:

$$\begin{aligned} \text{Setting 8:} \quad & \mathbf{V}_{ij} = \mathbf{I}_d + 0.5(\mathbf{J}_d - \mathbf{I}_d), \quad i, j = 1, 2, \\ \text{Setting 9:} \quad & \mathbf{V}_{ij} = \left((0.6)^{|r-s|} \right)_{r,s=1}^d, \quad i, j = 1, 2, \\ \text{Setting 10:} \quad & \mathbf{V}_{ij} = \mathbf{I}_d \cdot \ell + 0.5(\mathbf{J}_d - \mathbf{I}_d), \quad \ell = 1, \dots, 4, \\ \text{Setting 11:} \quad & \mathbf{V}_{ij} = \left((0.6)^{|r-s|} \right)_{r,s=1}^d + \mathbf{I}_d \cdot \ell, \quad \ell = 1, \dots, 4. \end{aligned}$$

Again, setting 10 and 11 combined with sample sizes $\mathbf{n}^{(2)}$ and $\mathbf{n}^{(3)}$ represent settings with positive and negative pairing, respectively. In this scenario, we consider three different null hypotheses of interest:

- (1) The hypothesis of *no effect of factor A*

$$H_0^\mu(A) : \{ \bar{\boldsymbol{\mu}}_{\cdot 1} = \bar{\boldsymbol{\mu}}_{\cdot 2} \} = \{ \mathbf{H}_A \boldsymbol{\mu} = \mathbf{0} \},$$

- (2) The hypothesis of *no effect of factor B*

$$H_0^\mu(B) : \{ \bar{\boldsymbol{\mu}}_{1 \cdot} = \bar{\boldsymbol{\mu}}_{2 \cdot} \} = \{ \mathbf{H}_B \boldsymbol{\mu} = \mathbf{0} \},$$

- (3) The hypothesis of *no $A \times B$ interaction effect*

$$H_0^\mu(AB) : \{ (\mathbf{P}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d) \boldsymbol{\mu} = \mathbf{0} \},$$

where $\mathbf{H}_A = \mathbf{P}_a \otimes b^{-1} \mathbf{J}_b \otimes \mathbf{I}_d$ and $\mathbf{H}_B = a^{-1} \mathbf{J}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d$.

The simulation results for factor A and B as well as the interaction between the factors are in Tables 10 – 12, respectively. Due to the larger total sample size N in this scenario, the asymptotic results come into play and therefore all methods lead to more accurate results than in the one-way layout. The behavior of the tests is similar to the one-way layout: Both MATS and WTS with the parametric bootstrap approach control the type-I error accurately in most scenarios. The χ^2 -approximation of the WTS is very liberal again, while the wild bootstrap MATS also shows a slightly liberal behavior. In situations with negative pairing (covariance setting 10 and 11 with sample size vector $\mathbf{n}^{(3)}$), the parametric bootstrap MATS improves the slightly liberal behavior of the WTS, see e.g., Table 10 for the normal distribution, where the WTS (PBS) leads to a type-I error of 6.1%, while the MATS (PBS) is at 4.9%. The nonparametric bootstrap is again slightly more conservative than the parametric bootstrap, see, e.g., Table 11 with χ_3^2 -distribution and covariance settings S10 and S11. For the interaction hypothesis with χ_3^2 -distribution, the WTS (PBS), MATS (PBS) and MATS (NPBS) show more conservative results than for the hypotheses about the main effects.

Table 7: Type-I error rates in % (nominal level $\alpha = 5\%$) for the parametric bootstrap (PBS) of the WTS and the MATS in the one-way layout for the log-normal distribution.

d	Cov	n	WTS (PBS)	MATS (PBS)
$d = 4$	S1	(10, 10)	2.0	3.5
		(10, 20)	4.4	5.3
		(20, 10)	4.3	5.5
		(20, 20)	3.5	4.0
	S2	(10, 10)	1.9	3.2
		(10, 20)	4.5	4.9
		(20, 10)	4.4	5.4
		(20, 20)	3.4	3.9
	S3	(10, 10)	6.8	6.1
		(10, 20)	4.2	3.6
		(20, 10)	14.9	13.8
		(20, 20)	8.4	6.7
	S4	(10, 10)	7.0	6.2
		(10, 20)	4.2	3.6
		(20, 10)	15.4	13.5
		(20, 20)	8.7	6.8
$d = 8$	S1	(10, 10)	2.9	4.1
		(10, 20)	3.9	5.6
		(20, 10)	4.2	5.3
		(20, 20)	3.0	4.0
	S2	(10, 10)	2.8	2.5
		(10, 20)	4.0	4.8
		(20, 10)	4.1	4.1
		(20, 20)	3.0	3.1
	S3	(10, 10)	6.3	6.0
		(10, 20)	3.5	3.6
		(20, 10)	15.0	12.7
		(20, 20)	7.7	6.7
	S4	(10, 10)	7.0	6.0
		(10, 20)	3.6	2.8
		(20, 10)	15.5	13.6
		(20, 20)	8.0	7.0

Table 8: Type-I error rates in % (nominal level $\alpha = 5\%$) for the parametric bootstrap (PBS) of the WTS and the MATS in the one-way layout for the t_3 distribution.

d	Cov	n	WTS (PBS)	MATS (PBS)
$d = 4$	S1	(10, 10)	3.6	4.4
		(10, 20)	4.8	4.8
		(20, 10)	3.9	4.3
		(20, 20)	4.0	4.5
	S2	(10, 10)	3.7	4.1
		(10, 20)	4.7	4.6
		(20, 10)	4.1	4.2
		(20, 20)	4.0	4.3
	S3	(10, 10)	4.1	3.1
		(10, 20)	4.3	4.2
		(20, 10)	4.9	3.3
		(20, 20)	3.9	3.8
	S4	(10, 10)	4.2	3.1
		(10, 20)	4.4	3.9
		(20, 10)	4.9	3.2
		(20, 20)	4.0	4.0
$d = 8$	S1	(10, 10)	3.5	4.7
		(10, 20)	4.8	4.7
		(20, 10)	5.0	4.2
		(20, 20)	3.9	4.7
	S2	(10, 10)	3.4	3.8
		(10, 20)	4.8	4.1
		(20, 10)	5.0	3.5
		(20, 20)	3.9	3.9
	S3	(10, 10)	5.2	3.2
		(10, 20)	3.5	4.2
		(20, 10)	8.6	2.7
		(20, 20)	4.0	3.6
	S4	(10, 10)	5.0	2.6
		(10, 20)	3.6	3.6
		(20, 10)	8.3	2.5
		(20, 20)	3.8	3.2

Table 9: Type-I error rates in % (nominal level $\alpha = 5\%$) for the parametric bootstrap (PBS) of the WTS and the MATS in the one-way layout for the double-exponential distribution.

d	Cov	n	WTS (PBS)	MATS (PBS)
$d = 4$	S1	(10, 10)	4.1	4.5
		(10, 20)	5.0	4.8
		(20, 10)	4.1	4.2
		(20, 20)	5.1	5.6
	S2	(10, 10)	4.1	4.5
		(10, 20)	4.9	4.8
		(20, 10)	4.0	4.2
		(20, 20)	5.1	5.4
	S3	(10, 10)	4.5	3.3
		(10, 20)	4.3	4.4
		(20, 10)	5.1	3.2
		(20, 20)	4.7	4.9
	S4	(10, 10)	4.6	3.7
		(10, 20)	4.6	4.3
		(20, 10)	5.0	3.3
		(20, 20)	4.7	4.7
$d = 8$	S1	(10, 10)	3.4	4.6
		(10, 20)	5.2	4.8
		(20, 10)	4.7	4.9
		(20, 20)	4.3	4.7
	S2	(10, 10)	3.6	3.4
		(10, 20)	5.3	4.3
		(20, 10)	4.6	4.3
		(20, 20)	4.3	4.3
	S3	(10, 10)	4.8	3.0
		(10, 20)	4.1	4.2
		(20, 10)	8.1	3.0
		(20, 20)	5.0	3.6
	S4	(10, 10)	4.7	2.5
		(10, 20)	4.2	3.7
		(20, 10)	7.9	2.5
		(20, 20)	4.9	3.5

Table 10: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) when testing for an effect of factor A in a two-way layout with $d = 4$ dimensional observations.

distr	Cov	n	WTS (χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS (NPBS)
normal	S8	$\mathbf{n}^{(1)}$	10.4	4.7	6.1	4.6	4.4
		$\mathbf{n}^{(2)}$	9.5	4.5	6.4	5.1	4.9
		$\mathbf{n}^{(3)}$	11.2	5.4	6.6	5.3	4.8
		$\mathbf{n}^{(4)}$	6.8	5.0	5.6	5.3	5.2
	S9	$\mathbf{n}^{(1)}$	10.4	4.8	6.2	4.7	4.4
		$\mathbf{n}^{(2)}$	9.5	4.4	6.6	5.3	4.9
		$\mathbf{n}^{(3)}$	11.2	5.5	6.5	4.9	4.6
		$\mathbf{n}^{(4)}$	6.8	4.9	5.7	5.1	5.1
	S10	$\mathbf{n}^{(1)}$	9	4.7	5.8	4.5	4.2
		$\mathbf{n}^{(2)}$	11.1	5.2	6.4	4.8	4.3
		$\mathbf{n}^{(3)}$	14.8	6.1	7.3	4.9	4
		$\mathbf{n}^{(4)}$	7.4	5	5.5	4.7	4.7
	S11	$\mathbf{n}^{(1)}$	9.2	4.7	5.8	4.4	4.1
		$\mathbf{n}^{(2)}$	10.2	4.9	6	4.4	4.1
		$\mathbf{n}^{(3)}$	13.3	6	7	4.8	4.2
		$\mathbf{n}^{(4)}$	7	5.1	5.4	4.6	4.5
χ_3^2	S8	$\mathbf{n}^{(1)}$	9.7	4.6	6.7	5.1	4.5
		$\mathbf{n}^{(2)}$	9.2	3.9	6.7	5.1	4.5
		$\mathbf{n}^{(3)}$	10.7	4.7	6.7	5	4.4
		$\mathbf{n}^{(4)}$	6.6	4.3	4.7	4.3	4.1
	S9	$\mathbf{n}^{(1)}$	9.7	4.6	6.5	4.7	4.1
		$\mathbf{n}^{(2)}$	9.2	3.9	7	5.1	4.3
		$\mathbf{n}^{(3)}$	10.7	4.8	6.6	5	4.2
		$\mathbf{n}^{(4)}$	6.6	4.5	4.9	4.4	4.2
	S10	$\mathbf{n}^{(1)}$	8.8	4.5	6.1	4.3	3.6
		$\mathbf{n}^{(2)}$	11.2	5	7.6	5.2	4.1
		$\mathbf{n}^{(3)}$	16	7	9.4	6.5	5.3
		$\mathbf{n}^{(4)}$	7.7	5.5	5.6	4.9	4.5
	S11	$\mathbf{n}^{(1)}$	8.5	4.2	5.6	3.7	3
		$\mathbf{n}^{(2)}$	9.9	4.3	6.7	4.7	3.8
		$\mathbf{n}^{(3)}$	13.9	6.4	8.7	5.9	4.4
		$\mathbf{n}^{(4)}$	7.2	5	5.6	4.9	4.3

Table 11: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) when testing for an effect of factor B in a two-way layout with $d = 4$ dimensional observations.

