Marléne Baumeister · Dissertation

technische universität
dortmund

fakultät
statistik

Empirical Process Theory for Robust Inference in Multivariate Analyses and Multiple Testing

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der Technischen Universität Dortmund

Marléne Baumeister

Dortmund, November 10, 2025

Present Dean of the Department
Prof. Dr. Philipp Doebler

Reviewers
Prof. Dr. Markus Pauly
Prof. Dr. Dennis Dobler

Date of Defense
September 23, 2025

Version November 10, 2025

This dissertation was typeset with LaTeX. KOMA-Script and the main font Latin Modern were used. For version control and project management git was used.

This work cites scientific works from the field of eugenics, because important statistical concepts were developed in this context. The author condemns the racist, ableist and in any way discriminatory aspects and conclusions contained therein.

Abstract

Motivation

This cumulative dissertation is about multivariate and multiple testing methods which are applicable for data beyond normality. In medicine, psychology, and biology, there is a need for testing procedures that are applicable in complex factorial designs and have robust properties at the same time. Most established approaches are less robust in the sense that they rely on strong assumptions such as normality and homogeneity, which cannot be justified in many quantitative research studies. Apart from single testing problems, which underlie ANOVA and MANOVA, multiple testing problems are equally relevant. The more complex the underlying factorial design, the more realistic it is that a multiple testing problem is of interest. Therefore, easy-interpretable and powerful solutions for single and multiple testing problems are of interest.

To have robust methods at hand, there is a recent trend to resampling-based alternatives to classical analysis-of-variance (ANOVA) and multivariate analysis-of-variance (MANOVA), which rely on fewer assumptions. While there already exist robust alternatives for well-known and even complexer mean-based factorial designs, the situation is different for estimands beyond the mean. Inference regarding quantiles, especially regarding the median, is a potentially robust and powerful alternative to mean-based inference, especially in context of skewed and heavy-tailed data. Another issue, where the use of resampling methods can be feasible, is the presence of covariates. For example in medicine and psychology, there is often is the desire or even a requirement to adjust for covariates. In this situation, the covariate-adjusted means can be an adequate estimand.

All research questions are motivated by real data examples from psychology or biology. They exemplify the high applicability of and the need for the respective methods.

Methods

This dissertation fills some gaps in the methodology for robust inference in multivariate analyses and multiple testing. The contributions are described as follows:

The first article introduces a MANOVA with quantiles as estimands for general factorial designs by implementing resampling tests based on quadratic form-type statistics. The asymptotic theory is based on empirical processes.

The second article includes a general comparison of one- and two-sided quantile-based simultaneous multiple testing procedures and extends some existing methods for quantile-based multiple testing to one-sided testing problems. An extensive Monte Carlo simulation was carried out, which provides a detailed view of the behaviour of the testing procedures under the null hypothesis and under the alternative.

In the third article, simultaneous multiple testing on covariate-adjusted means was made available through the implementation of multiple contrast test procedures in a semiparametric model for multivariate analysis-of-covariance (MANCOVA).

Results

In all three articles, it has been shown that the new methods are asymptotically valid and consistent resampling tests. To reach this conclusion, aspects of asymptotic statistics and empirical process theory have been applied. Apart from that, Monte Carlo simulations have been used to analyse the performance of the methods under the null hypothesis and under the alternative. This allows an insight into the behaviour of the new methods, especially on small sample sizes. For the first and the third article, the simulations allow an initial evaluation of the performance of the new methods. In the second article, the comparative simulation studies enable a deeper insight in which scenarios the behaviour of the methods is satisfactory and in which it is less good. In that context, some surprising results have been obtained.

Acknowledgements

At some point during my master's thesis I was irritated for the first time by the fact that there are so few suitable methods for doing inference in factorial designs. Markus Pauly and Marc Ditzhaus gave me the opportunity to work on the QMANOVA (which resulted in the first paper in this thesis) and I studied the theory of empirical processes for that. As a mathematician I had literally no idea about the aims and scopes of factorial designs, but I thought theory about this is an old hat. Then they gave me a paper from 2018 (Friedrich & Pauly, 2018) to read and I was wondering that essential theory and substantial improvement of MANOVA was developed less than five years ago. A little while before that happened, people (I think they were physicist) told me, that everything about mathematics has been found out, especially the aspects of mathematics that are useful for somebody or something in the real world.

And now I may call my work filling gaps in the statistical methodology about factorial designs. This includes talking to people who might want to use them. In doing so, I constantly notice that somehow a method does not exist yet. I am thankful to call this my profession and I am glad to have found something, where it is beneficial how I see the world, as a wild bunch of structure that has to be understood.

So many people have accompanied me on the way to this dissertation. My biggest thanks go to my supervisor Markus Pauly. You believed in me and my abilities like no one before and you are a great role model for me as a supervisor and leader. Often when I have rebelled against authority, I have done so with the knowledge that things can be done differently. I have always had this conviction, but I have this knowledge because of you. I have nothing but love for Marc Ditzhaus. I am grateful that I was able to be around your genius and your kindness. That you will never read this fills me with sadness. You will always be in my heart. Big thanks go to my second supervisor Dennis Dobler. You always managed to fill the gap that Marc has left in the supervisory and organisational sense. Thanks for tightening up my knowledge in empirical process theory and having the same worldview in some sense. I would like to thank Philipp Doebler and Kirsten Schorning for being part of my PhD committee.

Thanks to all my colleagues. We had as many fruitful discussions as we had fun. I will always associate Spezi with the 2nd floor in the math tower. Felix Fesca, I appreciate your openness, your optimism, and your reflective nature, it was a pleasure to spend most of my time in the same room with you for four years. Paavo Sattler, you are the perfect older scientific brother who makes nonsense with me and I love that sometimes meaningful stuff happens then. Ludger Sandig, thank you for your advice regarding LaTeX and typography

in general. I really appreciate what we share with each other. Thanks to all my co-authors for the successful collaboration, especially Merle Munko and Kai-Philipp Gladow. Kai, I am proud of the fact that we are walking this path together in some sense and I hope that we will continue to do so in the future. It is very similar with you, Merle, I really appreciate your scientific opinion and I was very pleased that we were able to get close to each other while being co-authors.

I am very thankful for the invitation of Arne Bathke and Georg Zimmermann to visit Salzburg during my PhD and it was a great possibility to make scientific friends. That brought me to Konstantin Emil Thiel. I have no words for what you mean to me. I love your kindness, your sensitivity and what we create together. Please let us climb many more mountains.

The staff at the *Graduiertenzentrum* at TU Dortmund University helped me a lot with general questions about doing a doctorate and how to actually work. Andrea Hellbusch from DoBuS has often given me an absurd amount of courage. Many hugs and kisses go out to my favourite women from mentoring[3]. I have never been part of a girls crew before, but I love it. Often, it was very useful to have a mentor, thank you, Kathrin Möllenhoff, for your openness and the helpful advice in many situation. But above all, I would like to thank Jöran Krause and mobile e.V. for their support. You reminded me that I as an autistic person am allowed to be a human even if I sometimes feel like an alien.

I thank my family, especially my parents and my brother, for being able to accept that I have chosen the crazy path of science, even though they come from a completely different world. Even if this has meant that I have emancipated myself from you, it does not mean that I love you any less. Thanks to all my friends for enduring me while I wrote this dissertation. The biggest hug goes to Chris McGregor, you share your knowledge about English and educational theory with me and water my flowers, when I am on the way and I really like that.

For me, doing a PhD has always meant also being a Juso. Sometimes it felt like I had two lives, one at university and one in politics. And even if it comes from a different world, thinking about a more equitable world, questioning power structures, working with a wide variety of people, and the opportunity to lead has really shaped my life as a scientist and made me into the strong person I am today.

Wenn ich will, verkaufe ich Filme meiner Träume.

Heiner Pudelko (1992)

Contents

List of Symbols

$(\Omega, \mathcal{A}, P)$	a probability space with sample space Ω, σ-field $\mathcal{A}$ and probability measure P
$\mathbb{1}_A : \Omega \to \{0, 1\}$	the indicator function of a set $A \subset \Omega$
$\mathbb{R}$	the set of real numbers
$\mathbb{N}$	the set of natural numbers $\{1, 2, 3, \ldots\}$
k	number of groups in a factorial design
n_i	sample size per group, $i \in \{1, \ldots, k\}$
n	total sample size $\sum_{i=1}^{k} n_i$
Y_{ij}	jth realisation in group i of a real-valued random variable in a probability space $(\Omega, \mathcal{A}, P)$, $j \in \{1, \ldots, n_i\}$, $i \in \{1, \ldots, k\}$.
F_i	distribution function of random variables in group $i \in \{1, \ldots, k\}$
d	dimension of a random variable
$\mathbf{Y}_{ij}$	jth d-dimensional vector of realisations in group i of a multivariate random variable in a probability space $(\Omega, \mathcal{A}, P)$, $j \in \{1, \ldots, n_i\}$, $i \in \{1, \ldots, k\}$
$\boldsymbol{F}_i$	multivariate distribution function of random variables in group $i \in \{1, \ldots, k\}$
$\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\Sigma}}$	estimator for vector $\boldsymbol{\rho}$ or matrix $\boldsymbol{\Sigma}$
$\lvert\mathbf{e}\rvert$	Euclidean norm of the vector $\mathbf{e}$
$\bar{\mathbf{e}}_{i\cdot}$	mean over the dotted index
$\mathrm{E}(Y)$	expectation of random variable Y
$\mathrm{Var}(Y)$	variance of random variable Y
$\mathrm{Cov}(Y, X)$	covariance between random variable Y and random variable X
c	number of covariates in a linear model
u	number of estimands
r	number of tests in a multiple testing problem
$\mathbf{A}'$	transpose of matrix $\mathbf{A}$
$\mathbf{A}^+$	Moore-Penrose-Inverse of matrix $\mathbf{A}$
$\mathbf{A}^{-\frac{1}{2}}$	square root of matrix $\mathbf{A}$
$\mathbf{A}_0$	matrix containing the diagonal elements of $\mathbf{A}$ and zero otherwise
$\mathbf{1}_t$	t-dimensional column vector containing ones, $t \in \mathbb{N}$
$\mathbf{0}_t$	t-dimensional column vector containing zeros, $t \in \mathbb{N}$
$\mathbf{I}_t$	$t \times t$ identity matrix, $t \in \mathbb{N}$
$\mathbf{J}_t$	$t \times t$-dimensional matrix of ones, i.e. $\mathbf{J}_t = \mathbf{1}_t \mathbf{1}_t'$, $t \in \mathbb{N}$

$\mathbf{P}_t$	a $t \times t$-dimensional projection matrix, $\mathbf{P}_t = \mathbf{I}_t - \frac{1}{t}\mathbf{J}_t$, $t \in \mathbb{N}$
$\otimes$	Kronecker product of matrices
$\oplus$	direct sum of matrices
$\text{rank}(\mathbf{A})$	rank of matrix $\mathbf{A}$
$\text{tr}(\mathbf{A})$	trace of a square matrix $\mathbf{A}$
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
`blubb, blup`	Marc Ditzhaus' generalised variable for everything

List of Publications

The following three papers and preprints are part of this cumulative dissertation:

Article 1

Baumeister, M., Ditzhaus, M., & Pauly, M. (2024). *Quantile-Based MANOVA: A New Tool for Inferring Multivariate Data in Factorial Designs.* Journal of Multivariate Analysis, 199, 105246. https://doi.org/10.1016/j.jmva.2023.105246

Contribution of the author: The author of this thesis conducted the mathematical proofs, the methodology and implemented all simulations under the supervision of Marc Ditzhaus and Markus Pauly. She also prepared and structured the manuscript with input from the co-authors.

Article 2

Baumeister, M., Munko, M., Gladow, K.-P., Ditzhaus, M., Chakarov, N., & Pauly, M. (2025). *Early and Late Buzzards: Comparing Different Approaches for Quantile-Based Multiple Testing in Heavy-Tailed Wildlife Research Data.* Biometrical Journal, 67(4), e70065. https://doi.org/10.1002/bimj.70065

Contribution of the author: The idea of this project was given by Markus Pauly and Marc Ditzhaus and the author worked on the methodology together with Merle Munko. The allocation of tasks was such that the author was responsible for the project management, while Merle Munko did the main work in the programming of the simulation study. The data example has been investigated by Kai-Philipp Gladow and the author, while she calculated the data analysis. She also prepared and structured the manuscript with input from all co-authors.

Article 3

Baumeister, M., Thiel, K. E., Matits, L., Zimmermann, G., Pauly, M., & Sattler, P. (2025). *Multivariate and Multiple Contrast Testing in General Covariate-adjusted Factorial Designs* arXiv:2506.15292.

Contribution of the author: The idea for the project arose when the author spoke with Lynn Matits about the methodology of intervention studies in psychology. Markus Pauly and Georg Zimmermann provided methodological input. The author developed the mathematical proofs and the simulation studies with the help of Konstantin Emil Thiel under the supervision of Markus Pauly and Paavo Sattler. The data analysis was performed by Konstantin Emil Thiel with the help of the author. She prepared and structured the manuscript with input from the co-authors.

Part I.

Introduction and Statistical Background

525,600 minutes
525,000 moments so dear
525,600 minutes
How do you measure, measure a year?

Seasons of Love, Larson (1996a)

1. Introduction and Motivation

In daylights, in sunsets, in midnights, in cups of coffee
In inches, in miles, in laughter, in strife
In 525,600 minutes
How do you measure a year in the life?

How about love?
Measure in love
Seasons of love

Seasons of Love, Larson (1996a)

In various fields of science, testing problems can be structured as factorial designs. A common example would be the situation, where the effect in an intervention group should be distinguished from the effect in a control group. This situation occurs for example in psychology, when the effect of a certain therapy has to be investigated. In many cases, more than two groups have to be compared. For example, if in medicine the effect of several treatments has to be investigated against a control. Sometimes, different factors and their interactions are of interest, e.g. gender and age categories of patients. Moreover, in some cases, multivariate models have to be applied, because several endpoints are of interest or a quantity of interest is measured over a vector of attributes. All these testing problems can be understood as complex factorial designs, e.g. as a design with two-factors or as a multivariate model. Having powerful multiple testing procedures at hand is as much as important as the possibility to test in general factorial designs, because in more complex designs it is natural that questions regarding more than one hypothesis occur. For instance, in the situation of more than two groups or if the factorial design is multivariate, certain pairwise comparisons are of interest.

As research questions like these have to be answered in any quantitative science, it is essential that robust methods are available for empirical scientists. If a closer look is taken at these research questions, it becomes clear that the classic methods to handle factorial designs, analysis-of-variance (ANOVA) (Fisher, 1919) and multivariate analysis-of-variance (MANOVA) (Wilks, 1932) are not sufficient in their performance as they rely on strong assumptions such as normality and certain types of variance homogeneity. For example, Konietschke et al. (2015) point out that the assumption of a multivariate normal distribution is almost never fulfilled and White (1980) name various problems of homoscedasticity-assuming estimators that are used in context

of heteroscedastic data. Tu sum it up, these assumptions are plainly not realistic for many research questions. If they were used anyway, this would lead to non reliable testing decisions in the end.

To overcome these difficulties, it has been shown that resampling tests can be an essential key. A resampling-based critical value allows to mimic the real distribution of the test statistics without using an asymptotic distribution. Under certain circumstances, with an asymptotic distribution, it is often not possible to mimic the real distribution of the test statistics well. This situation can also occur in simple models in context of small sample sizes. Resampling can improve the performance of statistical tests in terms of type I error control and power on small and unbalanced samples, see e.g. Konietschke et al. (2015). Apart form that, with resampling, it can be handled a wider class of testing problems. The underlying theory of resampling tests is usually characterised by arguments of asymptotic statistics (e.g. Lehmann & Romano, 2022, Part II), which allows to develop valid testing procedures with less assumptions. Certain assumptions of a specific distribution class like normality or the homogeneity of the variance and covariance structure do not have to be made, which results in more robust testing procedures. Janssen and Pauls (2003) proved the underlying theory for general bootstrap and permutation testing procedures. This theory and the possibility to carry out resampling procedures easily due to fast computers make it possible to develop resampling testing procedures for a wider class of testing problems. Apart from asymptotic statistics, the theory of empirical processes (van der Vaart & Wellner, 2000) is useful to analyse the asymptotic behaviour of resampling test statistics. In the end, the methodological development has lead to powerful resampling alternatives for ANOVA (e.g. Pauly et al., 2015) and MANOVA (e.g. Konietschke et al., 2015).

The fact that theory for resampling-based testing is available has two other advantages. Firstly, there is the possibility to consider estimands beyond the mean. Secondly, resampling techniques allows to mimic the distribution of more general test statistics. Consequently, with resampling it has been possible to develop testing procedures on relative effects (Neubert & Brunner, 2007) and on linear combinations of quantiles (Ditzhaus et al., 2021) as well as test statistics that allow for singularity of the limiting covariance matrix (Friedrich & Pauly, 2018). To handle covariate-adjusted means instead of means, it is also appropriate to consider a general model with fewer assumptions by adapting ideas of White (1980). The standard methods of Fisher (1919) and Wilks (1932) are applicable, because the adjustment for covariates leads only to another kind of linear model. But the generality of this methods and the small sample performance can also be improved by the use of resampling, which was done in Zimmermann et al. (2019) for the univariate case and in Zimmermann et al. (2020) for the multivariate one.

Apart from that, resampling methods allow for multiple testing in various complex factorial designs. In context of simultaneous multiple testing procedures, which produce coherent and consonant testing decisions (Bretz et al., 2011), the method of multiple contrast testing procedures (MCTPs) has turned out to be useful. (Bretz et al., 2001) introduced MCTPs by considering critical values from theoretic distributions. Similarly to the single testing procedures, the method can be extended to other scenarios and estimands beyond the mean by using resampling procedures. This has been done, e.g., for univariate covariate-adjusted means (Becher et al., 2025) and for quantiles (Segbehoe et al., 2022).

This cumulative dissertation contributes to the expansion of methodological variety for inference in single and multiple simultaneous factorial designs by developing and systematically comparing methods in different factorial designs. All considered testing procedures are resampling-based, which substantially improves the test's behaviour in terms of type I error control and statistical power. The dissertation consists of three articles. The first publication Baumeister et al. (2024) introduces a *quantile-based* MANOVA. The method can be understand as the multivariate extension of the quantile-based ANOVA of Ditzhaus et al. (2021). Empirical process theory is used to prove that the method is theoretically valid. The second publication Baumeister, Munko, et al. (2025a) includes a *systematic comparison of univariate quantile-based simultaneous multiple testing procedures* by a Monte Carlo simulation. It gives a broad overview about the methodology of simultaneous multiple testing also for one-sided hypotheses. A motivational data example exemplifies that quantiles as estimands are highly applicable on animal data. The third Preprint Baumeister, Thiel, Matits, Zimmermann, et al. (2025) implements the concept of *multiple contrast test procedures in a semiparametric* MANCOVA. This allows to infer multivariate multiple testing problems on covariate-adjusted means and is highly relevant as more than one correlated endpoints occur frequently in medicine and psychology, e.g. in intervention studies.