distr	Cov	n	WTS(χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS (NPBS)
normal	S8	$\mathbf{n}^{(1)}$	10.7	5.1	6.9	5.3	4.8
		$\mathbf{n}^{(2)}$	9.2	4.6	6	4.8	4.5
		$\mathbf{n}^{(3)}$	10.2	4.6	6.3	5	4.7
		$\mathbf{n}^{(4)}$	6.5	4.4	5.0	4.7	4.7
	S9	$\mathbf{n}^{(1)}$	10.7	5.2	6.6	5.1	4.8
		$\mathbf{n}^{(2)}$	9.2	4.7	6.1	4.9	4.8
		$\mathbf{n}^{(3)}$	10.2	4.5	6.2	4.5	4.3
		$\mathbf{n}^{(4)}$	6.5	4.5	5	4.5	4.4
	S10	$\mathbf{n}^{(1)}$	8.7	4.9	5.8	4.7	4.4
		$\mathbf{n}^{(2)}$	10.3	4.6	5.9	4.5	4.1
		$\mathbf{n}^{(3)}$	13.4	5.3	6.4	4	3.3
		$\mathbf{n}^{(4)}$	6.9	4.6	5.2	4.7	4.6
	S11	$\mathbf{n}^{(1)}$	9	4.9	5.7	4.4	4.3
		$\mathbf{n}^{(2)}$	9.4	4.7	6.1	4.4	4.1
		$\mathbf{n}^{(3)}$	12.4	5.1	6.4	4.1	3.4
		$\mathbf{n}^{(4)}$	6.4	4.6	5.2	4.4	4.3
χ^2_3	S8	$\mathbf{n}^{(1)}$	10.3	4.9	6.4	4.9	4.3
		$\mathbf{n}^{(2)}$	9.4	4	6.2	4.8	4.3
		$\mathbf{n}^{(3)}$	10.7	4.5	5.9	4.7	4.1
		$\mathbf{n}^{(4)}$	7.2	4.7	5.1	4.7	4.5
	S9	$\mathbf{n}^{(1)}$	10.3	4.9	6.8	4.8	4.2
		$\mathbf{n}^{(2)}$	9.4	4	6.3	4.6	4
		$\mathbf{n}^{(3)}$	10.7	4.5	6	4.4	4
		$\mathbf{n}^{(4)}$	7.2	4.9	5.4	5	4.8
	S10	$\mathbf{n}^{(1)}$	9	4.5	6.1	4.4	3.8
		$\mathbf{n}^{(2)}$	10.6	4.5	6.5	4.4	3.2
		$\mathbf{n}^{(3)}$	13.7	5.4	7	4.4	3.3
		$\mathbf{n}^{(4)}$	7.7	5.1	6	5.1	4.6
	S11	$\mathbf{n}^{(1)}$	9.2	4.6	6.2	4.4	3.7
		$\mathbf{n}^{(2)}$	9.7	4.2	6.3	4.1	2.9
		$\mathbf{n}^{(3)}$	12.4	5.1	6.4	4.2	3.1
		$\mathbf{n}^{(4)}$	7.3	5.2	6	5.1	4.3

Table 12: Type-I error rates in % (nominal level $\alpha = 5\%$) for the WTS with χ^2 -approximation and parametric bootstrap (PBS) and the MATS with wild bootstrap (wild), parametric bootstrap (PBS) and nonparametric bootstrap (NPBS) when testing the interaction hypothesis in a two-way layout with $d = 4$ dimensional observations.

distr	Cov	n	WTS(χ^2)	WTS (PBS)	MATS (wild)	MATS (PBS)	MATS(NPBS)
normal	S8	$n^{(1)}$	10.8	4.8	6.9	5.4	5
		$n^{(2)}$	10	4.7	6.5	5.4	5.2
		$n^{(3)}$	10	4.7	6.8	5.3	5
		$n^{(4)}$	6.3	4.3	5.0	4.6	4.6
	S9	$n^{(1)}$	10.8	4.9	6.9	5.1	4.9
		$n^{(2)}$	10	4.8	6.3	5.1	4.8
		$n^{(3)}$	10	4.8	6.4	5	4.8
		$n^{(4)}$	6.3	4.3	5	4.7	4.4
	S10	$n^{(1)}$	8.8	5.1	6.2	4.8	4.5
		$n^{(2)}$	10.9	4.8	6.6	4.9	4.5
		$n^{(3)}$	13.8	5.7	7.5	4.9	4.1
		$n^{(4)}$	6.9	4.5	5	4.3	4.2
	S11	$n^{(1)}$	9.1	4.9	6.4	4.8	4.4
		$n^{(2)}$	10.1	4.8	6.6	4.7	4.5
		$n^{(3)}$	12.9	5.4	7.3	4.7	4.1
		$n^{(4)}$	6.5	4.5	4.9	4.4	4.2
χ^2_3	S8	$n^{(1)}$	9.5	4.3	6.3	4.8	4.3
		$n^{(2)}$	9.4	4.3	6.3	4.8	4.4
		$n^{(3)}$	9.8	4.1	5.9	4.6	4
		$n^{(4)}$	6.6	4.6	4.8	4.4	4.2
	S9	$n^{(1)}$	9.5	4.3	6.3	4.6	4.1
		$n^{(2)}$	9.4	4.1	6.5	4.8	4.2
		$n^{(3)}$	9.8	4.2	5.9	4.6	4
		$n^{(4)}$	6.6	4.6	5.1	4.6	4.4
	S10	$n^{(1)}$	7.9	4.1	5.9	4.3	3.6
		$n^{(2)}$	10.3	4.2	5.9	3.9	2.9
		$n^{(3)}$	13	4.3	6.4	3.7	2.4
		$n^{(4)}$	6.8	4.7	5.5	4.6	4
	S11	$n^{(1)}$	8.2	4	5.8	4	3.4
		$n^{(2)}$	9.7	4.2	6	3.9	2.9
		$n^{(3)}$	11.7	4.3	6.2	3.5	2.4
		$n^{(4)}$	6.7	4.8	5.4	4.4	4.2

10. Another data example

As our second data example, we consider cardiological measurements in the left ventricle of 188 healthy patients, that were recorded at the University clinic Ulm, Germany. Variables of interest are the peak systolic and diastolic strain rate (PSSR and PDSR, respectively), measured in circumferential direction, the end systolic and diastolic volume (ESV and EDV, respectively) as well as the stroke volume (SV). The empirical covariance matrix is singular in this example, since stroke volume is calculated as the difference between end diastolic volume and end systolic volume. The empirical covariance matrices can be found in Section 10.1 below. Note that this data example is somewhat artificial, since the reason for the singularity of the empirical covariance matrix is known and one would usually drop one of the three variables involved in the collinearity. We consider a one-way layout analyzing the factor 'gender' (female vs. male). Some descriptive statistics of the measurements for this factor are displayed in Table 5. Boxplots of the systolic and diastolic measurements are in Figures 4 and 5, respectively.

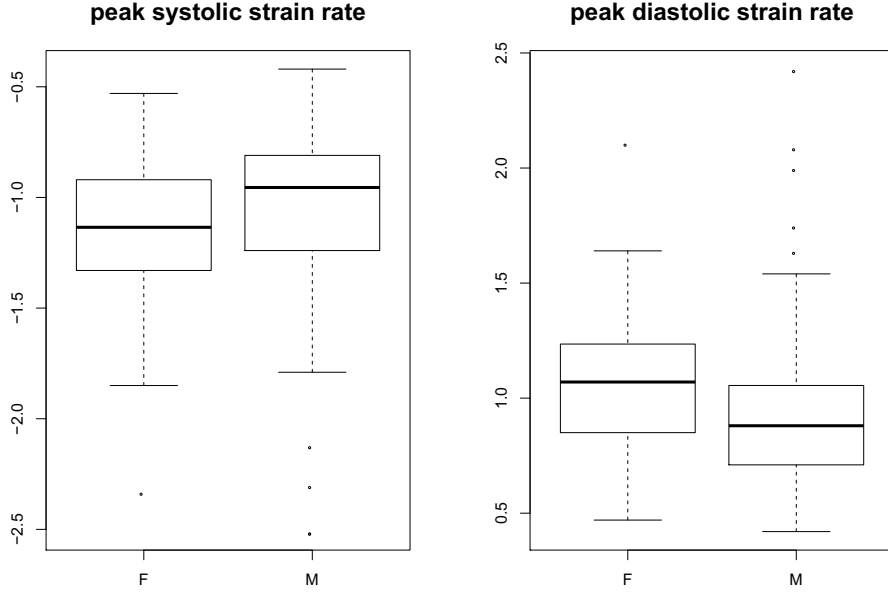


Figure 4: Boxplots of the systolic and diastolic peak strain rate for female and male patients.

Table 13: Descriptive statistics of the cardiology data. Volume measurements (EDV, ESV and SV) are in ml , peak strain rate measurements are in $1/sec$.

Gender	n	Mean					Sd				
		EDV	ESV	SV	PSSR	PDSR	EDV	ESV	SV	PSSR	PDSR
female	92	124.37	41.39	82.98	-1.16	1.07	25.25	13.38	15.77	0.32	0.30
male	96	157.02	54.98	102.04	-1.07	0.94	31.02	15.98	18.70	0.39	0.35

We now want to analyze, whether there is a significant difference in the multivariate means for female and male patients. The null hypothesis of interest thus is $H_0^{(1)} : \{(\mathbf{P}_2 \otimes \mathbf{I}_5)\boldsymbol{\mu} = \mathbf{0}\}$. Since the covariance matrix is singular in this example, we cannot apply the Wald-type test. Thus, we consider the parametric bootstrap approach of the MATS which yielded the best results in the simulation study. Computation of the MATS results in a value of $Q_N = 171.0011$ and the parametric bootstrap routine with 10,000 bootstrap runs gives a p -value of $p < 0.0001$, implying that there is indeed a significant effect of gender on the measurements.

In a second step we want to derive a confidence region for the factor 'Gender'. Here, we restrict our analyses to the strain rate measurements in order to be able to plot the confidence ellipsoids for the contrast of interest. That is, we consider the null hypothesis $H_0^{(2)} : \{\mathbf{T}\boldsymbol{\mu} = \mathbf{0}\} = \{\mu_{11} - \mu_{21} = \mu_{12} - \mu_{22}\} = \{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}\}$, where μ_{ij} is the corresponding mean value of measurement j (systolic vs. diastolic measurement) in group i (female vs. male) and

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

The parametric bootstrap procedure with 10,000 bootstrap runs leads to a p -value of $p = 0.0146$ for the MATS, i.e., there is a significant effect of gender on the peak strain rate. We can now construct a confidence ellipsoid as described in Section 4.2 based on the parametric bootstrap quantile $c_{1-\alpha}^*$. Therefore, we need to compute the eigenvalues λ_j and eigenvectors $\mathbf{e}_j, j = 1, 2$ of $\mathbf{T}\widehat{\mathbf{D}}_N\mathbf{T}$. The ellipse is centered at $\mathbf{T}\bar{\mathbf{X}} = (-0.097, 0.126)^\top$. For the eigendecomposition of $\mathbf{T}\widehat{\mathbf{D}}_N\mathbf{T}$, we obtain $\boldsymbol{\lambda} = (0.508, 0.412)$ as well as $\mathbf{e}_1 = (-1, 0)^\top$ and $\mathbf{e}_2 = (0, -1)^\top$, that is,