The structure of this dissertation is as follows. In Chapter 2, the statistical background of the considered methods is briefly explained. Global and local testing problems are introduced in general in Section 2.1. For this purpose, two versions of test statistics (Section 2.1.1 and 2.1.2) as well as single and multiple test problems are considered in a common framework (Section 2.1.3 and 2.1.4). In the end of this chapter, existing testing procedures are contextualized within this framework. This reveals that there are methodological gaps that still need to be filled in the world of general factorial designs. The statistical models that are considered in the three articles are briefly introduced in Section 2.2. A quantile-based model on factorial designs is presented in Section 2.2.1 and a semiparametric MANCOVA model is given in Section 2.2.2. How the three included articles intend to fill some of these gaps based on the models in Chapter 2.2 is briefly described in Chapter 3. The dissertation closes with a discussion and an outlook in Chapter 4. In Part II, the three articles are given in full.

2. Statistical Background

525,600 minutes
525,000 journeys to plan
525,600 minutes
How do you measure the life of a woman or man?

In truths that she learned, or in times that he cried
In bridges he burned, or the way that she died

Seasons of Love, Larson (1996a)

2.1. Global and Local Testing Problems

The aim of this chapter is to explain general concepts of building test statistics that infer local and global statistical hypotheses in factorial designs. This requires a general framework. Assume independent realizations of group-wise identical distributed random variables $\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in_i}$, $i \in \{1, \ldots, k\}$, in $k \in \mathbb{N}$ groups. The model allows to consider multivariate data. In that case, $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijd})'$ is a vector of realizations of d-dimensional identical distributed multivariate random variables, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n_i\}$. Here and throughout, $n_i \in \mathbb{N}$ represents the sample size of group $i \in \{1, \ldots, k\}$ and $n := \sum_{i=1}^{k} n_i$ denotes the total sample size, cf. the List of Symbols. All in all, the realisations of random variables can be pooled to the vector $\mathbf{Y} = (\mathbf{Y}'_{11}, \ldots, \mathbf{Y}'_{kn_k})' \in \mathbb{R}^{nd}$. Many quantitative research questions of interest can be translated into hypotheses formulated in terms of estimands $\boldsymbol{\rho}$. Here, $\boldsymbol{\rho}$ can, e.g. be a median, quantile or mean. In that general setting, a population estimand ρ_i for each group $i \in \{1, \ldots, k\}$ is considered separately. In the multivariate case, the estimand can be considered component-wise, then a vector $\boldsymbol{\rho}_i = (\rho_{i1}, \ldots, \rho_{id})'$ for every group $i \in \{1, \ldots, k\}$ is of interest. Furthermore, it may be interesting to consider $u \in \mathbb{N}$ estimands in one model instead of one estimand, as it is the situation in the univariate quantile-based model, compare Section 2.2.1. Then, the different estimands have to be defined separately and can be pooled in a vector $\boldsymbol{\rho}_i = (\rho_{i1}, \ldots, \rho_{iu})'$ for every group $i \in \{1, \ldots, k\}$. As both situations can be combined, it is assumed that $\boldsymbol{\rho} \in \mathbb{R}^{kdu}$. In general, the following assumption is used throughout this thesis. Here and subsequently, all limits are defined for $n \to \infty$.

Assumption 1. *The groups do not vanish, i.e.* $n_i/n \to \kappa_i > 0$, $i \in \{1, ..., k\}$.

In order to make inference about the estimand $\boldsymbol{\rho}$, it is presupposed that there exists $\hat{\boldsymbol{\rho}}$, a consistent estimator for $\boldsymbol{\rho}$, for which a corresponding central limit theorem holds:

$$\sqrt{n}\left(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\right) \xrightarrow{d} \mathbf{N} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \tag{2.1}$$

Here, $\mathbf{N}$ is a multivariate normal random vector with expectation $\mathrm{E}(\mathbf{N}) = \mathbf{0}$ and limiting covariance $\mathrm{Cov}(\mathbf{N}) := \boldsymbol{\Sigma} \in \mathbb{R}^{kdu}$, which is assumed to be estimated consistently by an estimator $\hat{\boldsymbol{\Sigma}}$.

Let $\alpha \in (0, 1)$ be the level of significance. To define hypotheses of interest, choose a matrix $\mathbf{H} \in \mathbb{R}^{r \times kdu}$, $r \in \mathbb{N}$, and a vector of constants $\mathbf{e} = (e_1, \ldots, e_r)' \in \mathbb{R}^r$, that jointly encode the underlying research question using the subsequent notation. A general hypothesis can then be formulated as

$$\mathcal{H}_0 : \mathbf{H}\boldsymbol{\rho} = \mathbf{e} \text{ vs. } \mathcal{H}_1 : \mathbf{H}\boldsymbol{\rho} \neq \mathbf{e}. \tag{2.2}$$

Various testing problems are covered by this representation. For example, the one-way ANOVA hypothesis of no group effect of estimands ρ_i, $i \in \{1, \ldots, k\}$, is obtained by selecting the projection matrix $\mathbf{H} = \mathbf{P}_k$ and $d = 1$:

$$\mathcal{H}_0 : \mathbf{P}_k\boldsymbol{\rho} = \mathbf{0} \Leftrightarrow \mathcal{H}_0 : \rho_1 = \cdots = \rho_k.$$

Turning a univariate testing problem into a multivariate one usually works by using the Kronecker product with the corresponding hypothesis matrix of the univariate testing problem and $\mathbf{I}_d$. For example, a MANOVA hypothesis of no group effect for d dimensions $\mathcal{H}_0 : \boldsymbol{\rho}_1 = \cdots = \boldsymbol{\rho}_k$ can be obtained by $\mathbf{H} = \mathbf{P}_k \otimes \mathbf{I}_d$. The Kronecker product of the hypothesis matrix $\mathbf{H}$ and a suitable adjustment matrix can be used to infer linear combinations of more than one estimand together. Complexer layouts, for example a two-way layout (Pukelsheim, 2006, Section 1.27) can be realized by splitting up the group index i into $i = (i_1, i_2)$ for $i_1 \in \{1, \ldots, a\}$ and $i_2 \in \{1, \ldots, b\}$ resulting in $k = a \cdot b$ (sub-)groups. Then, two factors A and B with a respective b levels can be inferred. The estimand per group $\boldsymbol{\rho}_i$ can be decomposed into a general effect $\boldsymbol{\rho}^{\mu}$ that applies to all groups, main effects $\boldsymbol{\rho}^{\alpha}_{i_1}$, $\boldsymbol{\rho}^{\beta}_{i_2}$ and an interaction effect $\boldsymbol{\rho}^{\alpha\beta}_{i_1 i_2}$ as

$$\boldsymbol{\rho}_i = \boldsymbol{\rho}_{(i_1, i_2)} = \boldsymbol{\rho}^{\mu} + \boldsymbol{\rho}^{\alpha}_{i_1} + \boldsymbol{\rho}^{\beta}_{i_2} + \boldsymbol{\rho}^{\alpha\beta}_{i_1 i_2},$$

assuming the usual side conditions

$$\sum_{i_1=1}^{a} \boldsymbol{\rho}^{\alpha}_{i_1} = \sum_{i_2=1}^{b} \boldsymbol{\rho}^{\beta}_{i_2} = \sum_{i_1=1}^{a} \boldsymbol{\rho}^{\alpha\beta}_{i_1 i_2} = \sum_{i_2=1}^{b} \boldsymbol{\rho}^{\alpha\beta}_{i_1 i_2} = \mathbf{0}$$

to ensure identifiability (Pukelsheim, 2006, Section 3.9). Then, null hypotheses for main effects can be formulated for $d = 1$ and $u = 1$ as follows:

$$\mathcal{H}_0(A) : \left(\mathbf{P}_a \otimes \frac{1}{b}\mathbf{J}_b\right)\boldsymbol{\rho} = \mathbf{0} \Leftrightarrow \bar{\boldsymbol{\rho}}_{1\cdot} = \cdots = \bar{\boldsymbol{\rho}}_{a\cdot} \Leftrightarrow \boldsymbol{\rho}^{\alpha}_{i_1} = 0,\ \forall i_1 \in \{1, \ldots, a\},$$

$$\mathcal{H}_0(B) : \left(\frac{1}{a}\mathbf{J}_a \otimes \mathbf{P}_b\right)\boldsymbol{\rho} = \mathbf{0} \Leftrightarrow \bar{\boldsymbol{\rho}}_{\cdot 1} = \cdots = \bar{\boldsymbol{\rho}}_{\cdot b} \Leftrightarrow \boldsymbol{\rho}^{\beta}_{i_2} = 0,\ \forall i_2 \in \{1, \ldots, b\}.$$

Higher-way layouts and also hierarchical designs with nested factors can be realized in a similar way, see e.g., Friedrich and Pauly (2018) and Pauly et al. (2015).

Furthermore, the hypothesis $\mathcal{H}_0$ can be interpreted as a global hypothesis of a multiple testing problem. In this situation the matrix $\mathbf{H}$ is split in meaningful rows $\mathbf{h}_s$, $s \in \{1, \ldots, r\} \in \mathbb{R}^{kdu}$: $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_r)' \in \mathbb{R}^{r \times kdu}$ corresponding to the local hypotheses. In fact, every vector $\mathbf{h}_s$, $s \in \{1, \ldots, r\}$, itself characterises a hypothesis of interest:

$$\mathcal{H}_{0,s} : \mathbf{h}_s' \boldsymbol{\rho} = e_s \text{ vs. } \mathcal{H}_{1,s} : \mathbf{h}_s' \boldsymbol{\rho} \neq e_s. \tag{2.3}$$

Building the intersections of the local hypotheses leads to the global hypothesis:

$$\bigcap_{s=1}^{r} \mathcal{H}_{0,s} = \{\forall s \in \{1, \ldots, r\} : \mathbf{h}_s' \boldsymbol{\rho} = e_s\} = \{\mathbf{H}\boldsymbol{\rho} = \mathbf{e}\} = \mathcal{H}_0.$$

The hypothesis $\mathcal{H}_0$, understood as a family of local hypotheses, covers various multiple testing problems of interest. For a single estimand $\boldsymbol{\rho}$ ($u = 1$), we can formulate hypotheses that are well known for vectors of means, compare Bretz et al. (2011) and Konietschke et al. (2013). For example, it is possible to consider the following test problems: all-pairs comparisons can be incorporated by choosing $u = 1$ and the Tukey-type matrix (Tukey, 1994) as contrast matrix $\mathbf{H}$. This leads to the hypotheses

$$\mathcal{H}_{0,s_1 s_2} : \rho_{\ell_1} - \rho_{\ell_2} = \epsilon_{s_1 s_2}$$

of all-pairs comparisons in one-way layouts, where $s_1, s_2 \in \{1, \ldots, k\}$, $s_1 > s_2$. Similarly, choosing the Dunnett-type matrix (Dunnett, 1955) gives the hypotheses

$$\mathcal{H}_{0,s} : \rho_s - \rho_1 = \epsilon_s$$

of many-to-one comparisons, $s \in \{2, \ldots, k\}$. Choosing the Grand-mean-type matrix (Djira & Hothorn, 2009) and $\boldsymbol{\epsilon} = \mathbf{0}$ instead leads to the hypotheses

$$\mathcal{H}_{0,s} : \rho_s = \bar{\rho},$$

where the estimand is compared to the mean $\bar{\rho} := k^{-1} \sum_{i=1}^{k} \rho_i$ of all group-wise estimands in one-way layouts, $s \in \{1, \ldots, k\}$.

In this framework, two different types of test statistics turned out to be feasible to infer the testing problems. The first one is the general t-type test statistic and the second one is the quadratic form-type statistic. Both are going to be explained in general in the next two sections.

2.1.1. t-Type Test Statistics

Consider the vectors $\mathbf{h}_s$ and the constants $\mathbf{e}$, $s \in \{1, \ldots, r\}$. Then, a test statistic can be defined as follows:

$$A_n(\mathbf{h}_s, e_s) = \sqrt{n}\, \frac{\mathbf{h}_s' \hat{\boldsymbol{\rho}} - e_s}{\sqrt{\mathbf{h}_s' \hat{\mathbf{D}} \mathbf{h}_s}}, \tag{2.4}$$

where $\hat{\mathbf{D}}$ is some symmetric matrix. For inference the following condition on these matrix can be useful.

Assumption 2. *The matrix $\hat{\mathbf{D}}$ estimates consistently some positive definite and symmetric matrix $\mathbf{D}$.*

Usually, the estimator $\hat{\mathbf{D}}$ is set to the estimator $\hat{\mathbf{D}} = \hat{\boldsymbol{\Sigma}}$, that consistently estimates the covariance matrix $\boldsymbol{\Sigma}$. For a more general framework, it can be set to $\hat{\mathbf{D}} = \hat{\boldsymbol{\Sigma}}_0$. If the test statistic in (2.4) is calculated for the matrix $\mathbf{H}$, the result is a vector of r test statistics:

$$A_n(\mathbf{H}, \mathbf{e}) = \left(\mathbf{H}\hat{\mathbf{D}}\mathbf{H}\right)_0^{-1/2} \sqrt{n}\, (\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}) = (A_n(\mathbf{h}_1, e_1), \ldots, A_n(\mathbf{h}_r, e_r))'.$$

Its asymptotic behaviour is analysed in the following theorem:

Theorem 3. *With the Central Limit Theorem (2.1) and under the Assumption 2, it holds:*

1. *Under $\mathcal{H}_0 : \mathbf{H}\boldsymbol{\rho} = \mathbf{e}$ the vector of test statistics $A_n(\mathbf{H}, \mathbf{e})$ converges in distribution to a multivariate normal distribution, i.e.*

$$A_n(\mathbf{H}, \mathbf{e}) = (A_n(\mathbf{h}_1, e_1), \ldots, A_n(\mathbf{h}_r, e_r))' \xrightarrow{d} \mathbf{A},$$

   where the random vector $\mathbf{A} = (A_1, \ldots, A_r)$ is r-dimensional and has the expectation $\mathrm{E}(\mathbf{A}) = \mathbf{0}$ and the covariance matrix

$$\mathbf{R} := \mathrm{Cov}\,(\mathbf{A}) = (\mathbf{H}\mathbf{D}\mathbf{H}')_0^{-\frac{1}{2}}\, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\, (\mathbf{H}\mathbf{D}\mathbf{H}')_0^{-\frac{1}{2}}.$$

2. *Under $\mathcal{H}_1 : \mathbf{H}\boldsymbol{\rho} \neq \mathbf{e}$, $A_n(\mathbf{H}, \mathbf{e})$ converges in probability to ∞.*

Proof. Under $\mathcal{H}_0$, it can be argued by Slutzky's (e.g. van der Vaart & Wellner, 2000, Example 1.4.7) and the Continuous Mapping (e.g. van der Vaart & Wellner, 2000, Thm. 1.11.1) Theorem, that

$$\begin{aligned} A_n(\mathbf{H}, \mathbf{e}) &= (\mathbf{H}\hat{\mathbf{D}}\mathbf{H}')_0^{-1/2} \sqrt{n}\, (\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}) \\ &= (\mathbf{H}\hat{\mathbf{D}}\mathbf{H}')_0^{-1/2} \mathbf{H} \sqrt{n}\, (\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}) \xrightarrow{d} (\mathbf{H}\mathbf{D}\mathbf{H}')_0^{-1/2} \mathbf{H}\mathbf{N} =: \mathbf{A}. \end{aligned}$$

From the distribution of the random variable $\mathbf{N}$ it follows that $\mathbf{A} = (A, \ldots, A_r)$ has a multivariate normal distribution with expectation $\mathrm{E}(A_n(\mathbf{H})) = \mathbf{0}$ and covariance matrix

$$\mathbf{R} := \mathrm{Cov}\left(A_n(\mathbf{H}, \mathbf{e})\right) = \left(\mathbf{H}\mathbf{D}\mathbf{H}'\right)_0^{-\frac{1}{2}} \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' \left(\mathbf{H}\mathbf{D}\mathbf{H}'\right)_0^{-\frac{1}{2}}.$$

This is the first assertion. Let $s \in \{1, \ldots, r\}$ be arbitrary and fixed. Then, $A_n(\mathbf{h}_s, e_s)$ is already under $\mathcal{H}_{0,s}$ asymptotically distributed like A_s. In general, it can be argued that $|A_n(\mathbf{h}_s, e_s)| / \sqrt{n}$ converges in probability:

$$\frac{1}{\sqrt{n}} |A_n(\mathbf{h}_s, e_s)| = \frac{|\mathbf{h}_s'\hat{\boldsymbol{\rho}} - e_s|}{\sqrt{\mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s}} \xrightarrow{p} \frac{|\mathbf{h}_s'\boldsymbol{\rho} - e_s|}{\left|\sqrt{\mathbf{h}_s'\mathbf{D}\mathbf{h}_s}\right|}.$$

Under $\mathcal{H}_{1,s}$, this limiting value is greater than zero. From this we can conclude that, under $\mathcal{H}_{1,s}$, the test statistics $A_n(\mathbf{h}_s, \mathbf{e})$ converges in probability to ∞ for all $s \in \{1, \ldots, r\}$, which is the second assertion. □

It is a consequence of Theorem 3 that the limiting random variables A_s, $s \in \{1, \ldots, r\}$, are normally distributed:

$$A_s \sim \mathcal{N}\left(0, \frac{\mathbf{h}_s'\boldsymbol{\Sigma}\mathbf{h}_s}{\mathbf{h}_s'\mathbf{D}\mathbf{h}_s}\right). \tag{2.5}$$

Therefore, the limiting distribution of $A_n(\mathbf{h}_s, e_s)$ can have a variance that is unequal to 1 using a general matrix $\hat{\mathbf{D}}$. If $\hat{\mathbf{D}} = \hat{\boldsymbol{\Sigma}}$ is chosen, the random variables A_s are asymptotically standard normally distributed and a critical value form the standard normal distribution can be used for testing. The advantage of this is explained in Section 2.1.4. In the general case (2.5), it is also possible to use appropriate resampling methods to determine the critical value for a test decision.

2.1.2. Quadratic Form-Type Statistics

Another possibility to infer the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_{0,s}$ is through quadratic form-type test statistics, which can be generally defined as

$$U_n(\mathbf{H}, \mathbf{e}) = n \left(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}\right)' \Xi\left(\mathbf{H}, \hat{\boldsymbol{\Sigma}}\right) \left(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}\right). \tag{2.6}$$

Here, Ξ is a function of $\mathbf{H}$ and $\hat{\boldsymbol{\Sigma}}$ and can be set to different forms. A classic version is the Wald-type statistic (WTS) (Wald, 1943):

$$\Xi\left(\mathbf{H}, \hat{\boldsymbol{\Sigma}}\right)_{\mathrm{WTS}} = \left(\mathbf{H}\hat{\boldsymbol{\Sigma}}\mathbf{H}'\right)^{+},$$

where inferring it usually requires that the covariance matrix $\mathbf{\Sigma}$ is non-singular. Two quadratic form-type statistics without that assumption are the ANOVA-type statistic (ATS) (Brunner et al., 1997) and the modified ANOVA-type statistic (MATS) (Friedrich & Pauly, 2018):

$$\Xi\Big(\mathbf{H}, \hat{\mathbf{\Sigma}}\Big)_{\text{ATS}} = \mathbf{I}_r / \text{tr}\Big(\mathbf{H}\hat{\mathbf{\Sigma}}\mathbf{H}'\Big), \quad \Xi\Big(\mathbf{H}, \hat{\mathbf{\Sigma}}\Big)_{\text{MATS}} = \Big(\mathbf{H}\hat{\mathbf{\Sigma}}_0\mathbf{H}'\Big)^{+}.$$

For the latter, only the diagonal elements of the covariance matrix $\mathbf{\Sigma}$ needs to be positive. In principle the function Ξ can be chosen as an arbitrary function, where the resulting matrix has to be $r \times r$-dimensional and symmetric, e.g.

$$\Xi\Big(\mathbf{H}, \hat{\mathbf{\Sigma}}\Big) = \Xi\Big(\mathbf{H}, \hat{\mathbf{\Sigma}}\Big)'.$$

To get a useful test statistic, but not for technical reasons, the resulting matrix $\Xi(\mathbf{H}, \hat{\mathbf{\Sigma}})$ has to be positive semidefinite. Moreover, the following assumption has to be made:

Assumption 4. *The function* $\Xi\Big(\mathbf{H}, \hat{\mathbf{\Sigma}}\Big)$ *is consistent for* $\Xi(\mathbf{H}, \mathbf{\Sigma})$.