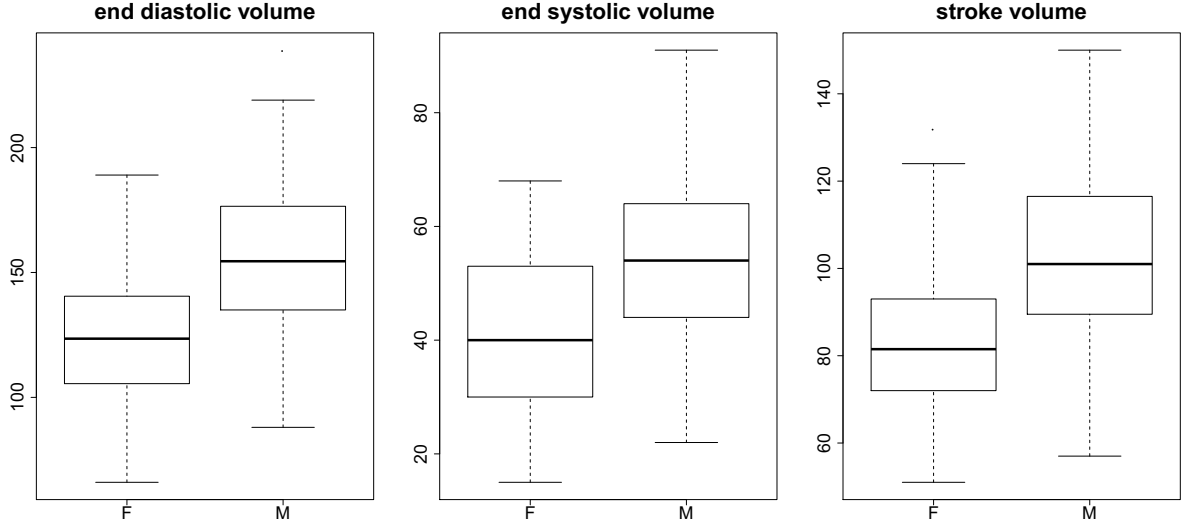


Figure 5: Boxplots of the end diastolic and systolic volume as well as stroke volume for female and male patients.

the confidence ellipse extends $\sqrt{\lambda_1 \cdot c_{1-\alpha}^* / N} = 0.013$ units in the direction of e_1 and 0.012 units in the direction of e_2 . The corresponding ellipse is displayed in Figure 6. It turns out that female patients have a lower systolic peak strain rate but a higher diastolic peak strain rate than male patients, a finding that is confirmed by the descriptive analysis in Table 5 and Figure 4.

10.1. Covariance Matrices

For convenience, we included the estimated covariance matrices of the data example for female and male patients and the five outcome variables end diastolic volume (EDV), end systolic volume (ESV), stroke volume (SV), peak systolic strain rate (PSSR) and peak diastolic strain rate (PDSR) in this section. Covariance matrices, rounded to three decimals, for female patients (EDV, ESV, SV, PSSR, PDSR):

$$\begin{pmatrix} 6.931 & 3.087 & 3.844 & 0.022 & -0.012 \\ 3.087 & 1.947 & 1.140 & 0.018 & -0.013 \\ 3.844 & 1.140 & 2.704 & 0.004 & 0.002 \\ 0.022 & 0.018 & 0.004 & 0.001 & -0.001 \\ -0.012 & -0.013 & 0.002 & -0.001 & 0.001 \end{pmatrix}$$

and male patients:

$$\begin{pmatrix} 10.025 & 4.522 & 5.503 & 0.043 & -0.021 \\ 4.522 & 2.661 & 1.862 & 0.027 & -0.018 \\ 5.503 & 1.862 & 3.641 & 0.016 & -0.003 \\ 0.043 & 0.027 & 0.016 & 0.002 & -0.001 \\ -0.021 & -0.018 & -0.003 & -0.001 & 0.001 \end{pmatrix}.$$

10.2. Problem with the ATS

Finally, we want to demonstrate the problems that can arise when using the ATS \tilde{Q}_N in multivariate data. Since the asymptotic distribution of the ATS depends on unknown parameters [5, 16], it is approximated by an \mathcal{F} -distribution. In particular, the scaled statistic

$$F_N = \frac{N}{\widehat{\text{tr}(\mathbf{T}\boldsymbol{\Sigma})}} \overline{\mathbf{X}} \cdot \mathbf{T} \overline{\mathbf{X}}.$$

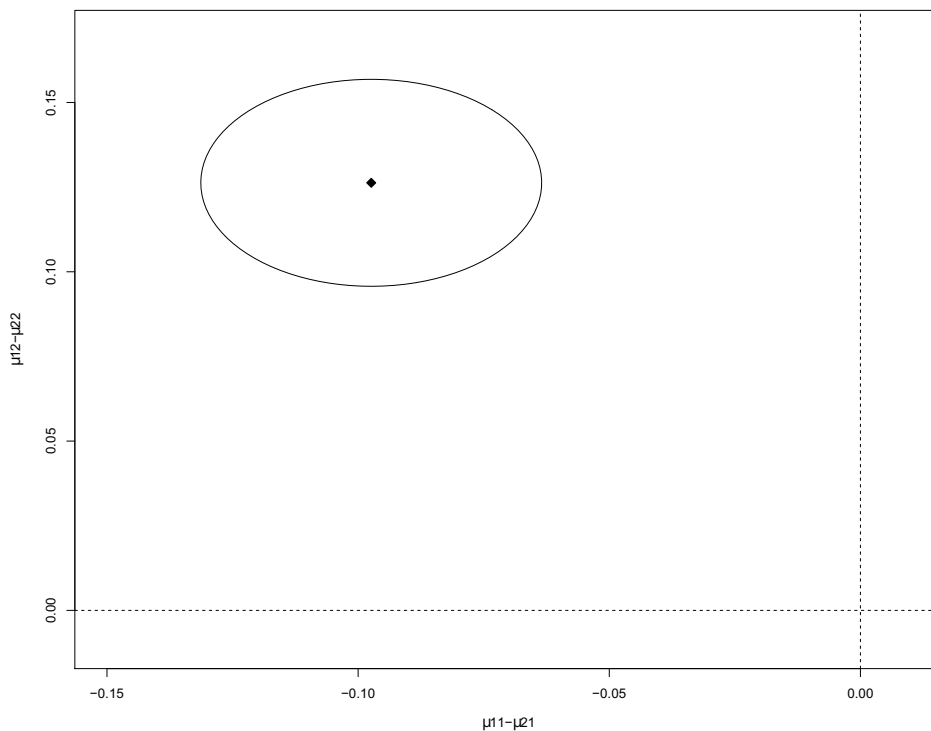


Figure 6: 95 % confidence ellipse for $\mu_1 - \mu_2$. The \blacklozenge denotes the center \overline{TX}_* of the ellipse. Female patients seem to have a lower systolic peak strain rate but a higher diastolic peak strain rate than male patients.

is approximated by an $\mathcal{F}(\widehat{\nu}, \infty)$ -distribution with $\widehat{\nu} = \text{tr}^2(\widehat{T\Sigma}) / \text{tr}(\widehat{T\Sigma})^2$ degrees of freedom. We consider only the peak strain rate measurements. The ATS then results in a test statistic of $F_N = 5.2$, while the corresponding quantile of the \mathcal{F} -distribution is 3.51, resulting in a p -value of 0.02. If we now change the units of the peak systolic strain rate from $1/\text{sec}$ to $1/\text{min}$, the test statistic becomes $F_N = 3.51$ with an \mathcal{F} -quantile of 3.84, resulting in a p -value of 0.06. By changing the units in one component, we have therefore changed the significance of the outcome at 5% level. Thus, the ATS should only be applied if observations are measured on the same scale as in repeated measurements but not to multivariate data in general.

References

- [1] M. S. Bartlett. A note on tests of significance in multivariate analysis. *Mathematical Proceedings of the Cambridge Philosophical Society*, 35(02):180–185, 1939. Cambridge University Press.
- [2] A. C. Bathke, S. W. Harrar, and L. V. Madden. How to compare small multivariate samples using nonparametric tests. *Computational Statistics & Data Analysis*, 52(11):4951–4965, 2008.
- [3] J. Beyersmann, S. D. Termini, and M. Pauly. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 40(3):387–402, 2013.
- [4] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.
- [5] E. Brunner. Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity. *Mathematical Statistics with Applications in Biometry*, 2001.
- [6] E. Brunner, F. Konietzschke, M. Pauly, and M. L. Puri. Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.
- [7] E. Brunner, U. Munzel, and M. L. Puri. Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis*, 70(2):286–317, 1999.
- [8] E. Brunner and M. L. Puri. Nonparametric methods in factorial designs. *Statistical papers*, 42(1):1–52, 2001.

- [9] A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.
- [10] A. C. Cameron and D. L. Miller. A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.
- [11] E. Chung and J. P. Romano. Multivariate and multiple permutation tests. *Journal of Econometrics*, 193(1):76–91, 2016.
- [12] S. Csörgö. On the law of large numbers for the bootstrap mean. *Statistics & probability letters*, 14(1):1–7, 1992.
- [13] R. Davidson and E. Flachaire. The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169, 2008.
- [14] A. P. Dempster. A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 29(4):995–1010, 1958.
- [15] A. P. Dempster. A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16(1):41–50, 1960.
- [16] S. Friedrich, E. Brunner, and M. Pauly. Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153:255–265, 2017.
- [17] S. Friedrich, F. Konietzschke, and M. Pauly. A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, 2016.
- [18] S. W. Harrar and A. C. Bathke. A modified two-factor multivariate analysis of variance: asymptotics and small sample approximations. *Annals of the Institute of Statistical Mathematics*, 64(1):135–165, 2012.
- [19] M. Hasler and L. A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 50(5):793–800, 2008.
- [20] H. Hotelling. A generalized t -test and measure of multivariate dispersion. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.
- [21] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.
- [22] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. 6th edition, Prentice Hall, 2007.
- [23] F. Konietzschke, A. Bathke, S. Harrar, and M. Pauly. Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, 140:291–301, 2015.
- [24] K. Krishnamoorthy and F. Lu. A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation*, 80(8):873–887, 2010.
- [25] D. Lawley. A generalization of fisher’s z test. *Biometrika*, 30(1-2):180–187, 1938.
- [26] D. Lin. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16(8):901–910, 1997.
- [27] C. Liu, A. C. Bathke, and S. W. Harrar. A nonparametric version of wilks’ lambda - asymptotic results and small sample approximations. *Statistics & Probability Letters*, 81(10):1502–1506, 2011.
- [28] R. Y. Liu. Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708, 1988.
- [29] E. Mammen. *When does bootstrap work? Asymptotic results and simulations*. Springer Science & Business Media, 1993.
- [30] R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.
- [31] M. Pauly, D. Ellenberger, and E. Brunner. Analysis of high-dimensional one group repeated measures designs. *Statistics*, 49:1243–1261, 2015.
- [32] F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.
- [33] F. Pesarin and L. Salmaso. A review and some new results on permutation testing for multivariate problems. *Statistics and Computing*, 22(2):639–646, 2012.
- [34] K. Pillai. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, 26(1):117–121, 1955.
- [35] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
- [36] C. Rao and S. Mitra. *Generalized inverse of matrices and its applications*. Wiley New York, 1971.
- [37] Ł. Smaga. Bootstrap methods for multivariate hypothesis testing. *Communications in Statistics-Simulation and Computation*, 2016. Just accepted.
- [38] E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5):641–656, 2008.
- [39] M. S. Srivastava and T. Kubokawa. Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis*, 115:204–216, 2013.
- [40] G. Vallejo and M. Ato. Robust tests for multivariate factorial designs under heteroscedasticity. *Behavior research methods*, 44(2):471–489, 2012.
- [41] G. Vallejo, M. Fernández, and P. E. Livacic-Rojas. Analysis of unbalanced factorial designs with heteroscedastic data. *Journal of Statistical Computation and Simulation*, 80(1):75–88, 2010.
- [42] S. Van Aelst and G. Willems. Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association*, 106(494):706–718, 2011.
- [43] S. Van Aelst and G. Willems. Fast and robust bootstrap for multivariate inference: the R package FRB. *Journal of Statistical Software*, 53(3):1–32, 2013.
- [44] S. S. Wilks. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, 17(3):257–281, 1946.
- [45] C.-F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.
- [46] L.-W. Xu, F.-Q. Yang, S. Qin, et al. A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, 115:172–180, 2013.