This assumption is usually concluded from the consistency of $\hat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$. Apart from that, one can observe that it holds $A_n(\mathbf{h}_s, e_s)^2 = U_n(\mathbf{h}_s, e_s)$ for $\Xi(\mathbf{h}_s, \hat{\mathbf{D}}) = 1/(\mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s)$, which highlights the similarities of the two approaches. In Section 2.1.3 and 2.1.4, the advantages of the different respective test statistics will be made clear. To prove the behaviour of the test statistic $U_n(\mathbf{H}, \mathbf{e})$ under the alternative, the following assumption is needed:

Assumption 5. *Under* $\mathcal{H}_1 : \mathbf{H}\boldsymbol{\rho} \neq \mathbf{e}$, *it holds* $(\mathbf{H}\boldsymbol{\rho} - \mathbf{e})' \, \Xi\,(\mathbf{H}, \mathbf{\Sigma})\,(\mathbf{H}\boldsymbol{\rho} - \mathbf{e}) > 0$.

The following can be stated about the asymptotic distribution of the test statistic $U_n(\mathbf{H}, \mathbf{e})$.

Theorem 6. *Assume the central limit theorem (2.1) and Assumptions 4 and 5.*

1. *Under the null hypothesis* $\mathcal{H}_0 : \mathbf{H}\boldsymbol{\rho} = \mathbf{e}$ *the test statistic* $U_n(\mathbf{H}, \mathbf{e})$ *converges in distribution to a weighted sum of* χ^2_1*-distributed random variables, i.e.*

$$U_n(\mathbf{H}, \mathbf{e}) \xrightarrow{d} \mathbf{U} = \sum_{s=1}^{r} \lambda_s \mathbf{U}_s,$$

where $\mathbf{U}_s \overset{iid}{\sim} \chi^2_1$ *and the weights* $\lambda_s \geq 0$, $s \in \{1, \ldots, r\}$, *are the eigenvalues of the matrix* $(\mathbf{H}\mathbf{\Sigma}\mathbf{H}')^{1/2}\Xi(\mathbf{H}, \mathbf{\Sigma})(\mathbf{H}\mathbf{\Sigma}\mathbf{H}')^{1/2}$.

2. Under $\mathcal{H}_1 : \mathbf{H}\boldsymbol{\rho} \neq \mathbf{e}$ *the test statistic* $U_n(\mathbf{H}, \mathbf{e})$ *converges in probability to* ∞*.*

Proof. To start with the first statement: under $\mathcal{H}_0$ and by the Continuous Mapping Theorem (van der Vaart & Wellner, 2000, Theorem 1.11.1), it follows from the CLT (2.1)

$$\begin{aligned}\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}) &= \sqrt{n}\Big[(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}) - (\mathbf{H}\boldsymbol{\rho} - \mathbf{e})\Big] \\ &= \mathbf{H}\sqrt{n}\left(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\right) \xrightarrow{d} \mathbf{HN} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)\end{aligned}$$

Due to Assumption 4, the consistency of the mid term $\Xi\Big(\mathbf{H}, \hat{\boldsymbol{\Sigma}}\Big)$, and a combination of Slutzky's theorem and the Continuous Mapping Theorem (van der Vaart & Wellner, 2000, Example 1.4.7), it follows

$$\begin{aligned}U_n(\mathbf{H}, \mathbf{e}) &= n\Big(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}\Big)'\Xi\Big(\mathbf{H}, \hat{\boldsymbol{\Sigma}}\Big)\Big(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}\Big) \\ &= \sqrt{n}\Big(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}\Big)'\Xi\Big(\mathbf{H}, \hat{\boldsymbol{\Sigma}}\Big)\sqrt{n}\Big(\mathbf{H}\hat{\boldsymbol{\rho}} - \mathbf{e}\Big) \xrightarrow{d} (\mathbf{HN})'\,\Xi\left(\mathbf{H}, \boldsymbol{\Sigma}\right)\mathbf{HN}.\end{aligned}$$

By Mathai and Provost (1992, p. 90), this has the same distribution as the random variable $\mathbf{U}$. The second statement follows immediately. It can be argued that $n^{-1}U_n(\mathbf{H}, \mathbf{e})$ always converges in probability to $(\mathbf{H}\boldsymbol{\rho} - \mathbf{e})'\,\Xi\left(\mathbf{H}, \boldsymbol{\Sigma}\right)(\mathbf{H}\boldsymbol{\rho} - \mathbf{e})$. From this, the assertion can be concluded with Assumption 5. □

If the WTS is considered, the quadratic form-type statistic has a simpler limiting distribution than the weighted sum of χ^2_1-distributed random variables stated in Theorem 6. As the whole covariance estimator $\hat{\boldsymbol{\Sigma}}$ is used in the function Ξ, it can be argued, that the WTS converges in distribution to a χ^2-distribution with $\operatorname{rank}(\mathbf{H})$ degrees of freedom: $\mathrm{WTS}(\mathbf{H}, \mathbf{e}) \xrightarrow{d} \mathbf{U}_{\mathrm{WTS}} \sim \chi^2_{\operatorname{rank}(\mathbf{H})}$. In contrast, the general limiting distribution of $U_n(\mathbf{H}, \mathbf{e})$ includes unknown parts. Therefore, for arbitrary choices of the function Ξ, it is not possible to consider a limiting distribution for the calculation of critical values. For this reason, resampling methods are generally utilized to determine the critical values for quadratic form-type statistics. In general, the choice of the hypothesis matrix $\mathbf{H}$ is not unique, several matrices leads to the same testing problem. Having only the global hypothesis in mind, and if $\mathbf{e} = \mathbf{0}$, there is the convention to use the unique projection matrix $\mathbf{T} = \mathbf{H}'(\mathbf{H}'\mathbf{H})^{+}\mathbf{H} \in \mathbb{R}^{kdu \times kdu}$ of $\mathbf{H}$, which is symmetric, unique, and idempotent. If $\mathbf{e} \neq \mathbf{0}$, the choice of the hypothesis matrix is not unique, see Sattler and Zimmermann (2024) and Sattler and Rosenbaum (2025) for a discussion on appropriate choices.

2.1.3. Global Testing Procedures

In this section, the aim is to infer a single hypothesis as given in Equation (2.2). In general, t-type and quadratic form-type statistics can be applied in this context.

Firstly, assume that the hypothesis matrix $\mathbf{H}$ has only one row ($r = 1$). Then, the t-type test statistic $A_n(\mathbf{H}, \mathbf{e}) = A_n(\mathbf{h}, e)$ can be used. For an appropriate critical value $w^A(\alpha)$, which depends in some way on the global level α, the test would be

$$\varphi_n = \mathbb{1}\left\{|A_n(\mathbf{h}, e)| > w^A(\alpha)\right\}.$$

If the critical value $w^A_{1-\alpha}$ is chosen in line with the asymptotic distribution of the test statistic $A_n(\mathbf{h}, e)$ under $\mathcal{H}_0$, e.g. it is a $(1-\alpha)$-quantile of an asymptotic distribution or of a resampling distribution under $\mathcal{H}_0$, it can be argued that φ_n is an asymptotic level-α test, i.e. $\mathrm{E}_{\mathcal{H}_0}(\varphi_n) \to \alpha$. From Theorem 3.2. it can be also concluded that φ_n is consistent for the alternative $\mathcal{H}_1 : \mathbf{h}'\boldsymbol{\rho} \neq e$. For further details, compare the results of Janssen and Pauls (2003).

Examples for t-Type Tests In the following, it is explained that some existing testing procedures are special cases of the testing procedure φ_n. In general, single estimands ($u = 1$) with one dimension ($d = 1$), and one or two groups ($k \in \{1, 2\}$) can be considered by φ_n. An important example of φ_n is the standard t-test (Student, 1908), which considers quantiles of t-distributions as critical values. In fact, any versions of asymptotic t-tests can be understood as special cases of φ_n. In the one-sample case ($k = 1$) of a t-test the one-dimensional mean μ is considered, what leads to the hypothesis $\mathcal{H}_0 : \mu = e$. In the two-sample case ($k = 2$), the hypothesis would be $\mathcal{H}_0 : \mu_1 = \mu_2$. Inferring the latter hypothesis, while at the same time allowing unequal variances in the two groups, is called the Behrens-Fisher problem (Fisher, 1935) and was for example solved by the Welch-Test (Welch, 1938). While the mentioned tests assume normally distributed data, the following tests do not assume any distribution class. Konietschke and Pauly (2014) consider resampling-based critical values in a mean-based paired hypotheses. There are also further testing procedures with estimands beyond the mean, which can be understood as special cases of the test φ_n. The Mann-Whitney U test (Deuchler, 1914) and the Brunner-Munzel test (Brunner & Munzel, 2000) as well as the version for paired samples (Munzel, 1999) are a nonparametric equivalent to the mean-based two-sample t-tests and consider the relative effect. For this estimand, there are also resampling-based versions available (Konietschke & Pauly, 2012; Neubert & Brunner, 2007). Bonett and Price (2002) describe tests like φ_n for linear combinations of medians $\mathbf{m}$, e.g. they consider the hypothesis $\mathcal{H}_0 : \mathbf{h}'\mathbf{m} = e$, and Chung and Romano (2013) discuss resampling methods for median-based tests.

If a hypothesis matrix $\mathbf{H}$ is considered ($r > 1$), the use of the test statistic $A_n(\mathbf{H}, \mathbf{e})$ leads to a vector of test statistics, as explained in Section 2.1.1. That the matrix $\mathbf{H}$ has more than one row occurs concretely, if more than two estimands are compared, e.g. if $k > 2$, $d > 1$, or $u > 1$ holds. In these situations, quadratic form-type statistic

$U_n(\mathbf{H}, \mathbf{e})$ can be used to create a single test:

$$\Phi_n = \mathbb{1}\left\{U_n(\mathbf{H}, \mathbf{e}) > w^U(\alpha)\right\}.$$

With the same arguments as for the testing procedure φ_n, it can be argued that the test Φ_n is an asymptotic level-α test and consistent for the alternative, if the critical value $w^U(\alpha)$ mimics the asymptotic distribution of the test statistics.

Examples for Univariate Testing with Quadratic Forms Firstly, I want to refer to univariate testing problems that are special cases of the testing procedure Φ_n, where it holds $d = 1$ and possibly more than one comparison has to be made. This means especially, that with the test Φ_n, it is possible to compare estimands of more than two groups ($k > 2$), e.g. it is possible to consider hypotheses like $\mathcal{H}_0 : \rho_1 = \cdots = \rho_k$. These hypotheses and more complex univariate layouts, as stated at the beginning of Section 2.1, are covered by the often-used analysis-of-variance (ANOVA), which was originally developed on means by Fisher (1919). There, normally distributed data is assumed, the critical value is taken from an F-distribution and a quadratic form-type statistic similar to the WTS (see Fahrmeir et al., 2013, Section 3.3) is considered. Therefore, these tests are versions of the test Φ_n. To infer the WTS as defined in Section 2.1.2, Wald (1943) introduced a quadratic form-type test for large samples, which uses a critical value from the χ^2-distribution. In principle, any meaningful hypothesis matrix $\mathbf{H}$ can be considered with this methodology. Pauly et al. (2015) also used the WTS and made it available for small sample sizes by developing a permutation approach. Ditzhaus et al. (2021) chose the WTS and a permutation approach to develop a quantile-based ANOVA in the univariate framework of Section 2.2.1. As explained there, the methods consider hypotheses with general linear combinations of more than one quantile. A nonparametric alternative to the ANOVA was introduced by Brunner et al. (1997). This method infers hypotheses regarding the distribution functions F_i, considers the relative effect and uses the ATS with an asymptotic approximation. The authors discuss also possibilities to infer a mean-based ATS and compare both approaches. Thinking of the F-test in linear models, it is possible to consider both factorial and regression parts together in one model. This has the advantage, that means in factorial designs can be adjusted by metric covariables, which is named as analysis-of-covariance (ANCOVA). In the end, such a model allows to do testing on covariate-adjusted means. This is explained for vectors of adjusted means in Section 2.2.2. Zimmermann et al. (2019) investigated bootstrap methods for a testing procedure based on the WTS for that situation. Bathke and Brunner (2003) created a nonparametric model for testing on covariate-adjusted estimands that uses a quadratic form-type test statistic and Thiel et al. (2024) improved the small-sample performance of this method through bootstrapping.

Examples for Multivariate Testing with Quadratic Forms Apart from that, with quadratic form-type statistics, it is possible to infer factorial designs on a multivariate estimand $\boldsymbol{\rho}_i \in \mathbb{R}^d$, $i \in \{1, \ldots, k\}$. The multivariate version of a two-sample t-test ($d > 1$, $k = 2$) is Hotelling's T^2 (Hotelling, 1931) with the mean-based hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_2 = \boldsymbol{\mu}_2$. It uses a quadratic form as test statistic and the Hotelling's T^2-distribution for the calculation of the critical values. To realise that normally distributed data is assumed. The generalizations of this method for more than two groups ($k > 2$) are Hotelling-Lawley Trace (Hotelling, 1951; Lawley, 1938), Pillai-Barlett Trace (Bartlett, 1939; Pillai, 1955), Roy's Largest Root (Roy, 1953; Roy, 1945) and Wilk's Lambda (Wilks, 1932), which are based on quadratic form-type statistics as well. General multivariate factorial designs are as well considered by Konietschke et al. (2015). They used a semiparametric model, similar to the model considered in Section 2.2.2, investigated the WTS as test statistic and several bootstrap procedures as resampling techniques. Friedrich and Pauly (2018) generalized this model for singular data and considered the MATS and bootstrapping. Sattler and Pauly (2018) investigated the use of quadratic form-type statistics in the situation of high-dimensional data. Zimmermann et al. (2020) implemented the MATS in the semiparametric multivariate analysis-of-covariance (MANCOVA) model that is stated in Section 2.2.2. Thereby they generalized the methods of Friedrich and Pauly (2018) in a way that it is applicable for covariate-adjusted means. A solution for the multivariate nonparametric Behrens-Fisher problem is discussed in Brunner et al. (2002) and WTS and ATS as well as asymptotic critical values are used. A nonparametric MANOVA on general factorial designs in presented in Dobler et al. (2020), where they consider an ATS and several bootstrap techniques.

2.1.4. Simultaneous Multiple Testing Procedures

Apart from that, it is also possible to consider simultaneous multiple testing procedures within that framework. For an appropriate family of hypotheses $\bigcap_{s=1}^{r} \mathcal{H}_{0,s}$ with corresponding hypothesis vectors $\mathbf{h}_s$, $s \in \{1, \ldots, r\}$, the resulting matrix $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_r) \in \mathbb{R}^{r \times kdu}$ and the vector of constants $\mathbf{e}$, it is possible to consider a multiple testing problem by using the test statistics $A_n(\mathbf{h}_s, e_s)$ for the local hypotheses. For appropriate critical values $w_s^A(\alpha)$, $s \in \{1, \ldots, r\}$, the following test decisions for the multiple testing problem is received:

1. for each $s \in \{1, ..., r\}$, $\mathcal{H}_{0,s}$ is rejected if and only if $|A_n(\mathbf{h}_s, e_s)| > w_s^A(\alpha)$,

2. the global hypothesis $\mathcal{H}_0 = \bigcap_{s=1}^{r} \mathcal{H}_{0,s}$ is rejected if and only if at least one $\mathcal{H}_{0,s}$ is rejected, i.e. if $\max_{s \in \{1,...,r\}} |A_n(\mathbf{h}_s, e_s)| / w_s^A(\alpha) > 1$.

Equivalently, for every local hypothesis $\mathcal{H}_{0,s}$, $s \in \{1, ..., r\}$, the test

$$\varphi_{n,s}^{M} = \mathbb{1}\Big\{|A_n(\mathbf{h}_s, e_s)| > w_s^A(\alpha)\Big\}$$

can be defined and for the global hypothesis $\mathcal{H}_0$ the test would be

$$\varphi_{n}^{M} = \max_{s\in\{1,...,r\}} \mathbb{1}\left\{\frac{|A_n(\mathbf{h}_s, e_s)|}{w_s^A(\alpha)} > 1\right\}.$$

Similarly to the single testing procedures, for an appropriate choice of the critical value $w_s^A(\alpha)$ it can be argued that the local tests are asymptotic level-α tests, which are consistent for the alternative. Thereby, the idea of this simultaneous testing procedure is to redefine the rejection of the global hypothesis: the global hypothesis is rejected simultaneously if at least one of the local hypotheses is rejected. Then, it is possible to define the critical value in a way that the multiple testing procedure controls the family-wise type I error rate (FWER). Furthermore, it is transparent by construction, which of the local tests are responsible for the global rejection and both local and global test decisions are coherent and consonant, see Bretz et al. (2011). Other approaches, e.g. other type of post hoc t-tests, do not have these advantages. Konietschke et al. (2013) pointed out that MCTPs provide more information about the process of rejecting the hypotheses. All in all, this testing problem complies with the union-intersection principle introduced by Roy (1953). The test φ_n^M is in fact a so-called max-t tests (Bretz et al., 2001) for an arbitrary estimand, see Pigeot (2000) for a methodologically overview about multiple testing. It is characteristic for simultaneous testing procedures that it is possible to define equivalent simultaneous confidence intervals for the vectors $\mathbf{h}_s'\boldsymbol{\rho}$:

$$\left[\mathbf{h}_s'\hat{\boldsymbol{\rho}} - \sqrt{\mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s}\frac{w_s^A(\alpha)}{\sqrt{n}}, \mathbf{h}_s'\hat{\boldsymbol{\rho}} + \sqrt{\mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s}\frac{w_s^A(\alpha)}{\sqrt{n}}\right], \quad s \in \{1, ..., r\}.$$

There are several options to determine a critical value for answering the testing procedures φ_n^M and $\varphi_{n,s}^M$. An intuitive and well-known method is the Bonferroni correction (Dunn, 1961), where each individual hypothesis is tested at a smaller local level of α/r. The fact that this method controls the FWER can be derived from the Bonferroni inequality. In the framework of this dissertation, one must consider the $(1-\alpha/(2r))$-quantile of the limiting distribution $|A_s|$ of the test statistics $|A_n(\mathbf{h}_s, e_s)|$ for every local hypothesis $s \in \{1, \ldots, r\}$. An advantage of this method is that it can be easily combined with various critical values in any kind of model, but by construction it is also clear that it leads to conservative test decisions when a large number of tests is inferred ($r \gg 1$). Another disadvantage can be that dependencies between local tests are not leveraged with the Bonferroni correction.