Table 5: Descriptive statistics of the data example: Reported are the sample sizes and the 7-dimensional mean vectors for each of the 43 states.

State	n	PSI045214	SEX25214	RHI125214	RHI225214	RHI325214	RHI425214	RHI525214
AK	29	25404.55	45.73	52.51	1.92	31.89	5.68	0.55
AL	67	72378.76	51.26	68.26	28.66	0.80	0.73	0.11
AR	75	39551.59	50.54	80.41	16.13	0.89	0.78	0.10
AZ	15	448765.60	49.63	78.99	2.33	14.76	1.45	0.20
CA	58	669008.62	49.52	81.59	3.57	3.16	7.50	0.39
CO	64	83685.41	48.05	92.62	1.81	2.06	1.30	0.12
FL	67	296914.88	48.60	80.61	15.00	0.73	1.70	0.10
GA	159	63505.30	50.31	68.19	28.36	0.49	1.30	0.12
IA	99	31385.11	50.15	95.69	1.43	0.46	1.13	0.08
ID	44	37146.91	49.26	94.36	0.58	2.03	0.82	0.16
IL	102	126280.20	49.93	91.67	5.30	0.35	1.23	0.03
IN	92	71704.95	50.20	94.42	2.85	0.36	0.99	0.04
KS	105	27657.34	49.74	93.64	2.11	1.23	0.91	0.08
KY	120	36778.81	50.25	93.87	3.86	0.29	0.57	0.06
LA	64	72651.19	49.95	64.63	32.08	0.84	0.94	0.05
MD	24	249016.96	50.88	73.14	20.58	0.44	3.41	0.10
ME	16	83130.56	50.89	95.67	0.98	0.87	0.89	0.02
MI	83	119396.11	49.67	91.18	4.05	1.69	1.03	0.03
MIN	87	62726.13	49.88	92.93	1.59	2.24	1.47	0.06
MO	115	52726.86	50.11	93.08	3.71	0.62	0.74	0.12
MS	82	36513.16	51.01	56.48	41.21	0.68	0.57	0.04
MT	56	18278.20	49.16	88.81	0.41	8.00	0.48	0.05
NC	100	99439.64	50.82	74.15	20.80	1.93	1.26	0.10
ND	53	13952.49	48.69	90.26	0.79	6.79	0.59	0.04
NE	93	20231.22	49.82	95.43	0.93	1.73	0.59	0.07
NJ	21	425627.38	51.06	76.97	13.19	0.56	7.16	0.10
NM	33	63199.15	49.37	86.02	1.83	8.70	1.11	0.14
NV	17	167005.82	47.41	87.34	2.88	4.27	2.26	0.31
NY	62	318487.53	50.22	87.43	6.98	0.70	2.90	0.06
OH	88	131751.85	50.38	92.70	4.33	0.28	0.97	0.02
OK	77	50364.30	49.84	78.41	3.78	11.18	0.86	0.14
OR	36	110284.42	50.04	91.49	0.89	2.48	1.77	0.30
PA	67	190853.87	50.03	91.99	4.86	0.27	1.46	0.03
SC	46	105053.96	50.93	60.75	36.10	0.64	0.93	0.09
SD	66	12926.89	49.48	82.51	0.78	14.02	0.63	0.04
TN	95	68940.55	50.46	89.75	7.50	0.45	0.73	0.05
TX	254	106129.76	49.20	89.24	6.82	1.17	1.16	0.08
UT	29	101479.38	49.12	92.88	0.66	3.30	1.04	0.36
VA	134	62136.49	50.25	75.67	18.84	0.51	2.09	0.07
WA	39	181064.87	49.85	88.87	1.61	3.01	2.79	0.34
WI	72	79966.17	49.64	92.55	1.71	2.96	1.31	0.04
WV	55	33642.29	50.12	95.57	2.38	0.25	0.49	0.01
WY	23	25397.96	49.04	94.04	1.17	2.13	0.83	0.10

Appendix: The GFD package

Friedrich, S., Konietschke, F. and Pauly, M. (2017). GFD: An R package for the Analysis of General Factorial Designs. *Journal of Statistical Software, Code Snippets*, **79**(1), 1–18, DOI: 10.18637/jss.v079.c01.

This work is licensed under the licenses

*Paper: Creative Commons Attribution 3.0 Unported License
(<https://creativecommons.org/licenses/by/3.0/de/legalcode>)*

Code: GNU General Public License (at least one of version 2 or version 3) or a GPL-compatible license.



GFD: An R Package for the Analysis of General Factorial Designs

Sarah Friedrich
Ulm University

Frank Konietzschke
UT Dallas

Markus Pauly
Ulm University

Abstract

Factorial designs are widely used tools for modeling statistical experiments in all kinds of disciplines, e.g., biology, psychology, econometrics and medicine. For testing null hypotheses in this framework, ANOVA methods are widely used. However, the corresponding F tests are only valid for normally distributed data with equal variances, two assumptions which are often not met in practice. The R package **GFD** provides an implementation of the Wald-type statistic (WTS), the ANOVA-type statistic (ATS) and a studentized permutation version of the WTS. Both the WTS and the permuted WTS do not require normally distributed data or variance homogeneity, whereas the ATS assumes normality. All methods are available for general crossed or nested designs and all main and interaction effects can be plotted. Additionally, the package is equipped with an optional graphical user interface to facilitate application for a wide range of users. We illustrate the implemented methods for a range of different designs.

Keywords: factorial designs, non-normal data, heteroscedasticity, permutation, R, GUI.

1. Introduction

Originated in the agricultural sciences factorial designs are widely used tools for modeling statistical experiments in a variety of disciplines, e.g., biology, econometrics, medicine, ecology or psychology. For testing null hypotheses formulated in terms of means, analysis-of-variance (ANOVA) methods are well known, and preferred for making statistical inference. ANOVA methods are implemented in R within the function `aov` in the R package **stats** (R Core Team 2017). The `anova` function in this package as well as `Anova` in the **car** package (Fox and Weisberg 2011) provide clearly arranged ANOVA tables for fitted models. The corresponding F tests, however, are only valid under the assumption of normally distributed errors and equal variances across the different treatment groups. These assumptions are hard to verify in practice and often not met. A violation usually inflates the type-I or -II errors of the F statistics.

The accuracy of the F tests depends on the actual data distributions, sample size allocations, and the degree of variance heteroscedasticity. For normally distributed errors, several procedures for heteroscedastic data have been proposed, e.g., the generalized Welch-James test (Johansen 1980), the approximate degrees of freedom test (Zhang 2012) or the ANOVA-type test proposed by Brunner, Dette, and Munk (1997), see also Bathke, Schabenberger, Tobias, and Madden (2009). These tests control the type-1 error level in heteroscedastic designs quite accurately, but are in general not asymptotically exact for non-normal data. In comparison to that, the Wald-type statistic, see Equation 2 below, is asymptotically exact in general factorial designs without assuming variance homogeneity or normally distributed error terms. It is well known, however, that the Wald-type statistic requires large sample sizes to control the pre-assigned type-I error, see e.g., Vallejo, Fernández, and Livacic-Rojas (2010). Its small sample behavior may be improved by applying an adequate permutation procedure, see Pauly, Brunner, and Konietzschke (2015) for the theoretical background. The only comparable test included in the R function `oneway.test` is the Welch (1951) test for heteroscedastic one-way layouts. Furthermore, an ANOVA-type test based on ranks is also implemented in the R package `asbio` (Aho 2017) within the functions `BDM` and `BDM.2way` for nonparametric one- and two-way layouts, respectively.

For a user friendly application of these rather robust methods in statistical data sciences, the R package **GFD** has been developed. The use of the main function `GFD` as well as its output are very similar to the `aov` function from the R package `stats` or the `Anova` function from the R package `car` (Fox and Weisberg 2011). Its application provides a descriptive overview of the data as well as the complete ANOVA-tables according to the `formula` input, which allows the modeling of arbitrary high-way layouts. Hereby the Wald-type statistic, a permuted version thereof as well as the ANOVA-type statistic for these general factorial designs are implemented. Both the Wald-type statistic as well as the permutation test neither assume normality nor homogeneous variances, while the ANOVA-type statistic assumes normality. Furthermore, all main and interaction effects can be plotted along with $(1 - \alpha)$ confidence intervals. In addition, the package is equipped with a graphical user interface (GUI) to facilitate application for a wide audience of statisticians, practitioners, and educational purposes. The package is freely available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=GFD>.

The paper is organized as follows: In Section 2 we describe the statistical model and the tests used in this setting. In Section 3 we provide various examples for different settings which are statistically evaluated with the R package **GFD**. Finally, we discuss the results in Section 4 and provide an outlook to future work.

Throughout the paper we use the following notation: We denote by $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$ the a -dimensional centering matrix, \mathbf{I}_a is the a -dimensional unit matrix and \mathbf{J}_a denotes the $a \times a$ matrix of 1's, i.e., $\mathbf{J}_a = \mathbf{1}_a\mathbf{1}_a^\top$, where $\mathbf{1}_a = (1, \dots, 1)^\top$ is the a -dimensional column vector of 1's.

2. Statistical model and inference methods

In order to cover different factorial designs, we consider the following general linear model

$$Y_{ik} = \mu_i + \varepsilon_{ik}, \quad (1)$$

where $k = 1, \dots, n_i$ is the experimental unit within class $i = 1, \dots, a$. Note that different

sample sizes n_i are admitted. For each fixed i the error terms ε_{ik} are independent and identically distributed with $E(\varepsilon_{i1}) = 0$ and $\text{VAR}(\varepsilon_{i1}) = \sigma_i^2 > 0$. Note that we neither assume normality of the error terms nor variance homoscedasticity. In this setting, a higher way factorial structure with crossed or nested factors can be achieved by splitting up the index i into sub-indices i_1, i_2, \dots, i_p . In our notation, the components $i = 1, \dots, a$ can be considered as a lexicographic order of the factor level combinations.

In this framework we like to test general linear null hypotheses

$$H_0^\mu : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$$

about the mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_a)^\top$. Here \mathbf{H} denotes an adequate hypothesis contrast matrix of interest.

Let $\bar{\mathbf{Y}} = (\bar{Y}_1, \dots, \bar{Y}_a)^\top$ denote the vector of group means and let $\mathbf{V}_N = \text{COV}(\sqrt{N}\bar{\mathbf{Y}}) = \text{diag}(\frac{N}{n_i}\sigma_i^2 : i = 1, \dots, a)$ denote the covariance matrix of $\sqrt{N}\bar{\mathbf{Y}}$. Then \mathbf{V}_N is consistently estimated by $\widehat{\mathbf{V}}_N = \text{diag}(\frac{N}{n_i}\widehat{\sigma}_i^2)$, where $\widehat{\sigma}_i^2 = \frac{1}{n_i-1} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_i)^2$ denotes the empirical variance of the sample $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$.