The concept of multiple contrast test procedures (MCTPs) overcomes these problems by exploiting the joint distribution of the vector of test statistics $A_n(\mathbf{H}, \mathbf{e})$. One strategy to do this was introduced by Gabriel (1969) giving a general introduction of simultaneous testing procedures and was made numerically available by Bretz et al. (2001). As explained in Section 2.1.1, if $\mathbf{D} = \mathbf{\Sigma}$ is chosen, the local test statistics $A_n(\mathbf{h}_1, e_1), \ldots, A_n(\mathbf{h}_r, e_r)$ have the same marginal limit distribution, and it is possible to consider the same critical value for all local hypotheses. To take the correlation of the local test statistics into account, the multivariate limiting distribution $\mathbf{A}$ of the test statistics $A_n(\mathbf{h}_s, e_s)$, $s \in \{1, \ldots, r\}$ in Theorem 3 is used. Precisely, the $(1-\alpha)$-quantile of the conditional distribution of $\max_{s\in\{1,\ldots,r\}}|A_s|$ is denoted as equicoordinate quantile of $|\mathbf{A}|$ and can be considered as critical value.

Examples for Multiple Contrast Test Procedures The concept of MCTPs has been implemented in various methods and all of them can be understood as special cases of the testing procedure φ_n^M. In Bretz et al. (2001), the mean is considered as estimand and, apart from the normal distribution, a comparison with the multivariate t-distribution is discussed. Hasler and Hothorn (2008) extended the method for heteroscedastic data. There are methods for high dimensional and functional scenarios with metric outcomes (Konietschke et al., 2021; Munko et al., 2023). Hasler and Hothorn (2011) and Hasler (2014) introduced MCTPs for multiple endpoints, while Becher et al. (2025) introduced covariate-adjusted MCTPs for univariate outcomes. Moreover, there are even rank-based MCTPs for univariate (Konietschke et al., 2012; Noguchi et al., 2020), repeated measures (Umlauft et al., 2019) and other complex designs (Rubarth et al., 2022). Segbehoe et al. (2022) introduced MCTPs for quantiles. If a multiple testing problem containing vectorized hypotheses as local ones has to be considered, the union intersection principle can be applied to quadratic form-type statistics, which is discussed in Sattler et al. (2025).

2.2. Statistical Models

The three articles contained in this dissertation are based on two different models, a quantile-based model and a semiparametric multivariate analysis-of-covariance (MANCOVA) model. The quantile-based model is considered in its multivariate version in the first article (Baumeister et al., 2024) and in the second second article in its univariate version (Baumeister, Munko, et al., 2025a). The third article (Baumeister, Thiel, Matits, Zimmermann, et al., 2025) refers to the semiparametric MANCOVA model. In the following two subsections, the models are explained in detail. Compare the List of Symbols for details about the notation.

2.2.1. A Model for Inferring (Multivariate) Quantiles in Factorial Designs

Univariate Model Assume $k \in \mathbb{N}$ mutually independent samples $Y_{i1}, \ldots, Y_{in_i} \sim F_i$, $i \in \{1, \ldots, k\}$, where F_i are distribution functions and the corresponding density is denoted by f_i. In the univariate quantile-based model it is possible to consider more than one quantile of interest at a time. To this end, let $0 < p_1 < \ldots < p_u < 1$ denote $u \in \mathbb{N}$ pre-specified quantile levels with corresponding quantiles

$$q_{iv} := F_i^{-1}(p_v) = \inf\{t \in \mathbb{R} \mid F_i(t) \geq p_v\}, \quad i \in \{1, \ldots, k\},\, v \in \{1, \ldots, u\}.$$

This means in particular that this model considers potentially one or more percentiles. Typical statistical questions of interest correspond to $p = 0.5$ and $u = 1$ for the median or to $p_1 = 0.25$, $p_2 = 0.75$ with $u = 2$ for the inter quartile range (IQR). The pooled quantile vector is denoted by $\mathbf{q} := (q_{11}, \ldots, q_{1u}, q_{21}, \ldots, q_{ku})'$. For asymptotic analyses with empirical quantiles, the following is assumed:

Assumption 7. *The group-wise distribution function F_i is continuously differentiable at q_{iv} with positive derivatives $f_i(q_{iv}) > 0$ for all $i \in \{1, \ldots, k\}$, $v \in \{1, \ldots, u\}$.*

Then, an estimator for the quantile q_{iv} is given by the empirical quantile

$$\hat{q}_{iv} := \hat{F}_i^{-1}(p_v) = \inf\Big\{t \in \mathbb{R} \;\Big|\; \hat{F}_i(t) \geq p_v\Big\} = Y^{(i)}_{\lceil n_i p \rceil : n_i},$$

where $\hat{F}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbb{1}\{Y_{ij} \leq t\}$ denotes the empirical distribution function and $Y^{(i)}_{1:n_i} \leq \ldots \leq Y^{(i)}_{n_i:n_i}$ denotes the group-wise ordered random variables. For this estimator, Serfling (1980) proved a central limit theorem:

Proposition 8. *Under Assumptions 1 and 7, the following convergence in distribution holds:*

$$\sqrt{n}\,(\widehat{q}_{iv} - q_{iv})_{v \in \{1,\ldots,u\}} \xrightarrow{d} \mathbf{N}_i^q \sim \mathcal{N}\Big(\mathbf{0}, \mathbf{Q}^{(i)}\Big) \tag{2.7}$$

for all $i \in \{1, \ldots, k\}$, where the group-wise covariance matrix $\mathbf{Q}^{(i)} = \left(\mathbf{Q}^{(i)}_{vw}\right)_{v,w \in \{1,\ldots,u\}}$ is given by the entries

$$\mathbf{Q}^{(i)}_{vw} := \kappa_i^{-1} \frac{1}{f_i(q_{iv}) f_i(q_{iw})} \left(\min\{p_v, p_w\} - p_v p_w\right). \tag{2.8}$$

A proof based on empirical process theory is given in Ditzhaus et al. (2021). There, the assertion is deduced from a central limit theorem for distribution functions with the functional delta-method for empirical processes (van der Vaart & Wellner, 2000, Thm. 3.9.4). Working with empirical processes has the advantage that corresponding results for resampling procedures can be obtained more easily.

Multivariate Model To adapt this model for more than one dimension, consider mutually independent d-dimensional observation vectors $\mathbf{Y} = (\mathbf{Y}'_{11}, \ldots, \mathbf{Y}'_{kn_k})'$ for individuals from k different (sub-)groups. In detail, the jth observation vector in group i is denoted by

$$\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijd})' \sim \mathbf{F}_i, \quad i \in \{1, \ldots, k\},\, j \in \{1, \ldots, n_i\}.$$

Then, $\mathbf{F}_i$ is a multivariate distribution function. According to that, the marginal distribution functions $F_{i\ell}$ of $Y_{i1\ell}$ are assumed to be continuous with existing density function $f_{i\ell}$, $i \in \{1, \ldots, k\}$, $\ell \in \{1, \ldots, d\}$. The joint distribution function of two entries Y_{ijm} and $Y_{ij\ell}$ is denoted by $\mathbf{F}_{i\ell m}$, $i \in \{1, \ldots, k\}$, $m, \ell \in \{1, \ldots, d\}$. In a multivariate quantile-based factorial design, a vector of marginal quantiles $\mathbf{q}_i = (q_{i1}, \ldots, q_{id})'$ is considered (Babu & Rao, 1988), where

$$q_{i\ell} = F_{i\ell}^{-1}(p) = \inf\left\{t \in \mathbb{R} \mid F_{i\ell}(t) \geq p\right\}, \quad i \in \{1, \ldots, k\},\, \ell \in \{1, \ldots, d\},$$

for a pre-specified quantile level $p \in (0, 1)$. In the multivariate model, the number of quantile levels is set to one ($u = 1$), therefore, d quantiles are considered per group. In line with the consideration of the marginal distribution function in the multivariate model, the Assumption 7 has to be adapted for that:

Assumption 9. *Let the group-wise distribution function of the ℓth component $F_{i\ell}$ be continuously differentiable at $q_{i\ell}$ with positive derivative $f_{i\ell}(q_{i\ell}) > 0$ for every $\ell \in \{1, \ldots, d\}$ and $i \in \{1, \ldots, k\}$.*

Equivalently to the univariate model, the marginal quantiles are estimated via the empirical quantiles

$$\hat{q}_{i\ell} = \hat{F}_{i\ell}^{-1}(p) = \inf\Big\{t \in \mathbb{R} \;\Big|\; \hat{F}_{i\ell}(t) \geq p\Big\} = Y^{(i\ell)}_{\lceil n_i p \rceil : n_i},\; i \in \{1, \ldots, k\},\, \ell \in \{1, \ldots, d\},$$

where $Y^{(i\ell)}_{1:n_i} \leq \cdots \leq Y^{(i\ell)}_{n_i:n_i}$ are the ordered random variables of the ℓth component within group i and

$$\hat{F}_{i\ell}(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{1}\{X_{ij\ell} \leq t\}$$

is the respective marginal empirical distribution function. Additionally, there is also a central limit theorem for the marginal quantiles:

Proposition 10 (Theorem 2.1 of Babu and Rao (1988))**.** *Under Assumptions 1 and 9, there is convergence in distribution*

$$\sqrt{n}\,(\hat{q}_{i\ell} - q_{i\ell})_{\ell \in \{1,\ldots,d\}} \xrightarrow{d} \widetilde{\mathbf{N}}^q_i, \quad i \in \{1, \ldots, k\},$$

where $\widetilde{\mathbf{N}}_i^q$ *is a zero-mean, multivariate normal distributed random vector with covariance matrix* $\mathbf{G}^{(i)} = (\mathbf{G}_{\ell m}^{(i)})_{\ell,m\in\{1,\ldots,d\}}$ *given by the entries*

$$\mathbf{G}_{\ell m}^{(i)} = \begin{cases} \dfrac{1}{\kappa_i}\dfrac{1}{f_{i\ell}^2(q_{i\ell})}\left(p-p^2\right), \ \ell = m, \\ \dfrac{1}{\kappa_i}\dfrac{1}{f_{i\ell}(q_{i\ell})f_{im}(q_{im})}\left\{\mathbf{F}_{i\ell m}\left(F_{i\ell}^{-1}(p), F_{im}^{-1}(p)\right) - p^2\right\}, \ \ell \neq m. \end{cases} \tag{2.9}$$

This has been proven by Babu and Rao (1988) without the use of empirical processes. For the sake of exposition, in Baumeister et al. (2024, summarised in Section 3.1) colleagues and I present an empirical processes-based proof. Within this theory, it is possible to prove a CLT for the multivariate distribution function and then, identically to the approach in the univariate situation (Ditzhaus et al., 2021), the assertion can be deduced by the functional delta method (van der Vaart & Wellner, 2000, Thm. 3.9.4). Resampling procedures can also be obtained by empirical process theory.

2.2.2. A Semiparametric MANCOVA Model

In order to infer covariate-adjusted means, a general MANCOVA set-up is presented in the following. Consider again a vector of realisations of d-dimensional random variables $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijd})'$ representing the outcome of individual $j \in \{1, \ldots, n_i\}$ in group $i \in \{1, \ldots, k\}$. In contrast to the quantile-based models in Section 2.2.1, there is a c-dimensional individual-specific covariate vector $\mathbf{z}_{ij} = (z_{ij1}, \ldots, z_{ijc})'$ attached to each outcome vector. All outcomes and covariate vectors are pooled in the n-dimensional vector $\mathbf{Y} = (\mathbf{Y}_{11}', \ldots, \mathbf{Y}_{kn_k}')'$ and the $n \times c$ matrix $\mathbf{Z} = (\mathbf{z}_{11}, \ldots, \mathbf{z}_{kn_k})'$. Having a linear model (e.g. Stapelton, 1995) in mind, the following quantities have to be defined: a vector of n error variables $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{11}', \ldots, \boldsymbol{\epsilon}_{kn_k}')'$, $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijd})'$, a vector of k adjusted means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_k')'$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})'$ and a vector of c regression coefficients $\boldsymbol{\nu} = (\boldsymbol{\nu}_1', \ldots, \boldsymbol{\nu}_c')'$, $\boldsymbol{\nu}_a = (\nu_{a1}, \ldots, \nu_{ad})'$, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n_i\}$, $a \in \{1, \ldots, c\}$. Then the semiparametric MANCOVA (cf. Zimmermann et al., 2020) model is given by

$$\mathbf{Y} = \widetilde{\mathbf{M}}\boldsymbol{\mu} + \widetilde{\mathbf{Z}}\boldsymbol{\nu} + \boldsymbol{\epsilon}, \tag{2.10}$$

where $\widetilde{\mathbf{M}} = \bigoplus_{i=1}^k (\mathbf{1}_{n_i} \otimes \mathbf{I}_d)$ and $\widetilde{\mathbf{Z}} = \mathbf{Z} \otimes \mathbf{I}_d$. Thereby, $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{M}}, \widetilde{\mathbf{Z}})$ is the *design matrix* of a linear model, where the matrix $\widetilde{\mathbf{M}}$ characterises the factorial part and matrix $\widetilde{\mathbf{Z}}$ the regression part. The covariance of random vector $\boldsymbol{\epsilon}$ is given by $\mathbf{S} = \mathrm{Cov}(\boldsymbol{\epsilon}) = \bigoplus_{i=1}^k (\mathbf{I}_{n_i} \otimes \mathbf{S}_i)$. In the model equation (2.10), the regression coefficients in the vector $\boldsymbol{\nu}$ depend on the dimension $\ell \in \{1, \ldots, d\}$ but not on the group $i \in \{1, \ldots, k\}$. Therefore, the model does not allow unequal regression coefficients for different groups. Allowing for unequal regression coefficients leads

to non-interpretable coefficients as the magnitude of the treatment effect is not the same at different levels of the matrix of covariates $\mathbf{Z}$ (Huitema, 2011). The following assumptions represent the semiparametric MANCOVA model and are similarly assumed in Zimmermann et al. (2020).

Assumption 11. *Let the following statements hold:*

(M1) *The errors* $\boldsymbol{\epsilon}_{ij}$ *are independent and identically distributed in every group* i *with* $\mathrm{E}(\boldsymbol{\epsilon}_{ij}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\epsilon}_{ij}) = \mathbf{S}_i$, *and* $\mathrm{E}(\|\boldsymbol{\epsilon}_{ij}\|^4) < \infty$ *for all* $i \in \{1, \ldots, k\}$ *and* $j \in \{1, \ldots, n_i\}$.

(M2) *The marginal variances are positive, i.e.* $\sigma^2_{i\ell} := \mathrm{Var}(\epsilon_{ij\ell}) > 0$ *for all* $i \in \{1, \ldots, k\}$ *and* $\ell \in \{1, \ldots, d\}$.

(M3) *The matrix of covariates* $\mathbf{Z}$ *has full column rank, i.e. the columns of* $\mathbf{Z}$ *are linearly independent of each other, and they should be independent of the columns of* $\bigoplus_{i=1}^{k} \mathbf{1}_{n_i}$.

(M4) $1/n_i \sum_{j=1}^{n_i} z_{ija} \to \boldsymbol{\gamma}_{ia} \in \mathbb{R}$ *for all* $i \in \{1, \ldots, k\}$ *and* $a \in \{1, \ldots, c\}$.

(M5) $1/n_i \sum_{j=1}^{n_i} \mathbf{z}_{ij}\mathbf{z}'_{ij} \to \Gamma_i \in \mathbb{R}^{c \times c}$ *for all* $i \in \{1, \ldots, k\}$.

From (M1) it follows that the observations in $\mathbf{Y}$ are assumed to be independent and identically distributed per group, but no specific distribution class (such as normality) is assumed. In particular, the distributions can differ between groups and components. The singularity of the covariance matrix $\mathbf{S}$ is allowed through (M2). Assumption (M3) avoids collinearity and (M4) and (M5) have other technical reasons.

As suggested in Zimmermann et al. (2020), the vector $\boldsymbol{\mu}$ can be estimated by the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}'_1, \ldots, \hat{\boldsymbol{\mu}}'_k)'$, that is

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{Y}}_{i\cdot} - (\bar{\mathbf{z}}_{i\cdot} \otimes \mathbf{1}'_d)\, \hat{\boldsymbol{\nu}}. \tag{2.11}$$

The vector $\hat{\boldsymbol{\nu}} = (\hat{\boldsymbol{\nu}}'_1, \ldots, \hat{\boldsymbol{\nu}}'_c)'$ is the OLS estimator of $\boldsymbol{\nu}$, where $\hat{\boldsymbol{\nu}}_a = (\hat{\nu}_{a1}, \ldots, \hat{\nu}_{ad})'$ for every $a \in \{1, \ldots, c\}$. To define the estimator $\hat{\boldsymbol{\nu}}$ in matrix notation, the matrices $\mathbf{M} = \bigoplus_{i=1}^{k} \mathbf{1}_{n_i}$, $\mathbf{P}_M := \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'$ and $\mathbf{W} := (\mathbf{I}_n - \mathbf{P}_M)\mathbf{Z}$ are used. Consequently, the estimator is given by $\hat{\boldsymbol{\nu}} := [(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' \otimes \mathbf{I}_d]\mathbf{Y}$, where $(\mathbf{I}_n - \mathbf{P}_M)$ adjusts the covariates $\mathbf{Z}$ in such a way that they are correctly multiplied with the factorial part and the multivariate OLS estimator is calculated by $\mathbf{W}$. The connection with the traditional formulation in linear models becomes clear, if one realizes that the OLS estimator for $\boldsymbol{\beta} = (\boldsymbol{\mu}', \boldsymbol{\nu}')'$ may also be written as

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\nu}}')' = \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\mathbf{Y}.$$

Nevertheless, statistical inference of $\boldsymbol{\mu}$ is of further interest. In Zimmermann et al. (2020), the following central limit theorem for the OLS estimator $\hat{\boldsymbol{\beta}}$ is stated and proven:

Proposition 12. *Let* $\boldsymbol{\beta} = (\boldsymbol{\mu}', \boldsymbol{\nu}')'$ *and* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\nu}}')'$. *Then, under (M1), (M3), (M4), (M5), and Assumption 1 it holds*

$$\sqrt{n}\Big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\Big) \xrightarrow{d} \mathbf{N}^{\mu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}),$$

where $\boldsymbol{\Lambda} := \lim_{n\to\infty} n\Big(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\Big)^{-1}\tilde{\mathbf{X}}'\mathbf{S}\tilde{\mathbf{X}}\Big(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\Big)^{-1}$.

3. Summary of the Articles

It's time now to sing out
Though the story never ends
Let's celebrate, remember a year
In the life of friends

Seasons of Love, Larson (1996a)

3.1. Quantile-based MANOVA: A New Tool for Inferring Multivariate Data in Factorial Designs

The aim of the Paper Baumeister et al. (2024) has been to introduce a quantile-based alternative to infer general multivariate factorial designs. The recent trend to more general and flexible MANOVA methods, as mentioned in Section 2.1.3, is extended to multivariate quantiles as estimands. Considering quantiles in statistical analyses, in particular considering the median, poses a robust alternative to mean-based analyses in many situations. Especially for heavy-tailed distributions and in case of outliers, the median is the preferred statistical estimand (Maronna et al., 2006). Bonett and Price (2002) pointed out: "Every student of introductory statistics is taught that the population median may be more meaningful than the population mean when the distribution is skewed." This has been a motivation for the development of quantile-based inference, furthermore, it is a known fact that inference on quantiles leads to more powerful test decisions in context of heavy-tailed data (cf. Ditzhaus et al., 2021). The method has been realized in the multivariate model stated in Section 2.2.1; consequently, it does not assume normality and allows for potential heterogeneity. As this model assumes the existence of differentiable densities, it is appropriate in situations, where metric data can be assumed.

In fact, the method applies the theory of quadratic from-type statistics (Section 2.1.3) on Proposition 10 and the multivariate quantile-based model in Section 2.2.1 in terms of the multivariate quantiles $\sqrt{n}(\hat{\mathbf{q}} - \mathbf{q})$. For this purpose, an appropriate estimator for the unknown limiting covariance matrix $\mathbf{G} := \bigoplus_{i=1}^{k} \mathbf{G}^{(i)}$ is needed. Three approaches that were introduced in Ditzhaus et al. (2021) have been adapted for the consistent estimation of $\mathbf{G}$: a kernel-based approach by Nadaraya (1965), a bootstrap approach by Chung and Romano (2013) using classic Efron's bootstrap (Efron, 1979) and an interval-based approach by Price and Bonett (2001). For the

sake of simplicity, $\hat{\mathbf{G}}$ denotes any consistent estimator among the three mentioned options. As explained at the beginning of Section 2.1 the resulting QMANOVA is applicable for all kind of factorial designs, where the model follows the common practice to reformulate the Hypothesis (2.2) with the unique projection matrix $\mathbf{T} = \mathbf{H}'(\mathbf{H}\mathbf{H}')^{+}\mathbf{H} \in \mathbb{R}^{dk\times dk}$ to $\mathcal{H}_0 : \mathbf{T}\mathbf{q} = \mathbf{0}$. The limiting unknown covariance matrix $\mathbf{G}$ is not necessarily non-singular. Therefore, it is in general not possible to consider the WTS in this framework. Instead, the ATS and MATS are proposed as instances as versions of the test statistic in (2.6):

$$\mathbf{ATS}_n(\mathbf{T}) = n\frac{(\mathbf{T}\hat{\mathbf{q}})'\mathbf{T}\hat{\mathbf{q}}}{\operatorname{tr}\left(\mathbf{T}\hat{\mathbf{G}}\mathbf{T}\right)}, \quad \mathbf{MATS}_n(\mathbf{T}) = n(\mathbf{T}\hat{\mathbf{q}})'\left(\mathbf{T}\hat{\mathbf{G}}_0\mathbf{T}\right)^{+}\mathbf{T}\hat{\mathbf{q}}.$$

Since it can be argued that the respective functions $\Xi\left(\mathbf{T}, \hat{\mathbf{G}}\right)$ for ATS and MATS are consistent for $\Xi(\mathbf{T}, \mathbf{G})$, Theorem 6 is applicable and bootstrapping has to be used to approximate the unknown limiting distribution of ATS and MATS as explained in Section 2.1.2. Here, a group-wise nonparametric bootstrap approach is considered. A d-dimensional bootstrap sample $\{\mathbf{Y}^*_{i1}, \ldots, \mathbf{Y}^*_{in_i}\}$ is drawn with replacement from the original observation vectors $\{\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in_i}\}$ for every group $i \in \{1, \ldots, k\}$. The (conditional) asymptotic behaviour of the bootstrap counterparts of the test statistics $\mathbf{ATS}^*_n(\mathbf{T})$ and $\mathbf{MATS}^*_n(\mathbf{T})$, is investigated by proving a bootstrap equivalent of Proposition 10 by empirical processes and combine this result with further bootstrap techniques of van der Vaart and Wellner (2000). From this, it can be concluded that, under $\mathcal{H}_0$, given the data, the bootstrap test statistics $\mathbf{ATS}^*_n(\mathbf{T})$ and $\mathbf{MATS}^*_n(\mathbf{T})$ have asymptotically the same distribution as the test statistics $\mathbf{ATS}_n(\mathbf{T})$ and $\mathbf{MATS}_n(\mathbf{T})$, respectively. The proposed resampling tests use the empirical $(1-\alpha)$-quantile $b^*_\alpha(\mathbf{Y})$ of the conditional distribution function $y \mapsto P(U^*_n(\mathbf{T}) \leq y \,|\, \mathbf{Y})$ as critical values. This leads to the tests $\Phi^*_n = \mathbb{1}\{U_n(\mathbf{T}) > b^*_\alpha(\mathbf{Y})\}$. As explained in Section 2.1.3, it can be argued that Φ^*_n is an asymptotic level-α test, which is consistent for the alternative.

The theoretical results of the QMANOVA are complemented by an extensive simulation study for small and moderate sample sizes, which also compares mean-based methods with the new median-based approach. In the article, the simulation is focussed on data that fulfils the null hypothesis of a median-based one-way layout. Further simulations results are presented in the supplementary material, which can be found online. From the simulations, it is apparent that the bootstrap covariance estimator has the best performance regarding the type I error control. With the other covariance estimators, ATS and MATS show a quite conservative type I error control. As both considered test statistics have a similar type I error control, it can be recommended to use the ATS and MATS combined with the bootstrap covariance estimator. These two favourable QMANOVA methods, have been compared with the mean-based MATS of Friedrich and Pauly (2018) with two different bootstrap versions. This comparison is only possible on symmetric data, such that the mean-based

hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and the median-based hypothesis $\mathcal{H}_0 : \mathbf{m}_1 = \mathbf{m}_2$ coincide. From this, it can be observed that the mean- and median-based methods have a comparable type I error control. For the median-based methods, the ATS tends to a more conservative behaviour than the MATS.

In context of power the methods of the QMANOVA can have one advantage. In the power simulations, it can be observed that the median-based tests are more powerful in simulations scenarios, where t_2-distributed data is considered, while the exact opposite occurs with the normal distribution. This may happen, because the sample median is the better location estimator in case of heavy-tailed data like the t_2-distribution, but for normally distributed data the situation is reversed. This observation is in line with the power simulation results in Ditzhaus et al. (2021) and leads to the conclusion that median-based estimation and inference is recommended in context of heavy-tailed data. This was also made clear by the illustrative data example, where a multivariate data set about Egyptian skulls in a three-way MANOVA setting is analysed. It contains four measures of the skulls that characterise their basic shape with considerably heavy-tailed data. Consequently, the median-based test is able to detect differences in this data set, while the mean-based based methods does not, which exemplifies also the high applicability of quantiles in skewed data.

3.2. Early and Late Buzzards: Comparing Different Approaches for Quantile-based Multiple Testing in Heavy-Tailed Wildlife Research Data

In the Paper Baumeister, Munko, et al. (2025a), different possibilities to infer quantile-based multiple testing procedures in general factorial designs have been compared.To incorporate this, the univariate model for more than one quantile from Section 2.2.1 has been considered. This model explicitly includes inference regarding the median, the IQR, or both together. The motivation for a comparison on methods for quantile-based multiple testing has been that in medical, ecological and psychological research, multiple testing problems occur as often as heavy-tailed and skewed data. To illustrate the advantages of quantiles as estimands, colleagues and I have considered a real data example of wild life animals, where heavy-tails can be observed. As already explained in the summary of the QMANOVA (Section 3.1), for this kind of data, the median is the preferred measure of location and the IQR is an adequate alternative to the variance as a measure of dispersion.The aim of the data example is to identify years between 2006 and 2022 with an earlier and a later hatching phenology, which is important to compare weather conditions and population characteristics between years with early and late breedings. In context of increasing temperatures and

extreme weather events due to climate change, it is of general interest to understand these connections. In the end, this ecological question leads to a non-inferiority multiple testing procedure. That is why, additionally, non-inferiority testing problems have been considered in the Monte Carlo simulations.

To be able to compare different methods reasonably well, the hypotheses in (2.3) and the simultaneous testing problems from Section 2.1.4 are defined in the univariate quantile-based model in Section 2.2.1. The two-sided quantile-based multiple testing problem can be formulated as follows:

$$\mathcal{H}_{0,s} : \mathbf{h}_s'\mathbf{q} = \epsilon_s \text{ vs. } \mathcal{H}_{1,s} : \mathbf{h}_s'\mathbf{q} \neq \epsilon_s, \quad s \in \{1, \ldots, r\}. \tag{3.1}$$

A one-sided non-inferiority multiple testing problem can be defined as follows:

$$\mathcal{H}^I_{0,s} : \mathbf{h}_s'\mathbf{q} \leq \epsilon_s \text{ vs. } \mathcal{H}^I_{1,s} : \mathbf{h}_s'\mathbf{q} > \epsilon_s, \quad s \in \{1, \ldots, r\}. \tag{3.2}$$

Then, the corresponding global hypotheses are $\mathcal{H}_0 : \mathbf{Hq} = \boldsymbol{\epsilon}$ and $\mathcal{H}^I_0 : \mathbf{Hq} \leq \boldsymbol{\epsilon}$, respectively. The motivation to consider both types of hypotheses has been that they have widely different interpretations despite the similar methodology. Furthermore, possibilities to infer equivalence hypotheses (Hauck & Anderson, 1984) for quantiles are stated but not discussed in detail. To infer the testing problem in (3.1), the testing procedure

$$\varphi^q_n = \max_{s\in\{1,\ldots,r\}} \mathbb{1}\left\{\frac{|A^q_n(\mathbf{h}_s, e_s)|}{w^q_s(\alpha)} > 1\right\} \tag{3.3}$$

is used for appropriate critical values $w^q_s(\alpha)$, where the one-sided version (3.2) uses another appropriate critical value $w^{qI}_s(\alpha)$ in the test decisions:

$$\varphi^{qI}_n = \max_{s\in\{1,\ldots,r\}} \mathbb{1}\left\{\frac{A^q_n(\mathbf{h}_s, e_s)}{w^{qI}_s(\alpha)} > 1\right\}. \tag{3.4}$$

The in the simulation study compared methods are based on Bonferroni-corrections with an asymptotic and a permutation critical value as well as an MCTP approach with bootstrap and asymptotic critical values. As stated in Theorem 3, it can be argued, that the vector of quantile-based t-type test statistics $A^q_n(\mathbf{H}, \mathbf{e})$ is asymptotically multivariate normally distributed for a consistent estimator $\hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$ and a single test statistic $A^q_n(\mathbf{h}_s, e_s)$ is standard normally distributed for every $s \in \{1, \ldots, r\}$. Then, the $(1-\alpha/(2r))$-quantile of the standard normal distribution $w^q_s(\alpha) = z_{1-\alpha/(2r)}$ for the two-sided multiple testing problem or $w^{qI}_s(\alpha) = z_{1-\alpha/r}$ for the non-inferiority multiple testing problem, is the suitable critical value for the asymptotic Bonferroni-adjusted procedure. To yield a potentially better small sample performance the permutation approach considered in Ditzhaus et al. (2021) is adapted for t-type test statistics.

Here, the idea is to draw the permuted samples $Y_{i1}^{\pi}, \ldots, Y_{in_i}^{\pi}$, $i \in \{1, \ldots, k\}$, without replacement from the pooled sample $Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{kn_k}$. The permutation Bonferroni-adjusted approach is derived by using permutation-based critical values instead of the standard normal quantiles. The theory behind this is adapted from the permutation approach in Ditzhaus et al. (2021) and is based on empirical process theory, similar to the theory of the QMANOVA (Section 3.1). The MCTP approach is an extension of the method in Segbehoe et al. (2022), which introduced MCTPs for one quantile. In the article, we extend the method to be feasible for more than one quantile of interest and to one-sided testing problems. For the asymptotic approach, the main idea is to consider the asymptotic multivariate distribution of the test statistics in the way that is explained in Section 2.1.4. For the limiting random variables $\mathbf{A}^q$ of $A_n^q(\mathbf{H}, \mathbf{e})$, it holds $\mathbf{A}^q = (\mathbf{A}_1^q, \ldots, \mathbf{A}_r^q)' \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^q)$ with

$$\mathbf{R}^q = (\mathbf{HQH}')_0^{-\frac{1}{2}} \mathbf{HQH}' (\mathbf{HQH}')_0^{-\frac{1}{2}}$$

and $\mathbf{Q} := \bigoplus_{i=1}^k \mathbf{Q}^{(i)}$. The limiting covariance matrix $\mathbf{Q}$ can be replaced by an appropriate consistent estimator $\hat{\mathbf{Q}}$ and the equicoordinate $(1-\alpha)$-quantiles $q_{1-\alpha}$ and $q_{1-\alpha}^I$ of the resulting distributions limiting $|\mathbf{A}^q|$ and $\mathbf{A}^q$ are used as critical values. Equivalently to the QMANOVA (Baumeister et al., 2024, summarised in Section 3.1), a group wise nonparametric bootstrap is considered for a better small sample performance. As explained above, a nonparametric bootstrap sample $Y_{i1}^*, \ldots, Y_{in_i}^*$ is drawn with replacement from the original i-th sample $Y_{i1}, \ldots, Y_{in_i}$. From this, bootstrap test statistics $A_n^{q*}(\mathbf{H}, \mathbf{e})$ are calculated and $q_{1-\alpha}^*$ as well as $q_{1-\alpha}^{I*}$, the empirical equicoordinate $(1-\alpha)$-quantiles of $|A_n^{q*}(\mathbf{H}, \mathbf{e})|$ respective $A_n^{q*}(\mathbf{H}, \mathbf{e})$, are used as critical values. All approaches have been simulated combined with the three consistent covariance estimators (kernel, interval, and bootstrap-based), which were also used in the QMANOVA in Section 3.1 and in the QANOVA by Ditzhaus et al. (2021).

To investigate, which method is appropriate to answer the multiple testing problems (3.3) and (3.4), the explained methods formulated in terms of medians or IQRs are compared in an extensive simulation study. Details of the simulation study, especially additionally plots and tables as well as the R-scripts can be found online at TUDOdata (Baumeister, Munko, et al., 2025b). One part of the simulation study investigates the tests' performances on small sample sizes (15 per group on average) and $k = 4$ groups. Here, it can be observed that the FWER-control highly depends on the choice of the covariance estimator and on the testing problem for all methods excepted the Bonferroni-adjusted permutation approach. While the latter approach has a stable FWER-control in this scenarios, the Bonferroni-adjusted asymptotic testing procedures tend to be too conservative and the two MCTP appraches show a conservative or a liberal behaviour depending on the tetsing problem and the covariance estimator. This is surprising in some way, as the Bonferroni correction is known for a conservative behaviour and the FWER-control of MCTPs is usually better. As expected, the Bonferroni-corrected approaches show a conservative behaviour in

the simulation study, where $r = 17$ tests and larger sample sizes are considered. Here, the data was generated similarly to the data example. In this simulation scenarios, the bootstrap MCTP with bootstrap covariance estimator shows the most stable FWER-control. The power is in general very similar for all methods, if the significance level under the null hypothesis is maintained. The paper concludes with a data analysis of the motivating example. As there are many tests and moderate sample sizes considered, the bootstrap MCTP with bootstrap covariance estimator is, based on the simulation results, the recommended method to infer the testing problem.

3.3. Multivariate and Multiple Contrast Testing in General Covariate-adjusted Factorial Designs

In the Paper Baumeister, Thiel, Matits, Zimmermann, et al. (2025), the concept of MCTPs (Section 2.1.4) has been applied to the semiparametric MANCOVA model in 2.2.2. Especially for multivariate outcomes, there is a need of statistical methods that can deal with multiple testing and covariate adjustment, which is underscored by a data example, where all three issues occur together. A synthetic dataset (Thiel et al., 2025) based on original data from the intervention-based *HypnoTreat* study (Karrasch, Matits, et al., 2023; Karrasch, Mavioğlu, et al., 2023; Karrasch et al., 2022) poses this example. The study examined chronically stressed individuals and analysed the effects of a single relaxation hypnosis session on heart rate variability, which is a set of stress related physiological variables measured on different scales. The intervention-induced changes in the heart rate variability are observed in 5 variables measuring the same physiological construct. Consequently, an overall global effect and specific local effects are of interest, which requires testing of multiple hypotheses. Furthermore, joint modelling of these outcomes may be beneficial for statistical analyses, as the outcomes are correlated. This requires a multivariate approach. Additionally, the confounding covariates chronic stress and suggestibility (ability to be hypnotised) should be considered to explain some of the variability in the data. In the end, this leads to multivariate multiple testing on covariate adjusted means. For $k = 2$ groups and $d = 5$ dimensions, the adjusted means $\mu_{i,\ell}$ for every endpoint $\ell \in \{1, \ldots, 5\}$ can be compared: $\mathcal{H}_{0,\ell} : \mu_{1\ell} = \mu_{2,\ell}$.

This example illustrates that it is beneficial to model more than one covariate-adjusted endpoint together while having the possibility to test specific problems therein. The use of MCTPs seems to be appropriate to infer these testing problems as they are a appropriate simultaneous multiple testing procedure, which was also a result of Baumeister, Munko, et al. (2025a, summarised in Section 3.2). In the end, the aim of this article was to implement powerful multivariate multiple contrast

testing procedures on covariate-adjusted means that can deal with various multiple testing problems, see Section 2.1.

To reach this goal, the semiparametric MANCOVA model explained in Section 2.2.2 has been considered. As stated in Section 2.1.3, Zimmermann et al. (2020) developed global hypothesis tests in this model by using quadratic form-type statistics. In fact, the implementation of MCTPs in this model extends their method to the multiple testing problem

$$\Omega = \Big\{\mathcal{H}_{0,s} : \mathbf{h}_s'\boldsymbol{\mu} = \mathbf{0} \;\Big|\; s \in \{1,\dots,r\}\Big\}, \tag{3.5}$$

where $\boldsymbol{\mu}$ denotes the adjusted mean defined in Section 2.2.2. The family Ω corresponds to the global hypothesis $\mathcal{H}_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, which is a version of the null hypothesis in (2.2) on covariate-adjusted means. As a consequence of the model assumptions, the matrix with the diagonal elements of the limiting covariance matrix $\boldsymbol{\Lambda}_0$ and its estimator $\hat{\boldsymbol{\Lambda}}_0$ are used as the matrices $\mathbf{D}$ and $\hat{\mathbf{D}}$ in Section 2.1.1. Additionally, as the semiparametric MANCOVA model allows for heteroscedasticity, a heteroscedasticity-consistent adjustment is applied on the squared residuals $\mathbf{S}$ (see Section 2.2.2). The resulting covariance estimator for the upper left block matrix $\boldsymbol{\Lambda}_{11,0}$ of $\boldsymbol{\Lambda}_0$ is named as $\hat{\boldsymbol{\Lambda}}_{11,0}$. In contrast to the quantile-based MCTPs in Section 3.2, the test statistics (see Equation 2.6)

$$A_n^{\mu}(\mathbf{h}_s) = \sqrt{n}\frac{\mathbf{h}_s'\hat{\boldsymbol{\mu}}}{\sqrt{\mathbf{h}_s'\hat{\boldsymbol{\Lambda}}_{11,0}\mathbf{h}_s}},$$

are not standard normally distributed for every $s \in \{1,\dots,r\}$. In fact, their limiting distribution is

$$A_s^{\mu} \sim \mathcal{N}\left(0, \frac{\mathbf{h}_s'\boldsymbol{\Lambda}_{11}\mathbf{h}_s}{\mathbf{h}_s'\boldsymbol{\Lambda}_{11,0}\mathbf{h}_s}\right).$$

Therefore, it is not possible to apply equicoordinate quantiles of the limiting distribution as critical values directly. To overcome this problem, colleagues and I have considered resampling techniques and a generalized determination of local levels. A parametric and a wild bootstrap were proven and theoretically founded in Zimmermann et al. (2020) and are also suitable for the use in these methods. For ease of notation, both shall be described by $^\circ$ here.