In order to test the null hypotheses formulated above in this general framework, we consider two generalizations of the two-sample Welch t statistic: The Wald-type statistic (WTS) as discussed, e.g., in Pauly *et al.* (2015), and the ANOVA-type statistic (ATS) from Brunner *et al.* (1997). The WTS is given by

$$Q_N = N \bar{\mathbf{Y}}^\top \mathbf{H}^\top (\mathbf{H} \widehat{\mathbf{V}}_N \mathbf{H}^\top)^+ \mathbf{H} \bar{\mathbf{Y}}. \quad (2)$$

Here, \mathbf{M}^+ denotes the Moore-Penrose inverse of a matrix \mathbf{M} . It is well known that under rather weak assumptions the WTS has asymptotically a central χ_f^2 distribution with $f = \text{rank}(\mathbf{H})$ degrees of freedom under $H_0^\mu : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$. However, the WTS requires large sample sizes to get a satisfactory approximation by using the quantiles of the limiting χ^2 distribution (Akritas, Arnold, and Brunner 1997; Akritas and Brunner 1997; Vallejo *et al.* 2010; Pauly *et al.* 2015).

A second generalization of the two-sample Welch statistic is the ANOVA-type statistic (ATS) defined as

$$A_N = \frac{N}{\text{tr}(\mathbf{T} \widehat{\mathbf{V}}_N)} \bar{\mathbf{Y}}^\top \mathbf{T} \bar{\mathbf{Y}},$$

where $\mathbf{T} = \mathbf{H}^\top (\mathbf{H} \mathbf{H}^\top)^- \mathbf{H}$. Following Brunner *et al.* (1997) the distribution of the ATS can be approximated by an $F(\hat{f}, \hat{f}_0)$ -distribution such that the first two moments coincide, i.e., by choosing

$$\hat{f} = \text{tr}(\mathbf{T} \widehat{\mathbf{V}}_N)^2 / \text{tr}(\mathbf{T} \widehat{\mathbf{V}}_N \mathbf{T} \widehat{\mathbf{V}}_N)$$

and

$$\hat{f}_0 = \text{tr}(\mathbf{T} \widehat{\mathbf{V}}_N)^2 / \text{tr}(\mathbf{D}^2 \widehat{\mathbf{V}}_N^2 \boldsymbol{\Lambda}).$$

Here \mathbf{D} denotes the matrix of diagonal elements of \mathbf{T} and $\boldsymbol{\Lambda} = \text{diag}((n_1 - 1)^{-1}, \dots, (n_a - 1)^{-1})$ (Brunner *et al.* 1997; Brunner and Puri 2001). Note that in the two-sample case this approximation coincides with the Satterthwaite- t -approximation. However, the ATS is in general asymptotically exact only for normally distributed error terms.

Another possibility is to improve the small sample behavior of the WTS by applying a permutation procedure (Pauly *et al.* 2015). To describe this procedure in detail, let $\mathbf{Y}^\pi =$

$\pi(\mathbf{Y}_1, \dots, \mathbf{Y}_a)^\top$ denote a fixed but arbitrary permutation of \mathbf{Y} , i.e., $\pi \in \mathcal{S}_N$. Furthermore, let $\bar{\mathbf{Y}}^\pi = (\bar{Y}_1^\pi, \dots, \bar{Y}_a^\pi)^\top$ denote the vector of means and $\widehat{\mathbf{V}}_N^\pi = \text{diag}\left(\frac{N}{n_i}(\widehat{\sigma}_i^\pi)^2 : i = 1, \dots, a\right)$ the diagonal matrix of empirical variances $(\widehat{\sigma}_i^\pi)^2$ under this permutation. Then, the permuted Wald-type statistic (WTPS) is given by

$$Q_N^\pi = N(\bar{\mathbf{Y}}^\pi)^\top \mathbf{H}^\top (\mathbf{H} \widehat{\mathbf{V}}_N^\pi \mathbf{H}^\top)^+ \mathbf{H} \bar{\mathbf{Y}}^\pi,$$

which is the WTS as defined in Equation 2 calculated with the permuted observations. Now, a permutation test is achieved by the following steps:

1. Fix the data \mathbf{Y} and compute the WTS Q_N .
2. Permute the data randomly and obtain the value of Q_N^π . Save this in A_1 .
3. Repeat Step 2 J (say $J = 10,000$) times and obtain the values A_1, \dots, A_J .
4. Compute the p value by the (approximative) conditional permutation distribution (i.e., the empirical distribution of A_1, \dots, A_J) as

$$p \text{ value} = \frac{1}{J} \sum_{j=1}^J \mathcal{I}(Q_N \geq A_j).$$

Instead of computing the p value for making statistical inference, the original WTS Q_N can be compared with the $(1 - \alpha)$ quantile of the conditional distribution of Q_N^π given the data \mathbf{Y} , i.e., the empirical quantile of A_1, \dots, A_J . Pauly *et al.* (2015) have shown that this algorithm yields a valid permutation approach and consistent level α test, i.e., the conditional distribution of the WTPS always approximates the null distribution of Q_N . The test controls the preassigned level α under the null hypothesis and is even finitely exact if the pooled data is exchangeable under the hypothesis. Note that in the special case of a one-way layout the WTPS reduces to the permutation test for means of Chung and Romano (2013). The default value for the number of permutation runs in the R package **GFD** is $nperm = J = 10,000$.

For practical recommendations we briefly summarize the main properties of the three considered tests from Pauly *et al.* (2015): Mathematically, only the WTS and WTPS provide valid asymptotic procedures for general factorial designs. Nevertheless, simulation studies demonstrate that the ATS controls the α level for finite samples rather satisfactory. In case of non-normal data, however, the test tends to be conservative, which leads to loss of power. The WTS, in contrast, is quite liberal for small to moderate sample sizes. The WTPS is a rather accurate procedure even for non-normal data. When data is very skewed and heteroscedastic, the test tends to be liberal and to over-reject the hypothesis, in particular when the larger sample has the smaller variance (so called negative pairing). Its liberality is, however, not as pronounced as for the WTS.

Note that in comparison the **coin** package (Hothorn, Hornik, van de Wiel, and Zeileis 2008), which contains permutation tests for two- and multiple-sample problems, does not, e.g., handle heteroscedastic shift models. In our more general situation we allow for different variances and/or different distributions among the different groups. Furthermore, the Welch test from the function `oneway.test` is also only an approximation for normally distributed models that is known to perform worse than the ATS and the WTPS, see e.g., Vallejo *et al.* (2010) and

Pauly *et al.* (2015). Remark further, that the ANOVA-type tests from the R package **asbio** (Aho 2017) are based on ranks and test different null hypotheses formulated in terms of distribution functions instead of means.

For the calculation of the confidence intervals, we have used the corresponding quantiles of the t distribution.

2.1. Two-sample tests

A special case of model (1) is the heteroscedastic two-sample case, i.e., $a = 2$. This results in the extended Behrens-Fisher model

$$Y_{ik} = \mu_i + \varepsilon_{ik}, \quad i = 1, 2; \quad k = 1, \dots, n_i,$$

which is usually analyzed using a Welch's t test in the statistic

$$T_N = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}}. \quad (3)$$

Its distribution is approximated by a t_ν distribution with estimated Satterthwaite-Welch degree of freedom ν to account for variance heterogeneity. Another possibility to approximate the distribution of T_N as defined in Equation 3 is to employ the studentized permutation distribution of T_N , and to carry out the test as a permutation test as proposed by Janssen (1997, 2005).

Note that the Wald-type statistic Q_N , as well as the ATS A_N are the square of T_N in the two-sample case. Furthermore, both the statistics Q_N and A_N are identical in this setup; and the second degree of freedom \hat{f}_0 of the ATS is identical to the Satterthwaite-Welch degree of freedom. The first degree of freedom \hat{f} is equal to 1, by definition. Thus the ATS test is essentially Welch's t test and the WTPS test is in fact Janssen's permutation test.

3. Examples

In this section, we provide examples demonstrating how different factorial designs can be analyzed using the **GFD** package. The function **GFD** returns an object of class 'GFD' from which the user may obtain plots and summaries of the results using **plot()**, **print()** and **summary()** methods, respectively. Here, **print()** returns a short summary of the results, i.e., the values of the test statistics along with degrees of freedom and corresponding p values whereas **summary()** also displays some descriptive statistics such as the means and variances for the different factor level combinations. Plotting is based on **plotrix** (Lemon 2006). For two- and higher-way layouts, the factors for plotting can be additionally specified in the **plot** call, see the examples below.

```
GFD(formula, data = NULL, nperm = 10000, alpha = 0.05)
```

Note that the test statistics for the main effects considered in Section 2 are not changed by whether or not an additional interaction term is specified in **formula** since the tests are determined by the choice of the hypothesis matrix **H**. Only crossed and hierarchical (nested) designs are implemented – a mixture of both is up to date not available.

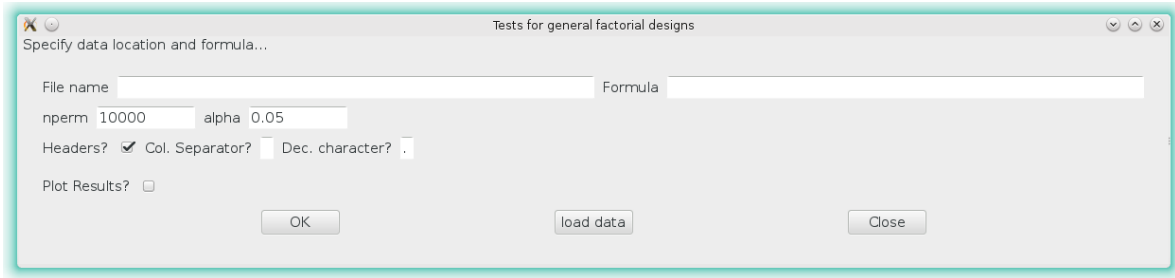


Figure 1: The GUI for tests in general factorial designs: The user can specify the data location, the formula, the number of permutations and the significance level α . One can additionally choose to plot the results.

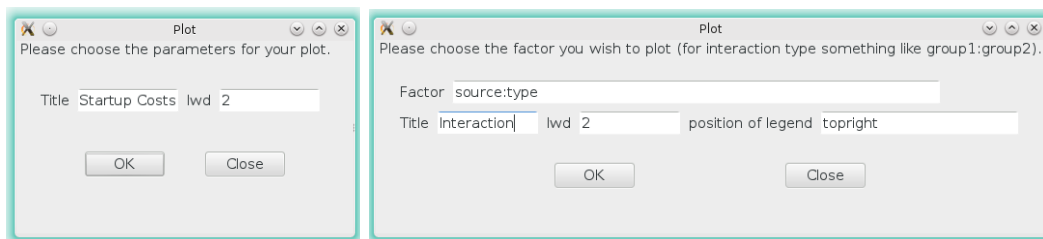


Figure 2: Graphical user interfaces for plotting: The left GUI is for the one-way layout (no choice of factors possible), the other one is for a higher-way layout. An example for plotting interactions is given in the right panel.

Furthermore, the **GFD** package is equipped with an optional GUI, based on **RGtk2** (Lawrence and Temple Lang 2010), which will be explained in detail in the next section.

3.1. Graphical user interface

The GUI is started in R with the command `calculateGUI()`. Note that the GUI depends on **RGtk2** and will only work if **RGtk2** is installed. The user can specify the data location (either directly or via the "load data" button), the formula, the number of permutations and the significance level α , see Figure 1. Additionally, one can specify whether or not headers are included in the data file, and which separator and character symbols are used for decimals in the data file. The GUI also provides a plotting option, which generates a new window for specifying the factors to be plotted (in higher-way layouts) along with a few plotting parameters, see Figure 2. Note that four- and higher way interactions cannot be plotted due to the increasing complexity of the plots.

```
R> library("GFD")
R> calculateGUI()
```

3.2. Two-sample tests

As an example of a two-sample problem we consider a subset of the `weightgain` data set (Hand, Daly, McConway, Lunn, and Ostrowski 1993) from the **HSAUR** package (Everitt and Hothorn 2017). The data contains information on the weight gain (in grams) of rats which

were randomized to one of four diets, distinguished by the amount of protein (high and low) and the source of protein (beef and cereal). For our purposes, we first restrict our analysis to the high protein group.

```
R> library("GFD")
R> data("weightgain", package = "HSAUR")
R> weightgain2 <- subset(weightgain, type == "High")
R> set.seed(123)
R> two_sample <- GFD(weightgain ~ source, data = weightgain2,
+   nperm = 10000, alpha = 0.05)
R> plot(two_sample, main = "Two-sample test", cex.axis = 1.5,
+   cex.lab = 1.5, cex.main = 1.5, lwd = 2)
R> two_sample
```

Call:

```
weightgain ~ source
```

Wald-Type Statistic (WTS):

Test statistic	df	p-value	p-value WTPS
4.37169244	1.00000000	0.03654068	0.05580000

ANOVA-Type Statistic (ATS):

Test statistic	df1	df2	p-value
4.37169244	1.00000000	17.99896078	0.05099558

Note that the results are identical with those using the `t.test` function:

```
R> t.test(weightgain ~ source, data = weightgain2)
```

Welch Two Sample t-test

```
data: weightgain by source
t = 2.0909, df = 17.999, p-value = 0.051
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.0679184 28.2679184
sample estimates:
 mean in group Beef mean in group Cereal
           100.0           85.9
```

As mentioned in Section 2.1 the p values obtained using the ATS and the Satterthwaite-Welch t test are identical. A reason for the smaller p value obtained with the WTS may be given due to its more liberal behavior in case of small sample sizes ($n_1 = n_2 = 10$), see Vallejo *et al.* (2010) and Pauly *et al.* (2015).

The data may also be analyzed using the GUI, see Figure 3 for an example. The corresponding plot of the effect is given in Figure 4.



Figure 3: Graphical user interface with formula for the `weightgain` data set.

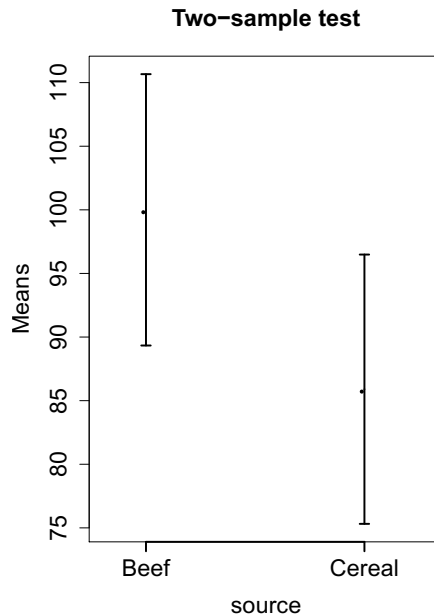


Figure 4: Mean weight gain for the two different sources of protein, beef and cereal, in the two-sample problem.

3.3. One-way layout

In a one-way layout,

$$Y_{ik} = \mu_i + \varepsilon_{ik}, \quad i = 1, \dots, a; \quad k = 1, \dots, n_i,$$

we are interested in the effect of factor A , i.e., we wish to test the null hypothesis $H_0 : \{\mu_1 = \dots = \mu_a\} = \{\mathbf{P}_a \boldsymbol{\mu} = \mathbf{0}\}$.

An example for such a model is the data set on startup costs of companies, which was selected from the Business Opportunities Handbook, see [Cengage College \(2008\)](#). The data represent business startup costs in thousands of dollars for five different kinds of shops.

```
R> library("GFD")
R> data("startup", package = "GFD")
R> set.seed(456)
R> model1 <- GFD(Costs ~ company, data = startup, nperm = 10000,
+   alpha = 0.05)
```

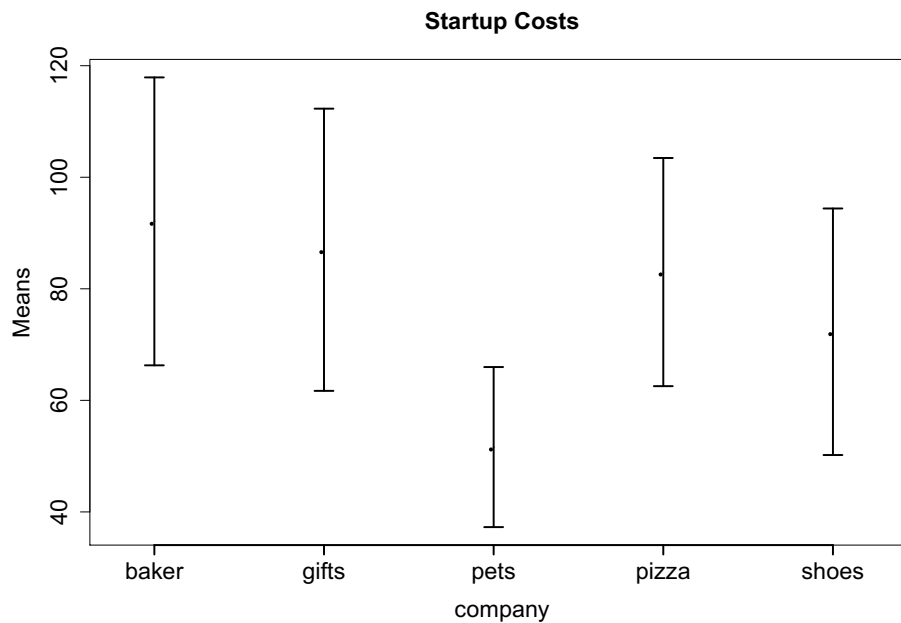


Figure 5: Mean startup costs for the five different companies in the `startup` data example.

```
R> summary(model1)
R> plot(model1, main = "Startup Costs", cex.axis = 1.5, cex.lab = 1.5,
+       cex.main = 1.5, lwd = 2)
```

```
Call:
Costs ~ company
```

Descriptive:

company	n	Means	Variances	Lower 95 % CI	Upper 95 % CI
1 baker	11	92.09091	1512.6909	66.28044	117.90138
2 gifts	10	87.00000	1289.1111	61.70193	112.29807
3 pets	16	51.62500	733.0500	37.27595	65.97405
4 pizza	13	83.00000	1165.1667	62.54732	103.45268
5 shoes	10	72.30000	983.7889	50.19995	94.40005

Wald-Type Statistic (WTS):

Test statistic	df	p-value	p-value WTPS
15.037830399	4.000000000	0.004623394	0.024600000

ANOVA-Type Statistic (ATS):

Test statistic	df1	df2	p-value
2.57248203	3.70623134	44.51042721	0.05456579

This example nicely demonstrates the liberal behavior of the WTS (p value = 0.0046) as well as the conservative behavior of the ATS (p value = 0.055). The WTPS, in contrast, is somewhere in between with a p value of 0.0246.

3.4. Two-way layout

In a two-way crossed design,

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

with $i = 1, \dots, a$; $j = 1, \dots, b$; $k = 1, \dots, n_{ij}$, one is interested in tests for the main effects of the factors A and B as well as for an interaction of the two, i.e.,

$$\begin{aligned} H_0(A) &: \{\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..} = 0 \ \forall i = 1, \dots, a\}, \\ H_0(B) &: \{\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..} = 0 \ \forall j = 1, \dots, b\}, \\ H_0(AB) &: \{\gamma_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} = 0 \ \forall i = 1, \dots, a, j = 1, \dots, b\}, \end{aligned}$$

or formulated with suitable contrast matrices:

$$\begin{aligned} H_0(A) &: \{\mathbf{H}_A \boldsymbol{\mu} = \mathbf{P}_a \otimes \frac{1}{b} \mathbf{1}_b^\top \cdot \boldsymbol{\mu} = \mathbf{0}\}, \\ H_0(B) &: \{\mathbf{H}_B \boldsymbol{\mu} = \frac{1}{a} \mathbf{1}_a^\top \otimes \mathbf{P}_b \cdot \boldsymbol{\mu} = \mathbf{0}\}, \\ H_0(AB) &: \{\mathbf{H}_{AB} \boldsymbol{\mu} = \mathbf{P}_a \otimes \mathbf{P}_b \cdot \boldsymbol{\mu} = \mathbf{0}\}. \end{aligned}$$

We will again consider the `weightgain` data set from package **HSAUR**. This time, however, we are interested in analyzing both factors, i.e., amount and source of protein.

```
R> library("GFD")
R> data("weightgain", package = "HSAUR")
R> set.seed(789)
R> model2 <- GFD(weightgain ~ source * type, data = weightgain)
R> summary(model2)
R> plot(model2, factor = "source:type", main = "Interaction", xlab = "Type",
+       cex.axis = 1.5, cex.lab = 1.5, cex.main = 1.5)
R> plot(model2, factor = "source", main = "Mean weight gain",
+       xlab = "source", cex.axis = 1.5, cex.lab = 1.5, cex.main = 1.5)
```

Call:

```
weightgain ~ source * type
```

Descriptive:

	source	type	n	Means	Variances	Lower 95 % CI	Upper 95 % CI
1	Beef	High	10	100.0	229.1111	89.33489	110.66511
3	Beef	Low	10	79.2	192.8444	69.41534	88.98466
2	Cereal	High	10	85.9	225.6556	75.31562	96.48438
4	Cereal	Low	10	83.9	246.7667	72.83158	94.96842

Wald-Type Statistic (WTS):

	Test statistic	df	p-value	p-value WTPS
source	0.9879494	1	0.32024407	0.3229
type	5.8123090	1	0.01591439	0.0204
source:type	3.9517976	1	0.04682133	0.0554

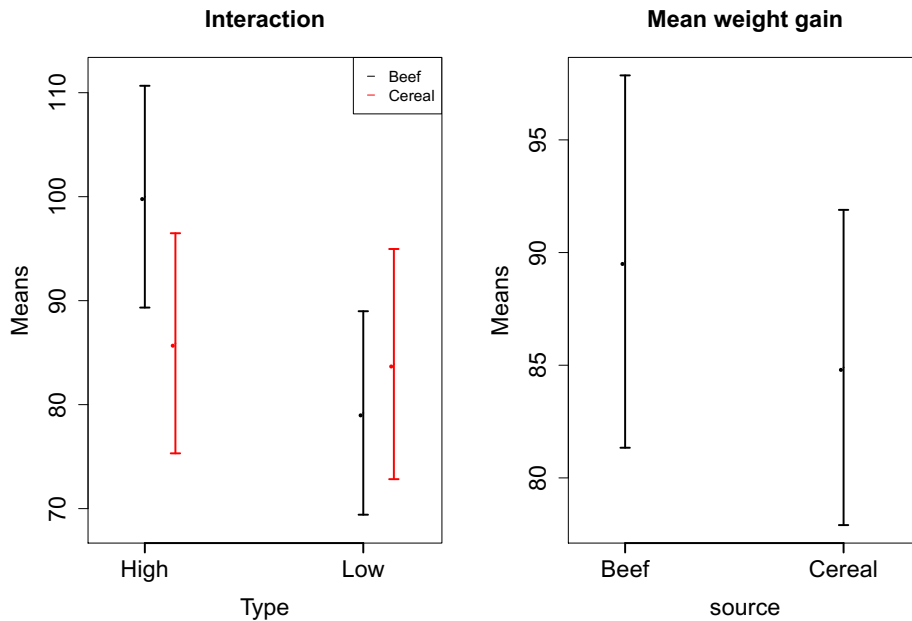


Figure 6: Plots of the interaction of factors `source` and `type` in the weight gain data (left) and for factor `source` alone (right).