The idea of the generalised determination of the local levels is as follows. The vector of test statistics, calculated on the bth bootstrap sample, $b \in \{1,\dots,B\}$, is named by $\left(A_n^{\mu,\circ,b}(\mathbf{h}_1),\dots,A_n^{\mu,\circ,b}(\mathbf{h}_r)\right)$ and the $(1-\gamma)$-quantile of $|A_n^{\mu,\circ,1}(\mathbf{h}_s)|, \dots, |A_n^{\mu,\circ,B}(\mathbf{h}_s)|$ by $q_{s,1-\gamma}^{\circ}$. Munko et al. (2024) introduced multiple contrast test procedures for the RMST by using a WTS and, based on ideas by Bühlmann (1998), they adjust the

significance level γ for each local test such that the level α is controlled globally. To realise this, for $\gamma \in [0,1]$ the estimated FWER has to be defined:

$$\mathrm{FWER}_n^\circ(\gamma) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{\exists\, s \in \{1,\dots,r\} : |A_n^{\mu,\circ,b}(\mathbf{h}_s)| > q_{s,1-\gamma}^\circ\right\}.$$

Then, the adjusted level $\gamma_n(\alpha)$ can be defined as:

$$\gamma_n(\alpha) := \max\left\{\gamma \in \left\{0, \frac{1}{B}, \dots, \frac{B-1}{B}\right\} \;\middle|\; \mathrm{FWER}_n^\circ(\gamma) \leq \alpha\right\}. \tag{3.6}$$

To give an intuition, $\gamma_n(\alpha)$ is chosen as the largest value such that $\mathrm{FWER}_n^\circ(\gamma)$ is bounded by the global level of significance α. The adjusted level $\gamma_n(\alpha)$ allows to define critical values for an appropriate testing procedure: for every $s \in \{1,\dots,r\}$, the testing problem in (3.5) can be inferred by

$$\varphi_{n,s}^{\mu,\circ} = \mathbb{1}\left\{|A_n^\mu(\mathbf{h}_s)| > q_{s,1-\gamma_n(\alpha)}^\circ\right\},$$

and in line with the testing procedure defined in Section 2.1.4 the global hypothesis $\mathcal{H}_0$ is rejected, if and only if at least one $\mathcal{H}_{0,s}$ is rejected:

$$\varphi_n^{\mu,\circ} = \max_{s\in\{1,\dots,r\}} \mathbb{1}\left\{\frac{|A_n^\mu(\mathbf{h}_s)|}{q_{s,1-\gamma_n(\alpha)}^\circ} > 1\right\}.$$

This formulation incorporates the different distributions of $A_n^\mu(\mathbf{h}_s)$, $s \in \{1,\dots,r\}$, by considering individual quantiles $q_{s,1-\gamma_n(\alpha)}^\circ$ as critical values, but uses the same level of significance $\gamma_n(\alpha)$ for all tests. In the paper, colleagues and I have shown that this testing procedure controls asymptomatically the family-wise type I error rate of Ω. Moreover, a theoretically valid calculation of local and global p-values that can be compared with the adjusted levels $\gamma_n(\alpha)$ respective the global level α is presented. It is shown that the tests $\varphi_{n,s}^{\mu,\circ}$ and $\varphi_n^{\mu,\circ}$ are consistent asymptotic level-α tests.

The paper includes an extensive simulation study regarding type I error and power, were $\varphi_{n,s}^{\mu,\circ}$ and $\varphi_n^{\mu,\circ}$ are compared with the Bonferroni-corrected tests of Zimmermann et al. (2020) and an asymptotic MCTP, which considers $\mathbf{D} = \mathbf{\Sigma}$ and requires a stronger assumption as (M2) in Assumption 11. Further details of the simulation study, especially additional plots and the results of further simulation results as well as the R-scripts can be found online at TUDOdata (Baumeister, Thiel, Matits, Pauly, et al., 2025). To get a broad overview about the performance of the testing procedures, small and moderate, balanced and unbalanced sample sizes as well as various distributions, covariance structures, and singular data is considered. The latter is considered to take the advantage of the semiparametric MANCOVA model into account. In the simulation study, the Bonferroni-corrected tests show the often

observed conservative behaviour, especially for testing problems with a large number of individual tests. The tests $\varphi_{n,s}^{\mu,\circ}$ and $\varphi_{n}^{\mu,\circ}$ show an overall good small sample performance and an FWER-control even on singular data, were the asymptotic MCTPs cannot be applied. In settings with negative pairing, a highly liberal behaviour occurs across all methods, which is an already described problem in mean-based semiparametric models (e,g. Konietschke et al., 2015). The power simulations include settings under shift, one-point, and trend alternatives. In each of these settings, the behaviour under the alternative is relatively similar for the compared methods that maintain the level of significance under the null hypothesis. It can be observed that the power in the singular settings is lower than for other settings, but $\varphi_{n,s}^{\mu,\circ}$ and $\varphi_{n}^{\mu,\circ}$ have an improved power in comparison to the other methods.

As an illustration of the developed methods, the synthetic *HypnoTreatSynth* data set (Thiel et al., 2025) is analysed by the methods that were compared in the simulation study. Here, the advantage of the simultaneous testing procedure of producing coherent and consonant local and global testing decisions is illustrated. Despite the fact that most methods are able to detect the added effect in the synthetic dataset, together with the simulation study it becomes clear that the developed resampling MCTPs are most reliable in the considered semi-parametric model.

4. Concluding Discussion and Outlook

Remember the love
Oh, you got to, you got to remember the love
Remember the love
You know that love is a gift from up above
Remember the love
Share love, give love, spread love
Measure in love
Measure, measure your life in love
Seasons of love

Seasons of Love, Larson (1996a)

4.1. Discussion

This dissertation includes three contributions to methodology for inference in general factorial designs. All contributions are motivated by the observation that there is a need of more generalised testing procedures in quantitative sciences. That is why none of the presented methods contain assumptions of normality or homoscedastic covariance structures and extends general concepts of single and multiple testing to more complex relevant testing problems. The developed methods have been proven through asymptotic statistics and empirical process theory and consider a resampling-based critical value, which substantially improves the general applicability of this methods. Furthermore, all methods have been examined in their behaviour under the null hypothesis and under the alternative and have been compared with existing methods by Monte Carlo simulations. Additionally, the applicability of the new methods is illustrated by data examples.

In the first article Baumeister et al. (2024) colleagues and I have implemented a quadratic form-type statistic to infer multivariate quantiles. This has made the use of resampling necessary. Resampling procedures for a broad class of estimands can be proven by empirical process theory and have turned out to be successful in that multivariate quantile-based framework. The availability of robust quantile-based testing procedures may increases the use of quantile-based analyses in medicine, psychology and biology, where skewed and heavy-tailed data occurs frequently, especially in experimental studies. In the included simulation study, it turned out that the introduced method is a robust and on heavy-tailed data a more powerful alternative to classical mean-based MANOVA. Therefore quantile-based testing can lead to more powerful testing procedures on an easy interpretable estimand.

That quantiles define highly applicable estimands has also been a motivation for the systematic comparison of quantile-based multiple testing procedures in the second article Baumeister, Munko, et al. (2025a). There, the well established concepts of Bonferroni-correction and MCTPs have been adapted and compared in a general quantile-based framework. The extension of the quantile-based ANOVA (Ditzhaus et al., 2021) to multiple testing also makes quantile-based testing more applicable in quantitative sciences and the existing method of Segbehoe et al. (2022) is checked for its applicability in certain scenarios. The simulation study has shown that the behaviour of the compared methods depends highly on the sample size of the simulated data. On small samples with a few compared tests, it turned out that MCTPs are not superior to the Bonferroni correction, while the situation is vice versa for bigger sample sizes and more tests. That the methodology of MCTPs is not applicable on smaller sample sizes in context of quantiles has been surprising, because this phenomenon was not observable on MCTPs with other estimands. Based in this observations, colleagues and I were able to give some recommendations for the use of quantile-based multiple testing procedures of certain scenarios. This can provide orientation for quantitative scientists while choosing the right statistical method for their research.

In the third paper, colleagues and I have implemented the concept of MCTPs in a semiparametric MANCOVA, which extends the methods of Zimmermann et al. (2020) to multiple testing problems. A motivating example has shown the practical relevance of the method as intervention studies can be modelled all-encompassing in a multivariate model that allows coherent and consonant multiple testing. Additionally, having valid methods at hand that allow for inclusion of covariates is very important in many fields of research. In general, the paper highlights the practical advantages of MCTPs to infer multiple endpoints simultaneously. To realise that, colleagues and I have implemented a generalized calculation of critical values inspired by Munko et al. (2024), which was necessary because we have considered a general model that allows certain types of singularity. This made also the consideration of resampling procedures necessary. As the article Baumeister, Munko, et al. (2025a) has shown that a systematic comparison with other simultaneous testing procedures is important to assess the methods performance, we have compared the new approaches with the Bonferroni-adjusted methods of (Zimmermann et al., 2020) and naive asymptotic MCTPs. The simulation study shows a good level-α control of the proposed method even on small samples and on singular data. The behaviour under the alternative is very similar for all methods that hold the level of significance in the respective scenario, this is why the MCTPs seems to be a suitable approach.

4.2. Outlook

Even though the three articles included in this dissertation have somewhat expanded the existence of factorial designs, there are many more gaps to fill in the world of factorial designs. These gaps are not exclusively interesting from a methodological point of view; on the contrary, there are still many research questions involving factorial designs and testing problems for which there is no methodological solution. In particular, solutions are missing if the statistical analysis has to deal with non normal, skewed, and heteroscedastic data at the same time.

To give some examples, it could be of interest to investigate how some multivariate extension of the considered multiple testing procedures in Baumeister, Munko, et al. (2025a) perform. The development of covariate-adjustment on quantiles could be a further extension of the world of quantile-based inference. This is of potential interest in uni- and multivariate models as well as in context of multiple testing.

To ensure that people can actually use the methods included in this thesis, R implementations are necessary. Together with a colleague, an R implementation of the methods to infer the semiparametric MANCOVA model of Zimmermann et al. (2020) including the MCTPs therein is in preparation. An R implementation of the quantile-based methods is also of interest.

In order to make the developed statistical methods truly applicable for scientists in the quantitative sciences, it is also necessary to investigate in further competitive simulation studies how the methods behave in realistic situations. Heinze et al. (2024) developed a model of four phases that a statistical method must have passed through before it is really usable in quantitative sciences. Analysing the behaviour of a method in competitive simulation studies plays an important role therein. Such an analysis as Baumeister, Munko, et al. (2025a) produced some surprising results that would not have been uncovered without the comparison made there. It may therefore be interesting to carry out further systematic simulation-based comparisons, particularly in the context of multiple testing problems. There is also a lack of a paper that systematically compares the performance of established current methods in the field of multiple testing. One comparative overview of concepts for multiple testing was given in Pigeot (2000), but does not include simulations. Konietschke et al. (2013) focus on the differences in methodology between MCTPs and ANOVA. To reach that, they explain the methodologically differences and compare the global power on theoretically aspects and in simulations, but they do not systematically compare different version of max-t tests.

To come to another area of testing problems, it would also be interesting to investigate the performance of different MANOVA methods in such a comparative simulation study similarly to the comparison on univariate two-sample problems that was done in Noguchi et al. (2021). This is motivated by the observation that there are lots of robust alternatives to the classical MANOVA methods (e.g. Anderson, 2001; Dobler et al., 2020; Friedrich & Pauly, 2018; Konietschke et al., 2015), but it seems that they are only used with hesitation. Reasons for this could be a low popularity of the new methods or limited knowledge about the benefits of them. A discussion of the method's benefits based on comparative simulations could resolve this matter.

Part II.

Publications and Preprints

In diapers, report cards
In spoke wheels, in speeding tickets
In contracts, dollars
In funerals, in births

In five hundred twenty-five thousand six hundred minutes
How do you figure our last year on earth?

Figure in love
Measure in love
Seasons of love

Seasons of Love B, Larson (1996b)

Article 1

Baumeister, M., Ditzhaus, M., & Pauly, M. (2024).

Quantile-based MANOVA:
A new tool for inferring multivariate data in factorial designs.

Journal of Multivariate Analysis, 199, 105246.
https://doi.org/10.1016/j.jmva.2023.105246

Journal of Multivariate Analysis 199 (2024) 105246

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Quantile-based MANOVA: A new tool for inferring multivariate data in factorial designs

Marléne Baumeister [a,b,*], Marc Ditzhaus [c], Markus Pauly [a,b]

[a] Department of Statistics, TU Dortmund University, Germany
[b] Research Center Trustworthy Data Science and Security, UA Ruhr, Germany
[c] Faculty of Mathematics, Otto von Guericke University Magdeburg, Germany

ARTICLE INFO

Article history:
Received 14 December 2022
Received in revised form 16 October 2023
Accepted 16 October 2023
Available online 27 October 2023

AMS 2020 subject classifications:
62H15
62G09
62G10

Keywords:
Efron's bootstrap
Factorial designs
Heteroscedasticity
Multivariate analysis of variance
Nonparametric inference
Quantile-based analysis

ABSTRACT

Multivariate analysis-of-variance (MANOVA) is a well established tool to examine multivariate endpoints. While classical approaches depend on restrictive assumptions like normality and homogeneity, there is a recent trend to more general and flexible procedures. In this paper, we proceed on this path, but do not follow the typical mean-focused perspective. Instead we consider general quantiles, in particular the median, for a more robust multivariate analysis. The resulting methodology is applicable for all kind of factorial designs and shown to be asymptotically valid. Our theoretical results are complemented by an extensive simulation study for small and moderate sample sizes. An illustrative data analysis is also presented.

1. Introduction

In various fields, e.g., biology, ecology, medicine, or psychology, several outcome variables are of simultaneous interest leading to multivariate data. For example, an ecologist may study the aggression against predators and the relative reproductive success (fitness) of birds grouped by sex and colour morph [5]. Other examples are psychological tests or different medical quantities, e.g., heart rate, blood pressure, weight, or height of a patient. As pointed out by Warne [51], multivariate analysis-of-variance (MANOVA) is "one of the most common multivariate statistical procedures in the social science literature". However, classical MANOVA [2,9,28,39,52] relies on restrictive assumptions as normality and homogeneity of covariances. But the "normality assumption becomes quasi impossible to justify when moving from univariate to multivariate observations" [26] and, similarly, homogeneity is often implausible. To overcome these, several remedies have been suggested for tackling at least one of both issues. Thereby solutions have been developed for specific layouts, e.g., one- or two-way [3,27,53,54] as well as for general factorial designs [18,26,46]. As common in statistical inference, all these proposals focus on the expectation (vector) and thus infer means or contrasts thereof. For heavy-tail distributions and in case of outliers the mean is not the appropriate statistical estimand [cf. 31]. Therefore, the present paper aims to introduce

* Corresponding author at: Department of Statistics, TU Dortmund University, Germany.
E-mail address: baumeister@statistik.tu-dortmund.de (M. Baumeister).

https://doi.org/10.1016/j.jmva.2023.105246

(i) a robust, quantile-based counterpart to mean-based MANOVA procedures
(ii) without assuming a specific distribution class (such as normality or sphericity)
(iii) while allowing for potential heterogeneity
(iv) in the framework of general factorial designs, including, e.g., higher-way layouts.

To achieve these aims, we extend the recently proposed QANOVA [quantile-based analysis-of-variance, 10] approach for univariate endpoints to multivariate settings. The QANOVA procedure is a powerful alternative to the commonly established quantile regression [24,25] and allows "the simple incorporation of interaction effects without a loss in power" [10]. Thus, it remains to answer the question how to extend QANOVA as there exists several possibilities to define multivariate quantiles, see, e.g., Small [47], Serfling [44] and Becker et al. [4]. For example, the R-Package `MNM` [35] allows for MANOVA analyses based on the spatial median by [36] and its affine equivariant version, the Hettmansperger–Randles median. `MNM` covers tests equivalent to Hotelling's T^2-Test for more than two samples and tests regarding randomized block designs [37].

Different to `MNM`, our method is based on the vector of marginal quantiles [1]. The marginal quantiles have the advantage, that they are computational more efficient and easier to interpret. In particular, it allows for compatible post hoc analyses on univariate components with the QANOVA. The proposed method will be established within a fully heterogeneous model and can be used for any quantile, not only for the median. For proofing correctness of our method we employ refined results on empirical quantile processes [10,49], and combine them with strategies and ideas from Friedrich and Pauly [18] for mean-based MANOVA.

The paper is structured as follows. The model and the estimators for the population quantiles are presented in Section 2 together with a brief introduction to general factorial designs. Section 3 presents the statistical methods. First (Section 3.1), the test statistics are constructed and their mathematical foundation is explained. Thereafter, covariance estimators based on kernel density estimators [34], bootstrapping [8,12] and an interval-based strategy [40] are proposed (Section 3.2). Finally, a (group-wise) bootstrap strategy is considered (Section 3.3) to estimate the unknown limit distribution of the test statistics. All proofs and some technical details are presented in the Appendix. To investigate the method's small sample properties and compare it with existing methods, an extensive simulation study is carried out in Section 4. An illustrative data analysis of Egyptian skulls complements our investigation (Section 5). The paper closes with a discussion and an outlook.