ANOVA-Type Statistic (ATS):

	Test statistic	df1	df2	p-value
<code>source</code>	0.9879494	1	35.72893	0.32692829
<code>type</code>	5.8123090	1	35.72893	0.02118641
<code>source:type</code>	3.9517976	1	35.72893	0.05452616

The factor `type`, i.e., high or low amount of protein in the food, has a significant impact on the weight gain at 5% level of significance using all three different tests. The source of the protein, in contrast, does not have a significant influence. The interesting part is the test for interaction: Here, the classical WTS results in a p value of 0.047, whereas both the ATS and WTPS provide a p value of 0.055. Thus, both the ATS and WTPS endorse a “borderline significance” at 5% level.

Figure 6 shows plots for the main effect of the factor `type` as well as the interaction between both factors.

3.5. Three-way layout

For the three-way example, we consider a data set on pizza delivery times (Mackisack 1994). The objective of the study was to see how the delivery time in minutes would be affected by three different factors: whether thick or thin crust was ordered (factor A), whether Coke was ordered with the pizza or not (factor B), and whether or not garlic bread was ordered as a side (factor C). The R code to analyze this data is given in the following statements:

```
R> library("GFD")
R> data("pizza", package = "GFD")
```

```
R> set.seed(1234)
R> model3 <- GFD(Delivery ~ Crust * Coke * Bread, data = pizza)
R> summary(model3)
R> plot(model3, factor = "Crust:Coke:Bread", legendpos = "center",
+       main = "Delivery time of pizza", xlab = "Bread", cex.axis = 1.5,
+       cex.lab = 1.5, cex.main = 1.5, lwd = 2)
R> plot(model3, factor = "Crust:Coke", legendpos = "topleft",
+       main = "Two-way interaction", xlab = "Coke", cex.axis = 1.5,
+       cex.lab = 1.5, cex.main = 1.5, lwd = 2)
```

Call:

```
Delivery ~ Crust * Coke * Bread
```

Descriptive:

	Crust	Coke	Bread	n	Means	Variances	Lower 95 % CI	Upper 95 % CI
1	thin	no	no	2	19.0	2.0	14.69735	23.30265
5	thin	no	yes	2	17.5	0.5	15.34867	19.65133
3	thin	yes	no	2	17.5	4.5	11.04602	23.95398
7	thin	yes	yes	2	15.0	2.0	10.69735	19.30265
2	thick	no	no	2	19.5	0.5	17.34867	21.65133
6	thick	no	yes	2	18.0	2.0	13.69735	22.30265
4	thick	yes	no	2	21.5	0.5	19.34867	23.65133
8	thick	yes	yes	2	18.5	0.5	16.34867	20.65133

Wald-Type Statistic (WTS):

	Test statistic	df	p-value	p-value	WTPS
Crust	11.56	1	0.0006738585		0.0089
Coke	0.36	1	0.5485062355		0.5613
Crust:Coke	6.76	1	0.0093223760		0.0286
Bread	11.56	1	0.0006738585		0.0073
Crust:Bread	0.04	1	0.8414805811		0.8153
Coke:Bread	1.00	1	0.3173105079		0.3457
Crust:Coke:Bread	0.04	1	0.8414805811		0.8212

ANOVA-Type Statistic (ATS):

	Test statistic	df1	df2	p-value
Crust	11.56	1	4.699248	0.02121110
Coke	0.36	1	4.699248	0.57625702
Crust:Coke	6.76	1	4.699248	0.05122842
Bread	11.56	1	4.699248	0.02121110
Crust:Bread	0.04	1	4.699248	0.84984482
Coke:Bread	1.00	1	4.699248	0.36598284
Crust:Coke:Bread	0.04	1	4.699248	0.84984482

We find a significant influence of the factors `Crust` and `Bread`. The WTS and WTPS also suggest a significant interaction between the factors `Crust` and `Coke` at 5% level, which is only

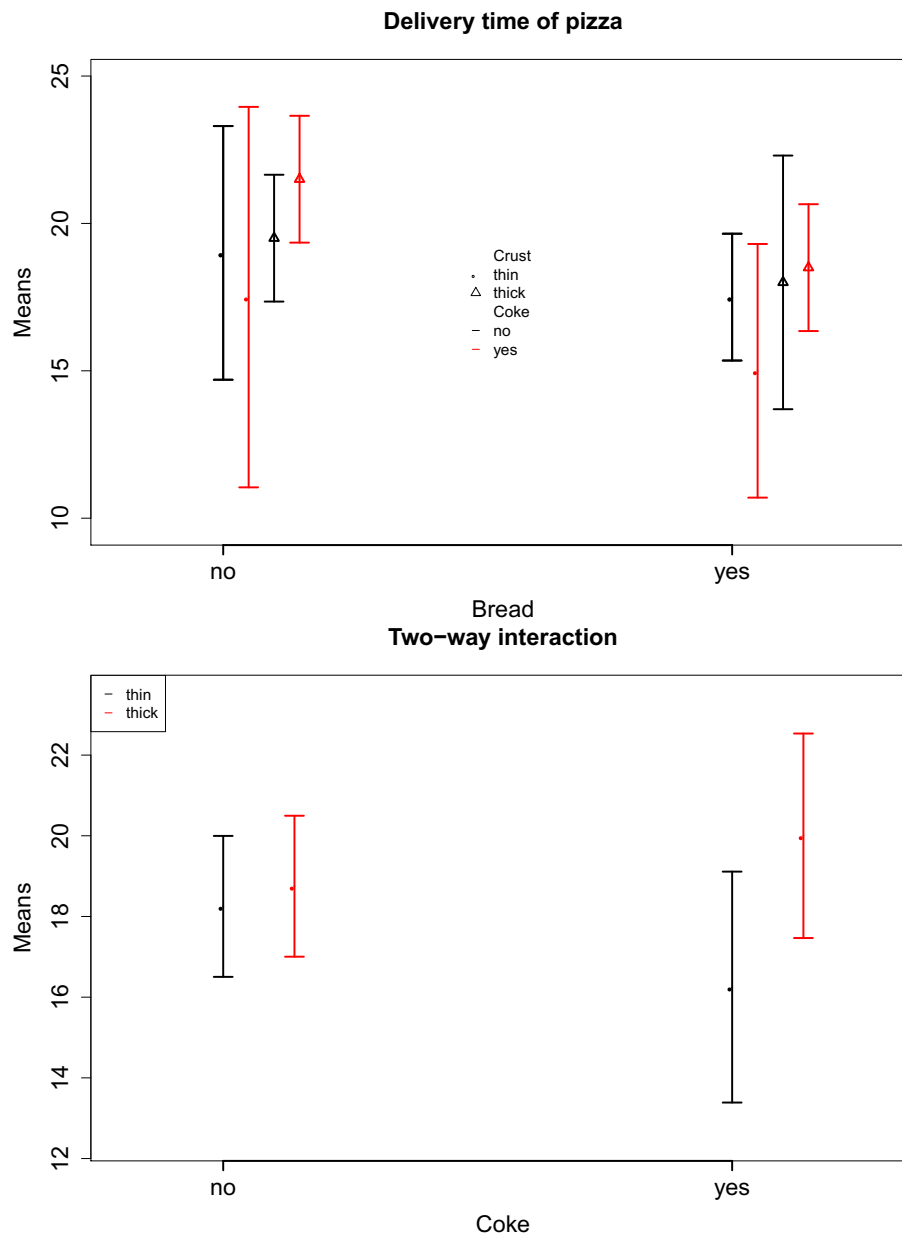


Figure 7: Plots of the three-way interaction (upper panel) and the two-way interaction between factors Coke and Crust (lower panel).

borderline significant when using the ATS. Figure 7 shows interaction plots of the three-way interaction as well as the two-way interaction between Crust and Coke.

3.6. Nested design

A nested design is covered by the model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},$$

where factor B is nested within the levels of factor A . As an example, we consider the curdies

data set (Quinn, Lake, and Schreiber 1996) included in the **GFD** package. The aim of the study was to describe basic patterns of variation in a small flatworm, *Dugesia*, in the Curdies River, Western Victoria. Therefore, worms were sampled at two different seasons and three different sites within each season. For our analyses we consider both factors as fixed (e.g., some sites may only be accessed in summer). The R code for analyzing this nested design is given in the following:

```
R> library("GFD")
R> data("curdies", package = "GFD")
R> set.seed(987)
R> nested <- GFD(dugesia ~ season + season:site, data = curdies)
R> summary(nested)
R> plot(nested, factor="season:site", xlab = "site", cex.axis = 1.5,
+       cex.lab = 1.5, cex.main = 1.5, lwd = 2)
```

Call:

```
dugesia ~ season + season:site
```

Descriptive:

	season	site	n	Means	Variances	Lower 95 % CI	Upper 95 % CI
1	SUMMER	4	6	0.4190947	0.4615290	-0.25954958	1.0977390
2	SUMMER	5	6	0.2290862	0.3148830	-0.33146759	0.7896401
3	SUMMER	6	6	0.1942443	0.0729142	-0.07549781	0.4639864
4	WINTER	1	6	2.0494375	4.0647606	0.03543415	4.0634408
5	WINTER	2	6	4.1819078	35.6801853	-1.78509515	10.1489107
6	WINTER	3	6	0.6782063	0.1910970	0.24151987	1.1148927

Wald-Type Statistic (WTS):

	Test statistic	df	p-value	p-value	WTPS
season	5.415180	1	0.01996239		0.0001
season:site	5.200991	4	0.26728919		0.3154

ANOVA-Type Statistic (ATS):

	Test statistic	df1	df2	p-value
season	5.415180	1.000000	6.447707	0.05593278
season:site	1.382224	1.217424	6.447707	0.29278958

In this setting, both WTS and WTPS detect a significant influence of the season whereas the ATS, again, only shows a borderline significance at 5% level. The effect of the site is not significant. A plot for the nested effect is given in Figure 8.

4. Conclusion and future work

The R package **GFD** implements a broad range of semi-parametric methods for the analysis of general factorial designs, i.e., linear models without the assumption of normality and/or homoscedastic variances across the treatment groups. Three different methods are implemented:

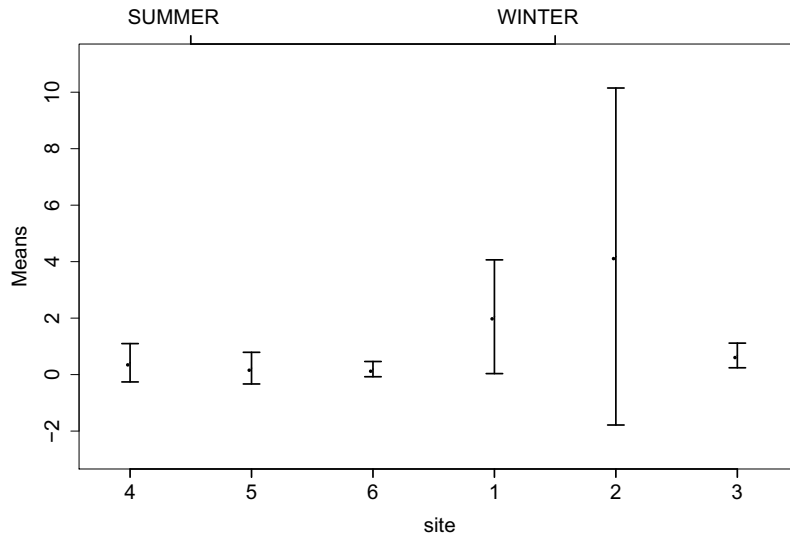


Figure 8: Plot for the effects in the nested design. The sites are nested within seasons.