2. Motivation and set-up

We consider a general, multivariate model based on mutually independent d-dimensional observation vectors for individuals from k different (sub-)groups, e.g., representing different treatments or different epochs of antique objects as in Section 5. In detail, the jth observations vector in group i is denoted by

$$\mathbf{X}_{ij} = \left(X_{ij1}, \dots, X_{ijd}\right)^T \sim \mathbb{P}_i, \quad i \in \{1, \dots, k\},\ j \in \{1, \dots, n_i\}.$$

Here, $\mathbb{P}_i$ denotes the joint distribution with corresponding multivariate distribution function $\mathbf{F}_i$. Furthermore, let $F_{i\ell}$ be the continuous marginal distribution function of $X_{i1\ell}$, $i \in \{1, \dots, k\}$, $\ell \in \{1, \dots, d\}$ with existing density function $f_{i\ell}$. The joint distribution function of two entries X_{ijm} and $X_{ij\ell}$ is denoted by $\mathbf{F}_{i\ell m}$, $i \in \{1, \dots, k\}$, $m, \ell \in \{1, \dots, d\}$. Throughout this paper, we like to infer the vector of marginal quantiles [cf. 1] $\mathbf{q}_i = (q_{i1}, \dots, q_{id})^T$, where

$$q_{i\ell} = F_{i\ell}^{-1}(p) = \inf\{t \in \mathbb{R} | F_{i\ell}(t) \geq p\}, \quad i \in \{1, \dots, k\},\ \ell \in \{1, \dots, d\},$$

for a pre-specified quantile level $p \in (0, 1)$, e.g., $p = 0.5$ for medians. This means in particular that we consider all percentiles in this framework. Within this setting, we want to develop testing procedures for the general null hypothesis

$$\mathcal{H}_0 : \mathbf{H}\mathbf{q} = \mathbf{0}_r, \quad \mathbf{q} = (\mathbf{q}_1^T, \dots, \mathbf{q}_k^T)^T, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{r \times dk}$ is a contrast matrix, i.e., $\mathbf{H}\mathbf{1}_{dk} = \mathbf{0}_r$, and $\mathbf{1}_r$ and $\mathbf{0}_r$ are the r-dimensional vectors of 0's and 1's, respectively. The concrete choice of $\mathbf{H}$ depends on the underlying research question and is similar to classical mean-based MANOVA. For example, the one-way MANOVA hypothesis of no group effect is obtained by selecting $\mathbf{H} = \mathbf{P}_k \otimes \mathbf{I}_d$, where $\otimes$ denotes the Kronecker product of matrices:

$$\mathcal{H}_0 : \left(\mathbf{P}_k \otimes \mathbf{I}_d\right)\mathbf{q} = \mathbf{0}_{dk} \Leftrightarrow \mathcal{H}_0 : \mathbf{q}_1 = \dots = \mathbf{q}_\mathbf{k}.$$

Turning to a two-way layout with factors A having a levels and B possessing b levels, we split up the group index i into $i = (i_1, i_2)$ for $i_1 \in \{1, \dots, a\}$ and $i_2 \in \{1, \dots, b\}$ resulting in $k = a \cdot b$ (sub-)groups. In a more lucid way, the multivariate quantile $\mathbf{q}_i$ can be decomposed into a general effect $\mathbf{q}^\mu$, main effects $\mathbf{q}_{i_1}^\alpha$, $\mathbf{q}_{i_2}^\beta$ and an interaction effect $\mathbf{q}_{i_1 i_2}^{\alpha\beta}$ as

$$\mathbf{q}_i = \mathbf{q}_{(i_1, i_2)} = \mathbf{q}^\mu + \mathbf{q}_{i_1}^\alpha + \mathbf{q}_{i_2}^\beta + \mathbf{q}_{i_1 i_2}^{\alpha\beta},$$

assuming the usual side conditions

$$\sum_{i_1=1}^{a} \mathbf{q}_{i_1}^\alpha = \sum_{i_2=1}^{b} \mathbf{q}_{i_2}^\beta = \sum_{i_1=1}^{a} \mathbf{q}_{i_1 i_2}^{\alpha\beta} = \sum_{i_2=1}^{b} \mathbf{q}_{i_1 i_2}^{\alpha\beta} = \mathbf{0}_d$$

to ensure identifiability. Then, null hypotheses for main and interaction effects can be formulated as follows:

$$\begin{aligned}
\mathcal{H}_0(A) &: \left(\mathbf{P}_a \otimes \frac{1}{b}\mathbf{J}_b \otimes \mathbf{I}_d\right)\mathbf{q} = \mathbf{0}_{dk} \Leftrightarrow \bar{\mathbf{q}}_{1\cdot} = \cdots = \bar{\mathbf{q}}_{a\cdot} \Leftrightarrow \mathbf{q}_{i_1}^{\alpha} = 0,\ i_1 \in \{1,\dots,a\},\\
\mathcal{H}_0(B) &: \left(\frac{1}{a}\mathbf{J}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d\right)\mathbf{q} = \mathbf{0}_{dk} \Leftrightarrow \bar{\mathbf{q}}_{\cdot 1} = \cdots = \bar{\mathbf{q}}_{\cdot b} \Leftrightarrow \mathbf{q}_{i_2}^{\beta} = 0,\ i_2 \in \{1,\dots,b\},\\
\mathcal{H}_0(AB) &: (\mathbf{P}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d)\,\mathbf{q} = \mathbf{0}_{dk} \Leftrightarrow \bar{\mathbf{q}}_{\cdot\cdot} - \bar{\mathbf{q}}_{i_1\cdot} - \bar{\mathbf{q}}_{\cdot i_2} + \mathbf{q}_{i_1 i_2} \equiv \mathbf{0}_d,\ i_1 \in \{1,\dots,a\}, i_2 \in \{1,\dots,b\}\\
&\qquad\qquad \Leftrightarrow \mathbf{q}_{i_1 i_2}^{\alpha\beta} = 0,\ i_1 \in \{1,\dots,a\},\ i_2 \in \{1,\dots,b\}.
\end{aligned}$$

Here, $\mathbf{P}_d = \mathbf{I}_d - \frac{1}{d}\mathbf{J}_d$ is the d-dimensional centring matrix, $\mathbf{I}_d$ is the d-dimensional identity matrix and $\mathbf{J}_d = \mathbf{1}_d^T\mathbf{1}_d$ is the $d \times d$ matrix consisting of 1's only. The $\bar{\mathbf{q}}_{i_1\cdot}$, $\bar{\mathbf{q}}_{\cdot i_2}$ and $\bar{\mathbf{q}}_{\cdot\cdot}$ are the means over the dotted indices. Higher-way layouts and also hierarchically designs with nested factors can be incorporated in a similar way, see e.g., [16,18,38] for mean-based testing strategies.

We like to stress that different matrices $\mathbf{H}$ can describe the same null hypotheses but may affect the statistic's outcome [43]. In the sequel we therefore follow the common practice to reformulate (1) as $\mathbf{Tq} = \mathbf{0}_{dk}$ with the unique projection matrix $\mathbf{T} = \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T\right)^+\mathbf{H} \in \mathbb{R}^{dk\times dk}$, where $\mathbf{A}^+$ denotes the Moore–Penrose inverse of the matrix $\mathbf{A}$. It is easy to check that $\mathbf{H}$ and $\mathbf{T}$ lead to equivalent null hypotheses while the matrix $\mathbf{T}$ has the advantage of being unique, symmetric and idempotent [3,15,26,38]. To infer (1) based on real-data, the marginal quantiles are estimated via the empirical quantiles

$$\hat{q}_{i\ell} = \hat{F}_{i\ell}^{-1}(p) = \inf\{t \in \mathbb{R} | \hat{F}_{i\ell}(t) \geq p\} = X^{(i\ell)}_{\lceil n_i p\rceil:n_i},\ i \in \{1,\dots,k\},\ \ell \in \{1,\dots,d\},$$

where $X^{(i\ell)}_{1:n_i} \leq \cdots \leq X^{(i\ell)}_{n_i:n_i}$ are the order statistics of the ℓ-th component within group i and

$$\hat{F}_{i\ell}(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{1}_{\{X_{ij\ell}\leq t\}}$$

is the respective marginal empirical distribution function evaluated at time $t \in \mathbb{R}$. Together with the pre-chosen contrast matrix and respective covariance estimators (discussed in the next section), they are used in quadratic form-type test statistics to infer (1).

3. Statistical methods

3.1. Construction of tests

As we want to develop an (at least) asymptotically valid method, we first recall the central limit theorem for marginal quantiles [1]. It relies on the following two standard regularity assumptions on the sample sizes and the distribution functions. Here and subsequently, all limits are meant as $n \to \infty$.

Assumption 1. The groups do not vanish, i.e., $(n_i/n) \to \kappa_i > 0$.

Assumption 2. Let $F_{i\ell}$ be continuously differentiable at $q_{i\ell}$ with positive derivative $f_{i\ell}(q_{i\ell}) > 0$ for every $\ell \in \{1,\dots,d\}$ and $i \in \{1,\dots,k\}$.

Proposition 1 (*Theorem 2.1 of Babu and Rao [1]*)**.** *Under Assumptions 1 and 2 we have convergence in distribution*

$$\sqrt{n}\left(\hat{q}_{i\ell} - q_{i\ell}\right)_{\ell\in\{1,\dots,d\}} \xrightarrow{d} \mathbf{Z}_i, \quad i \in \{1,\dots,k\},$$

where $\mathbf{Z}_i$ is a zero-mean, multivariate normal distributed random vector with covariance matrix $\mathbf{\Sigma}^{(i)} = (\mathbf{\Sigma}^{(i)}_{\ell m})_{\ell,m=1,\dots,d}$ given by the entries

$$\mathbf{\Sigma}^{(i)}_{\ell m} = \begin{cases} \dfrac{1}{\kappa_i}\dfrac{1}{f^2_{i\ell}(q_{i\ell})}\left(p - p^2\right), & \ell = m,\\ \dfrac{1}{\kappa_i}\dfrac{1}{f_{i\ell}(q_{i\ell})f_{im}(q_{im})}\left\{\mathbf{F}_{i\ell m}\left(F^{-1}_{i\ell}(p), F^{-1}_{im}(p)\right) - p^2\right\}, & \ell \neq m. \end{cases} \tag{2}$$

This was also proven by [1] without the use of empirical processes. For ease of convenience we present an empirical processes based proof for Proposition 1 in A.2. There, a central limit theorem for multivariate quantiles is deduced from Proposition 4 with the functional delta-method for empirical processes [49, Thm. 3.9.4]. In principle, Proposition 1 and the group's independence allow us to construct quadratic form test statistics in terms of the vector $\sqrt{n}(\hat{\mathbf{q}} - \mathbf{q})$. For this purpose, we only require an appropriate estimator for the unknown limiting covariance matrix $\mathbf{\Sigma} := \bigoplus_{i=1}^k \mathbf{\Sigma}^{(i)}$. Consistent

proposals are discussed in Section 3.2. Let us suppose for a moment, that $\widehat{\mathbf{\Sigma}}$ consistently estimates $\mathbf{\Sigma}$. Then we propose a so-called ANOVA-type statistic (ATS) [7] and a modified ANOVA-type statistic (MATS) [18] to infer $\mathbf{Tq} = \mathbf{0}_{dk}$:

$$\mathbf{ATS}_n(\mathbf{T}) = n\frac{(\mathbf{T}\hat{\mathbf{q}})^{\mathsf{T}}\mathbf{T}\hat{\mathbf{q}}}{\operatorname{tr}\left(\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T}\right)}, \quad \mathbf{MATS}_n(\mathbf{T}) = n(\mathbf{T}\hat{\mathbf{q}})^{\mathsf{T}}\left(\mathbf{T}\hat{\mathbf{\Sigma}}_{\mathbf{0}}\mathbf{T}\right)^{+}\mathbf{T}\hat{\mathbf{q}}.$$

Here, $\mathbf{\Sigma}_0$ and $\hat{\mathbf{\Sigma}}_0$ denote the matrices containing only the diagonal elements of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$, respectively. As described in Sattler et al. [43], both test statistics can be unified into the following general form

$$S_n(\mathbf{T}) = n(\mathbf{T}\hat{\mathbf{q}})^T \mathit{\Xi}\left(\mathbf{T}, \hat{\mathbf{\Sigma}}\right)\mathbf{T}\hat{\mathbf{q}},$$

where $\mathit{\Xi}(\mathbf{T}, \hat{\mathbf{\Sigma}}) = \operatorname{tr}(\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T})^{-1}\mathbf{I}_{dk}$ for the ATS and $\mathit{\Xi}(\mathbf{T}, \hat{\mathbf{\Sigma}}) = (\mathbf{T}\hat{\mathbf{\Sigma}}_{\mathbf{0}}\mathbf{T})^{+}$ for the MATS. In contrast to the QANOVA the limiting unknown covariance matrix $\mathit{\Sigma}$ is not necessarily non-singular. This complicates working with $\mathit{\Xi}(\mathbf{T}, \hat{\mathbf{\Sigma}}) = (\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T})^{+}$ as in common Wald-type statistics [50]. In particular, the corresponding Moore–Penrose inverse is in general no longer consistent, hampering the usual χ^2-inference. We therefore do not consider Wald-type statistics in our paper. For the following Proposition we need another technical assumption:

Assumption 3. The diagonal elements of $\mathit{\Sigma}^{(i)}$ are positive, i.e. $\mathit{\Sigma}^{(i)}_{\ell\ell} > 0,\ i \in \{1, \dots, k\},\ \ell \in \{1, \dots, d\}$.

Proposition 2. *For a consistent estimator $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ and under Assumption 3 the versions of $\mathit{\Xi}(\mathbf{T}, \hat{\mathbf{\Sigma}})$ in the ATS and the MATS are consistent for $\mathit{\Xi}(\mathbf{T}, \mathbf{\Sigma})$. Thus,*

(i) $tr\left(\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T}\right)^{-1}\mathbf{I}_{dk} \xrightarrow{P} tr\left(\mathbf{T}\mathbf{\Sigma}\mathbf{T}\right)^{-1}\mathbf{I}_{dk}$;
(ii) $\left(\mathbf{T}\hat{\mathbf{\Sigma}}_{\mathbf{0}}\mathbf{T}\right)^{+} \xrightarrow{P} \left(\mathbf{T}\hat{\mathbf{\Sigma}}_{\mathbf{0}}\mathbf{T}\right)^{+}$.

This technical result is used in the following theorem, which summarizes the asymptotic distribution of $S_n(\mathbf{T})$.

Theorem 1. *Let $\hat{\mathbf{\Sigma}}$ be a consistent covariance matrix estimator of $\mathbf{\Sigma}$ and assume that Assumption 1, 2 and 3 hold.*

(i) Under $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}_{dk}$, the test statistic S_n converges in distribution to a weighted sum of χ^2_1 distributed random variables, i.e.,

$$S_n(\mathbf{T}) \xrightarrow{d} B = \sum_{i=1}^{dk}\lambda_i B_i, \tag{3}$$

where $B_i \overset{iid}{\sim} \chi^2_1$ and $\lambda_i \geq 0,\ i \in \{1, \dots, dk\}$, are the eigenvalues of $(\mathbf{T}\mathbf{\Sigma}\mathbf{T})^{\frac{1}{2}}\mathit{\Xi}\left(\mathbf{T}, \mathbf{\Sigma}\right)(\mathbf{T}\mathbf{\Sigma}\mathbf{T})^{\frac{1}{2}}$.
(ii) Under $\mathcal{H}_1 : \mathbf{Tq} \neq \mathbf{0}_{dk}$, S_n converges in probability to ∞.

Let b_α be the $(1-\alpha)$-quantile of B in (3). From Theorem 1 we can deduce that the test $\varphi_n = \mathbf{1}\{S_n(\mathbf{T}) > b_\alpha\}$ is of asymptotic level α for $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}$. Furthermore, it is consistent for any alternative $\mathcal{H}_1 : \mathbf{Tq} \neq \mathbf{0}_{dk}$. However, the distribution of B depends on unknown parameters through $\mathbf{\Sigma}$. Thus, b_α is in general unknown and we consider a bootstrap procedure to approximate it, which we discuss in Section 3.3. But first we address the pending question regarding to the estimation of $\mathbf{\Sigma}$.

3.2. Estimation of the covariance matrix

Variance estimation of the median or general quantiles is not easy in the univariate setting and various strategies can be found in the literature [6,8,30,33]. At a first glance, the situation becomes even more complicated in the multivariate set-up. However, a careful observation of (2) yields a simple relationship between the diagonal and off-diagonal elements:

$$\mathbf{\Sigma}^{(i)}_{\ell m} = \sqrt{\mathbf{\Sigma}^{(i)}_{\ell\ell}\mathbf{\Sigma}^{(i)}_{mm}}\,\frac{\mathbf{F}_{i\ell m}(q_{i\ell}, q_{im}) - p^2}{p - p^2},\ \ell \neq m.$$

Thus, the known univariate strategies to estimate the variances $\mathbf{\Sigma}^{(i)}_{\ell\ell}$, $\ell \in \{1, \dots, k\}$, can be combined with an estimator for $\mathbf{F}_{i\ell m}(q_{i\ell}, q_{im})$. For the latter, let us first introduce the joint empirical distribution function $\hat{\mathbf{F}}_{i\ell m}$ defined by

$$\hat{\mathbf{F}}_{i\ell m}(t_1, t_2) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{1}_{(-\infty, t_1]\times(-\infty, t_2]}(X_{ij\ell}, X_{ijm})$$

for $i \in \{1, \dots, k\}$, $\ell, m \in \{1, \dots, d\}$ and $t_1, t_2 \in \mathbb{R}$. Under the following regularity assumption, consistency of $\hat{\mathbf{F}}_{i\ell m}(\hat{q}_{i\ell}, \hat{q}_{im})$ for $\mathbf{F}_{i\ell m}(q_{i\ell}, q_{im})$ holds:

Assumption 4. For every $i \in \{1, \dots, k\}$ and every $\ell, m \in \{1, \dots, d\}$, the joint distribution function $\mathbf{F}_{i\ell m}$ is continuous at $(q_{i\ell}, q_{im}) = (F^{-1}_{i\ell}(p), F^{-1}_{im}(p))$.

Proposition 3 (*Theorem 2.2 of [1]*). *Under Assumption 1 and 4, the estimator* $\hat{\mathbf{F}}_{i\ell m}(\hat{q}_{i\ell}, \hat{q}_{im})$ *converges in probability to* $\mathbf{F}_{i\ell m}(q_{i\ell}, q_{im})$.

Because Babu and Rao [1] proved this, we present a proof with empirical processes in Appendix A.5. Consequently, we obtain a general form of estimators for $\mathbf{\Sigma}^{(i)}$:

$$\hat{\mathbf{\Sigma}}^{(i)}_{\ell m} = \begin{cases} \dfrac{n}{n_i}\hat{\sigma}^2_{i\ell}(p), \ \ell = m, \\ \dfrac{n}{n_i}\hat{\sigma}_{i\ell}(p)\hat{\sigma}_{im}(p)\dfrac{\hat{\mathbf{F}}_{i\ell m}(\hat{q}_{i\ell}, \hat{q}_{im}) - p^2}{p - p^2}, \ \ell \neq m, \end{cases}$$

where $\hat{\sigma}^2_{i\ell}(p)$ is a consistent estimator for the asymptotic variance $\sigma^2_{i\ell}(p) = \kappa_i \mathbf{\Sigma}^{(i)}_{\ell\ell}$ of the marginal, centred empirical quantiles $\sqrt{n}(\widehat{q}_{i\ell} - q_{i\ell})$. We follow Ditzhaus et al. [10] and study three different choices for $\hat{\sigma}^2_{i\ell}(p)$ considered in the univariate case. All approaches produce consistent estimators under the respective assumptions for $\hat{\mathbf{\Sigma}}^{(i)}$ [10].