Wald-type statistic Q_N , ANOVA-type statistic A_N as well as a permutation approach proposed by Pauly *et al.* (2015). All methods can be used to test general hypotheses among the main and interaction effects. In particular, nested designs can be analyzed using **GFD**. From a practical point of view we recommend the WTPS procedure since it has been found in Pauly *et al.* (2015) to possess both good finite type-I error rate control and power behavior. The ATS and WTS, in comparison, are slightly conservative or rather liberal, respectively. Confidence interval plots are available for all effects of interest – except of four- and higher-way interactions.

A graphical user interface (GUI) has been implemented which allows a convenient use of the software in industry, academia, and educational purposes. We plan to update the **GFD** package on a regular basis with new procedures available for the analysis of general designs. So far, ANOVA-based methods are implemented, and an adjustment of the treatment effects for covariates is not possible. Furthermore, tests and simultaneous confidence intervals for multiple comparisons based on the permutation approach are not yet available. The extension of the implemented methods to covariates and multiple comparisons and their implementation will be part of future research.

Acknowledgments

The work of Sarah Friedrich and Markus Pauly was supported by the German Research Foundation project DFG-PA 2409/3-1.

References

Aho K (2017). *asbio: A Collection of Statistical Tools for Biologists*. R package version 1.4-2, URL <https://CRAN.R-project.org/package=asbio>.

- Akritis MG, Arnold SF, Brunner E (1997). “Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs.” *Journal of the American Statistical Association*, **92**(437), 258–265. doi:10.2307/2291470.
- Akritis MG, Brunner E (1997). “A Unified Approach to Rank Tests for Mixed Models.” *Journal of Statistical Planning and Inference*, **61**(2), 249–277. doi:10.1016/s0378-3758(96)00177-2.
- Bates D, Mächler M (2017). **Matrix: Sparse and Dense Matrix Classes and Methods**. R package version 1.2-10, URL <https://CRAN.R-project.org/package=Matrix>.
- Bathke AC, Schabenberger O, Tobias RD, Madden LV (2009). “Greenhouse-Geisser Adjustment and the ANOVA-Type Statistic: Cousins or Twins?” *The American Statistician*, **63**(3), 239–246. doi:10.1198/tast.2009.08187.
- Brunner E, Dette H, Munk A (1997). “Box-Type Approximations in Nonparametric Factorial Designs.” *Journal of the American Statistical Association*, **92**(440), 1494–1502. doi:10.1080/01621459.1997.10473671.
- Brunner E, Puri ML (2001). “Nonparametric Methods in Factorial Designs.” *Statistical Papers*, **42**(1), 1–52. doi:10.1007/s003620000039.
- Cengage College (2008). http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/owan/frames/frame.html. [Accessed 28-04-2016].
- Chung E, Romano JP (2013). “Exact and Asymptotically Robust Permutation Tests.” *The Annals of Statistics*, **41**(2), 484–507. doi:10.1214/13-aos1090.
- Everitt BS, Hothorn T (2017). **HSAUR: A Handbook of Statistical Analyses Using R (1st Edition)**. R package version 1.3-8, URL <https://CRAN.R-project.org/package=HSAUR>.
- Fox J, Weisberg S (2011). *An R Companion to Applied Regression*. 2nd edition. Sage, Thousand Oaks.
- Hand DJ, Daly F, McConway K, Lunn D, Ostrowski E (1993). *A Handbook of Small Data Sets*. CRC Press.
- Hankin RKS (2005). “Recreational Mathematics with R: Introducing The **magic** Package.” *R News*, **5**(1), 48–51.
- Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). “Implementing a Class of Permutation Tests: The **coin** Package.” *Journal of Statistical Software*, **28**(8), 1–23. doi:10.18637/jss.v028.i08.
- Janssen A (1997). “Studentized Permutation Tests for Non-lid Hypotheses and the Generalized Behrens-Fisher Problem.” *Statistics & Probability Letters*, **36**(1), 9–21. doi:10.1016/s0167-7152(97)00043-6.
- Janssen A (2005). “Resampling Student’s *t*-Type Statistics.” *The Annals of the Institute of Statistical Mathematics*, **57**(3), 507–529. doi:10.1007/bf02509237.

- Johansen S (1980). “The Welch-James Approximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression.” *Biometrika*, **67**(1), 85–92. doi:10.2307/2335320.
- Lawrence M, Temple Lang D (2010). “**RGtk2**: A Graphical User Interface Toolkit for R.” *Journal of Statistical Software*, **37**(8), 1–52. doi:10.18637/jss.v037.i08.
- Lemon J (2006). “**plotrix**: A Package in the Red Light District of R.” *R News*, **6**(4), 8–12.
- Mackisack M (1994). “What Is the Use of Experiments Conducted by Statistics Students?” *Journal of Statistics Education*, **2**(1), 1–15. URL <https://ww2.amstat.org/publications/jse/v2n1/mackisack.html>.
- Pauly M, Brunner E, Konietzschke F (2015). “Asymptotic Permutation Tests in General Factorial Designs.” *Journal of the Royal Statistical Society B*, **77**(2), 461–473. doi:10.1111/rssb.12073.
- Placzek M, Konietzschke F, Pauly M (2014). “Studentisierte Permutationstests für verbundene und unverbundene 2-Stichprobenprobleme.” In *KSFE 2014 – Konferenz der SAS-Anwender in Forschung und Entwicklung*.
- Quinn GP, Lake PS, Schreiber ESG (1996). “Littoral Benthos of a Victorian Lake and Its Outlet Stream: Spatial and Temporal Variation.” *Australian Journal of Ecology*, **21**(3), 292–301. <http://users.monash.edu.au/~murray/AIMS-R-users/downloads/data/curdies.csv>, [Accessed 28-04-2016].
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien, Austria. URL <https://www.R-project.org/>.
- Vallejo G, Fernández MP, Livacic-Rojas PE (2010). “Analysis of Unbalanced Factorial Designs with Heteroscedastic Data.” *Journal of Statistical Computation and Simulation*, **80**(1), 75–88. doi:10.1080/00949650802482386.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. doi:10.1007/978-0-387-21706-2.
- Welch BL (1951). “On the Comparison of Several Mean Values: An Alternative Approach.” *Biometrika*, **38**(3–4), 330–336.
- Wickham H (2011). “The Split-Apply-Combine Strategy for Data Analysis.” *Journal of Statistical Software*, **40**(1), 1–29. doi:10.18637/jss.v040.i01.
- Zhang JT (2012). “An Approximate Degrees of Freedom Test for Heteroscedastic Two-Way ANOVA.” *Journal of Statistical Planning and Inference*, **142**(1), 336–346. doi:10.1016/j.jspi.2011.07.023.

Affiliation:

Sarah Friedrich, Markus Pauly

Institute of Statistics

Ulm University

89081 Ulm, Germany

E-mail: sarah.friedrich@uni-ulm.de, markus.pauly@uni-ulm.de

Frank Konietschke

Department of Statistics

University of Texas at Dallas

Dallas, TX 75080, United States of America

Email: fxk141230@utdallas.edu

Sarah Friedrich

Persönliche Angaben

* **20. Juli 1989**, *Filderstadt*.

Ausbildung

Seit Feb 2015 **Promotionsstudentin**, *Institut für Statistik*, Universität Ulm.

- Sep 2015: Dreiwöchiger Forschungsaufenthalt an der University of Texas, Dallas, USA.

Okt 2012 – Dez 2014 **M.Sc. Mathematische Biometrie**, *Universität Ulm*.

- Titel der Masterarbeit: 'Estimation of pregnancy outcome probabilities in the presence of heavy left-truncation'
- Boehringer-Ingelheim-Preis für den besten Masterabschluss im Studiengang Mathematische Biometrie
- Young-Statisticians Award 2015 der Deutschen Region der Internationalen Biometrischen Gesellschaft (IBS-DR)

Okt 2009 – Sep 2012 **B.Sc. Mathematik**, *Universität Stuttgart*.

- Nebenfach: Technische Biologie
- Titel der Bachelorarbeit: 'Datenspektroskopie: Eigenräume von Faltungsoperatoren und Clustering'

Sep 2000 – Jul 2008 **Leibniz-Gymnasium**, Stuttgart-Feuerbach, Experimentalzug G8, Abschluss: Abitur.

Sonstiges

Okt 2008 – Sep 2009 **Auslandsaufenthalt in Australien**.

Publikationen und Preprints

- Dobler, D, Friedrich, S und Pauly, M (2017). Nonparametric MANOVA in Mann-Whitney effects. (Eingereicht bei *Journal of the American Statistical Association*).
- Friedrich, S, Konietschke, F, und Pauly, M (2017). Analysis of Multivariate Data and Repeated Measures Designs with the R Package MANOVA.RM. (Eingereicht bei *Computational Statistics and Data Analysis*.)
- Friedrich, S und Pauly, M (2017). MATS: Inference for potentially Singular and Heteroscedastic MANOVA. *arXiv preprint arXiv:1704.03731*. (Revision eingereicht bei *Journal of Multivariate Analysis*).
- Friedrich, S, Brunner, E und Pauly, M (2017). Permuting Longitudinal Data In Spite Of The Dependencies. *Journal of Multivariate Analysis*, **153**, 255–265.
- Friedrich, S, Beyersmann, J, Winterfeld, U, Schumacher, M und Allignol, A (2017). Nonparametric Estimation of Pregnancy Outcome Probabilities. *Annals of Applied Statistics*, **11**(2), 840–867.
- Friedrich, S, Konietschke, F und Pauly, M (2017). A Wild Bootstrap Approach for Nonparametric Repeated Measurements. *Computational Statistics and Data Analysis*, **113**, 38–52.
- Friedrich, S, Konietschke, F und Pauly, M (2017). GFD: An R package for the Analysis of General Factorial Designs. *Journal of Statistical Software, Code Snippets*, **79**(1), 1–18.
- Bathke, A, Friedrich, S, Konietschke, F, Pauly, M, Staffen, W, Strobl, N und Höller, Y (2016). Using EEG, SPECT, and Multivariate Resampling Methods to Differentiate Alzheimer Patients from Others. *arXiv preprint arXiv:1606.09004*. (Revision eingereicht *Multivariate Behavioral Research*).

R-Pakete

GFD: Tests for General Factorial Designs.

rankFD: Rank-Based Tests for General Factorial Designs.

MANOVA.RM: Analysis of Multivariate Data and Repeated Measures Designs.

Vorträge

- Resampling Approaches for Repeated Measures Designs and Multivariate Data - Theory, R-package and Applications, CEN ISBS 2017, Wien (invited talk).
- MATS: Inference for potentially singular and general heteroscedastic MANOVA, EMS 2017, Helsinki.
- Permuting longitudinal data despite all the dependencies, YES Workshop 2017, Eindhoven (Poster).
- Permuting longitudinal data despite all the dependencies, ISCB 2016, Birmingham.
- Permuting longitudinal data despite all the dependencies, YSM 2016, London.
- GFD: An R-package for the Analysis of General Factorial Designs, Statistical Computing 2016, Günzburg.
- Permuting longitudinal data despite all the dependencies, ISNPS 2016, Avignon.
- Permuting longitudinal data despite all the dependencies, DAGStat 2016, Göttingen.
- Permuting longitudinal data despite all the dependencies, GPSD 2016, Bochum.
- Estimation of pregnancy outcome probabilities in the presence of heavy left-truncation, ISCB 2015, Utrecht.
- Estimation of pregnancy outcome probabilities in the presence of heavy left-truncation, Biometrisches Kolloquium 2015, Dortmund, Young Statisticians Session.
- Estimation of pregnancy outcome probabilities in the presence of heavy left-truncation, Workshop ‘Non-parametric analyses of complex time to event data’, Ulm.

Name: Sarah Jasmin Friedrich

Matrikelnummer: 791605

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe.

Ulm, Juli 2017

Sarah Jasmin Friedrich