3.2.1. Kernel estimator

The *kernel-based approach* uses the strong consistent kernel density estimator by Nadaraya [34] to estimate the densities $f_{i\ell}$. It is given by

$$\hat{f}_{K,i,\ell}(x) = \frac{1}{n_i h_{n_i\ell}} \sum_{j=1}^{n_i} K_{i\ell}\left(\frac{x - X_{ij\ell}}{h_{n_i\ell}}\right), \ i \in \{1, \dots, k\}; \ \ell \in \{1, \dots, d\},$$

where $K_{i\ell}$ is a kernel and $h_{n_i\ell}$ is a bandwidth, $i \in \{1, \dots, k\}$; $\ell \in \{1, \dots, d\}$. For its strong consistency we require:

Assumption 5. Suppose for every $i \in \{1, \dots, k\}$; $\ell \in \{1, \dots, d\}$ that $K_{i\ell}$ is of bounded variation, f_i is uniformly continuous and the series $\sum_{m=1}^{\infty} \exp(-\gamma m h^2_{n_i\ell})$ converges for every choice of $\gamma > 0$.

This leads to the following consistent estimator for $\sigma^2_{i\ell}(p)$:

$$\hat{\sigma}^2_{i\ell,K}(p) = \frac{1}{\hat{f}^2_{K,i,\ell}(\hat{q}_{i\ell})}(p - p^2).$$

3.2.2. Bootstrap estimator

The *bootstrap approach* was originally proposed by [8], who borrowed the idea from [12]. To introduce it, consider the bootstrap samples $X^*_{i1\ell}, \dots, X^*_{in_i\ell}$, $i \in \{1, \dots, k\}$, $\ell \in \{1, \dots, d\}$, drawn mutually independent and with replacement from the observations $X_{i1\ell}, \dots, X_{in_i\ell}$. We denote all estimators based on the bootstrap sample by a *, e.g., $\hat{q}^*$. Then, the bootstrap sample quantile estimator $\hat{q}^*_{i\ell}$ can be calculated and its conditional mean squared error, given data, is given by

$$\left(\hat{\sigma}^*_{i\ell}(p)\right)^2 = n_i \sum_{j=1}^{n_i} \left(X^{(i\ell)}_{j:n_i} - \hat{q}_{i\ell}\right)^2 \underbrace{Pr^*\left(X^{(i\ell)*}_{\lceil n_i p\rceil:n_i} = X^{(i\ell)}_{j:n_i} | \mathbf{X}_{i\ell}\right)}_{:=P^*_{ij\ell}}. \tag{4}$$

As explained by Efron [12], $P^*_{ij\ell}$ can be rewritten for every $\ell \in \{1, \dots, d\}$ as

$$P^*_{ij\ell} = Pr\left(B_{n_i, \frac{j-1}{n_i}} \leq \lceil n_i p\rceil - 1\right) - Pr\left(B_{n_i, \frac{j}{n_i}} \leq \lceil n_i p\rceil - 1\right),$$

where $B_{n,p}$ denotes a binomial distributed random variable with size parameter n and success probability p. Ghosh et al. [20] proved that the estimator $\left(\hat{\sigma}^*_{i\ell}(p)\right)^2$ is consistent for $\sigma^2_{i\ell}(p)$ under the following moment condition.

Assumption 6. For some $\delta > 0$ we have $\max_{i\in\{1,\dots,k\},\, \ell\in\{1,\dots,d\}} \mathrm{E}\left(|X_{i1\ell}|^{\delta}\right) < \infty$.

3.2.3. Interval-based estimator

An *interval-based approach* was initially suggested by McKean and Schrader [33] and later modified by Price and Bonett [40] for the median. The methodology can easily be adapted to handle general quantiles [6,10]. The principle idea is to start with the asymptotic confidence interval $(X^{(i\ell)}_{l_i(p):n_i}, X^{(i\ell)}_{u_i(p):n_i})$ for $q_{i\ell}$. Its length (asymptotically) depends on the (unknown) standard deviation $\sigma_{i\ell}(p)$. Basic calculation yields the following estimator:

$$\hat{\sigma}^2_{i\ell,PB}(p) = \left(\sqrt{n_i}\,\frac{X^{(i\ell)}_{u_i(p):n_i} - X^{(i\ell)}_{l_i(p):n_i}}{2z_{\alpha^*_{n_i\ell}(p)/2} + 2n_i^{-1/2}}\right)^2,$$

where $l_i(p) = \max\{1, \lfloor n_i p - z_{\alpha/2}\sqrt{n_i p(1-p)}\rfloor\}$ is the lower and $u_i(p) = \min\{n_i, \lfloor n_i p + z_{\alpha/2}\sqrt{n_i p(1-p)}\rfloor\}$ is the upper limit of the binomial interval, and $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. Note that $l_i(p)$ and

$u_i(p)$ are independent of the dimension $\ell \in \{1, \dots, d\}$, and typically $\alpha = 0.05$ is chosen for their computation. Price and Bonett [40] set

$$\alpha^*_{n_i\ell}(p) = \begin{cases} \alpha_{n_i\ell}(p) = 1 - \sum_{j=l_i(p)+1}^{u_i(p)-1} \binom{n_i}{j} p^j (1-p)^{n_i-j}, & n_i \leq 100, \\ 0.05, & n_i > 100 \end{cases}$$

in the denominator of $\hat{\sigma}^2_{i\ell,PB}(p)$. The case distinction is motivated by the computation time of $\alpha_{n_i\ell}(p)$ which becomes quite demanding for larger sample sizes. Since we have $\alpha_{n_i\ell}(p) \longrightarrow \alpha$ by the central limit theorem, the use of $\alpha_{n_i\ell}(p)$ is only necessary for small to moderate sample sizes.

3.3. Bootstrapping the test statistic

Having one of the presented estimators for the covariance matrices at hand, we are able to calculate the respective ATS or MATS. However, the limiting distribution of B from Theorem 1 remains unknown and the $(1-\alpha)$-quantile b_α of B cannot be computed. That is why we consider a group-wise nonparametric bootstrap approach to approximate it. This resampling strategy was already applied by Friedrich and Pauly [18] for their mean-based MANOVA procedure and is known to be effective for various testing problems [11,26,29]. As in Section 3.2.2, we consider a d-dimensional bootstrap sample $\{\mathbf{X}^*_{i1}, \dots, \mathbf{X}^*_{in_i}\}$ drawn with replacement from the original observation vectors $\{\mathbf{X}_{i1}, \dots, \mathbf{X}_{in_i}\}$ for every group $i \in \{1, \dots, k\}$. Moreover, we add a * to all statistics which are calculated from the bootstrap sample, e.g., $\hat{\mathbf{q}}^*$ denotes the bootstrap quantile vector. Note that under the null hypothesis $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}_{dk}$ the test statistic can be written as

$$S_n(\mathbf{T}) = n\left[\mathbf{T}\left(\hat{\mathbf{q}} - \mathbf{q}\right)\right]^T \Xi\left(\mathbf{T}, \hat{\Sigma}\right)\mathbf{T}\left(\hat{\mathbf{q}} - \mathbf{q}\right).$$

For its bootstrap counterpart, the estimators $\hat{\mathbf{q}}$ and $\hat{\Sigma}$ are replaced by their bootstrap versions $\hat{\mathbf{q}}^*$ and $\hat{\Sigma}^*$ and the unknown quantile vector $\mathbf{q}$ is substituted by its the empirical counterpart $\hat{\mathbf{q}}$. Consequently, we obtain the bootstrap statistic

$$S^*_n(\mathbf{T}) = n\left[\mathbf{T}\left(\hat{\mathbf{q}}^* - \hat{\mathbf{q}}\right)\right]^T \Xi\left(\mathbf{T}, \hat{\Sigma}^*\right)\mathbf{T}\left(\hat{\mathbf{q}}^* - \hat{\mathbf{q}}\right). \tag{5}$$

We can derive the (conditional) asymptotic behaviour of $S^*_n(T)$ by slightly adopting the argumentation for Proposition 1 and combine that with the bootstrap results of van der Vaart and Wellner [49] to get an equivalent result to Theorem 1:

Theorem 2. *Let $\hat{\Sigma}$ be a consistent estimator for Σ and let $\hat{\Sigma}^*$ denote its consistent bootstrap version. Then, the bootstrap test statistic $S^*_n(\mathbf{T})$ given in (5) converges always, conditionally given the data, in distribution to a real-valued random variable B^*, i.e., we have under $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}_{dk}$ as well as under $\mathcal{H}_1 : \mathbf{Tq} \neq \mathbf{0}_{dk}$*

$$\sup_{x\in\mathbb{R}} \left| Pr\left(S^*_n(\mathbf{T}) \leq x | \mathbf{X}\right) - P\left(B^* \leq x\right)\right| \xrightarrow{P} 0.$$

Hereby, the distribution of B^* depends on the underlying setting and can be expressed by $B^* = \sum_{i=1}^{dk} \lambda^*_i B_i$ where $\lambda^*_i \geq 0$ and $B_i \sim \chi^2_1$. Under $\mathcal{H}_0$, the distribution of B^* coincides with the limit null distribution of $S_n(\mathbf{T})$, i.e., $B \overset{d}{=} B^*$.

Our proposed resampling test use $b^*_\alpha(\mathbf{X})$, the empirical $(1-\alpha)$-quantile of the conditional distribution function $x \mapsto P(S^*_n(\mathbf{T}) \leq x | \mathbf{X})$ as critical value. This leads to the test $\varphi^*_n = \mathbf{1}\{S_n(\mathbf{T}) > b^*_\alpha(\mathbf{X})\}$. Under $\mathcal{H}_0$, Theorem 2 implies that $b^*_\alpha(\mathbf{X})$ converges in probability to b_α given the data. Thus, combining Lemma 1 of Janssen and Pauls [23], Theorems 1 and 2, we obtain that the resampling test φ^*_n is asymptotically exact, i.e., $\mathrm{E}_{\mathcal{H}_0}(\varphi^*_n) \to \alpha$. Moreover, φ^*_n is even consistent for general alternatives $\mathcal{H}_0 : \mathbf{Tq} \neq \mathbf{0}_{dk}$. To accept this, we first deduce from Theorem 2 that $b^*_\alpha(\mathbf{X})$ converges in probability to some $\widetilde{b} \in \mathbb{R}$ given the data under $\mathcal{H}_1$. Combining this observation with Theorem 1(ii) and Theorem 7 of Janssen and Pauls [23] yields the desired consistency.

4. Simulations

To assess the tests' performances for small and moderate sample sizes, we conducted a simulation study. The idea of the data generation is as follows. We generate median-centred data $\mathbf{e}_{ij} = \left(e_{ij1}, \dots, e_{ijn_i}\right)^T$, $e_{ij\ell} = Y_{ij\ell} - \mathrm{median}(Y_{ij\ell})$ and choose for $Y_{ij\ell}$ the following five different distributions to cover symmetric and skewed scenarios: (a) the standard normal distribution $Y^{(1)}_{ij\ell} \sim N_{0,1}$, (b) the Student's t-distribution with 2 degrees of freedom $Y^{(2)}_{ij\ell} \sim t_2$, (c) the Student's t-distribution with 3 degrees of freedom $Y^{(3)}_{ij\ell} \sim t_3$, (d) the standard log-normal distribution $Y^{(4)}_{ij\ell} \sim LN_{0,1}$ and (e) the Chi-square distribution with 3 degrees of freedom $Y^{(5)}_{ij\ell} \sim \chi^2_3$. Certain homoscedastic and heteroscedastic covariance settings are realized by multiplying the square root of different covariance matrices to this data. Furthermore, we considered six different covariance matrices representing homoscedastic and heteroscedastic scenarios, which are displayed below for $d = 4$:

(i) $\Sigma^{(1)} = \mathbf{I}_d - \frac{1}{2}\left(\mathbf{J}_d - \mathbf{I}_d\right) = \Sigma^{(2)}$, (ii) $\Sigma^{(1)} = \left(0.6^{|a-b|}\right)_{a,b=1}^{d} = \Sigma^{(2)}$,

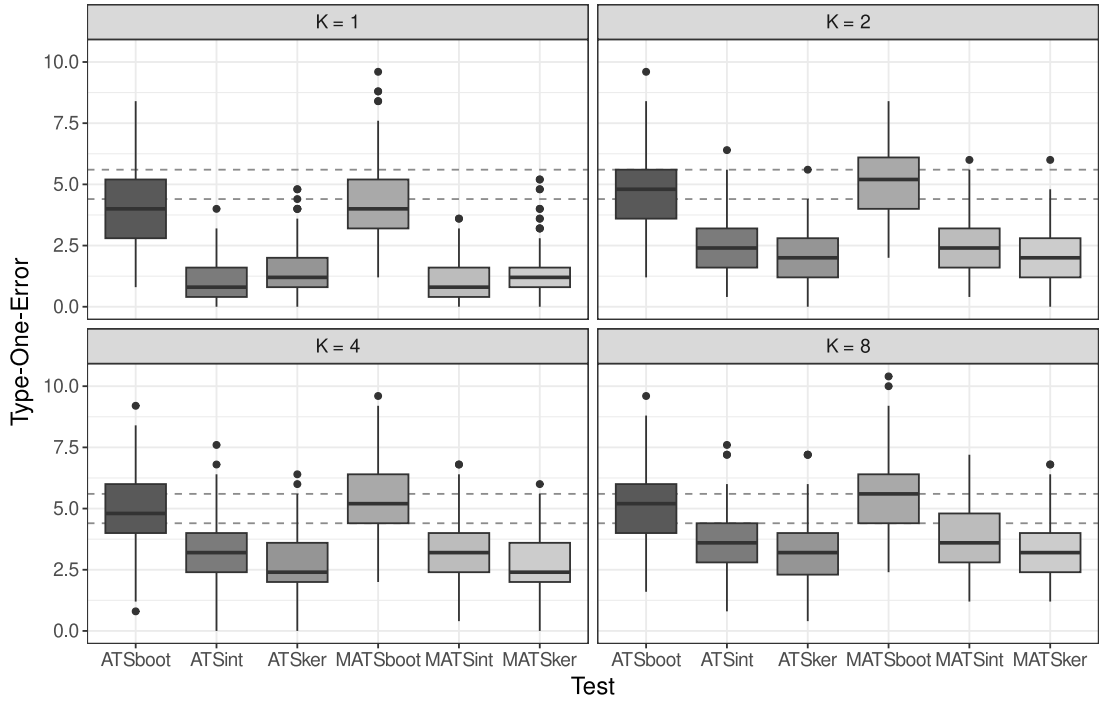

Fig. 1. Type I error in % of the six versions of the QMANOVA divided into two types (ATS, MATS) and three covariance estimator (boot, int, ker). The results are divided by the sample size factor $K \in \{1, 2, 4, 8\}$, i.e., $\mathbf{n} = K\mathbf{n}^{(r)}$, $r \in \{1, 2, 3\}$.

(iii) $\boldsymbol{\Sigma}^{(1)} = \mathbf{I}_d + \frac{1}{2}(\mathbf{J}_d - \mathbf{I}_d)$, $\boldsymbol{\Sigma}^{(2)} = 3\,\mathbf{I}_d + \frac{1}{2}(\mathbf{J}_d - \mathbf{I}_d)$, (iv) $\boldsymbol{\Sigma}^{(1)} = \left(0.6^{|a-b|}\right)_{a,b=1}^{d}$, $\boldsymbol{\Sigma}^{(2)} = \left(0.6^{|a-b|}\right)_{a,b=1}^{d} + 2\,\mathbf{I}_d$,

(v) $\boldsymbol{\Sigma}^{(1)} = \begin{pmatrix} 1 & 0.6 & 0.36 & 0.18 \\ 0.6 & 1 & 0.6 & 0.3 \\ 0.36 & 0.6 & 1 & 0.5 \\ 0.18 & 0.3 & 0.5 & 0.25 \end{pmatrix}$, $\boldsymbol{\Sigma}^{(2)} = \boldsymbol{\Sigma}^{(1)} + 0.5\,\mathbf{J}_d$,

(vi) $\boldsymbol{\Sigma}^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0.5 \end{pmatrix}$, $\boldsymbol{\Sigma}^{(2)} = \boldsymbol{\Sigma}^{(1)} + 0.5\,\mathbf{J}_d$.

Here, the fifth covariance setting is a modification form the second, where the elements σ_{1dd}, $\sigma_{1(d-1)d}$ and $\sigma_{1d(d-1)}$ of $\boldsymbol{\Sigma}^{(1)}$ are modified as described. The sixth setting is based on $\mathrm{diag}(\sqrt{2^s})$ for $s \in \{0, \ldots, d-1\}$, where the dth row and column is replaced by half the row or rather the column before. To create data which is median centred, we calculate the empirical median $\mathbf{M}_{ij}$ of $(\boldsymbol{\Sigma}^{(i)})^{\frac{1}{2}}\mathbf{e}_{ij}$ from an extra sample with the size $n = 10^7$ and withdraw $\mathbf{M}_{ij}$ from the data. Therefore, our simulated data can be described by the following model:

$$\mathbf{X}_{ij} = (\boldsymbol{\Sigma}^{(i)})^{\frac{1}{2}}\mathbf{e}_{ij} - \mathbf{M}_{ij} \sim \mathbf{F}_i,\ i \in \{1, \ldots, k\}, j \in \{1, \ldots, n_i\}.$$

The aforementioned data generating process and the choice of the different settings is adapted from the one-way layout simulation in Friedrich and Pauly [18, Sec. 5]. To consider small and large sample scenarios, we chose balanced and unbalanced small samples $\mathbf{n}^{(1)} = (10, 10)^T$, $\mathbf{n}^{(2)} = (10, 20)^T$ and $\mathbf{n}^{(3)} = (20, 10)^T$ as well as its multiples $K \cdot \mathbf{n}^{(r)}$ for $K \in \{2, 4, 8\}$. As a benchmark, we compare our method with the mean-based resampling MATS proposed by Friedrich and Pauly [18]. This method is implemented in the R-package `MANOVA.RM` [17] as the functions `MANOVA()` or `MANOVA.wide()` (same tests for different data formats). We simulated two versions of the mean-based MATS, one is characterized by a parametric bootstrap and the other by a wild bootstrap with Rademacher weights, which are both implemented in the aforementioned package. All simulations are calculated with the computing environment `R` [48], Version 4.0.0, for `nsim` = 5000 simulation runs and `nboot` = 2000 bootstrap iterations. As in Ditzhaus et al. [10], we used the classical Gaussian kernel for the kernel density estimation and calculate the bandwidth h by Silverman's rule-of-thumb [45, Eq. 3.31] with the R-function `bw.rnd0()`. The R-Code for the simulations as well as the data analysis script can be found in the following git-repository: https://gitlab.com/charlienriesz/qmanova.

Table 1
Type-I error rate in % (nominal level $\alpha = 5\%$) for testing $\mathcal{H}_0 : \mathbf{m}_1 = \mathbf{m}_2$ in an one-way layout. We present all settings with the sample size (10, 10) of MATS and ATS combined with the bootstrap covariance estimator. The two comparing meanMATS models are named with *meanMATS-p* for a parametric bootstrap and with *meanMATS-w* for the wild bootstrap approach. The results inside the binomial interval for α [4.4, 5.6] are printed in bold.

Σ	Distr	$d = 4$				$d = 8$			
		median		meanMATS		median		meanMATS	
		MATS	ATS	param	wild	MATS	ATS	param	wild
1	$N_{0,1}$	2.4	2.0	3.2	4.0	2.4	2.8	3.2	4.0
	t_2	1.6	**4.4**	2.0	4.0	4.0	6.0	4.0	6.8
	t_3	**4.8**	**5.6**	4.0	**4.4**	**4.4**	3.6	3.6	**4.4**
	$LN_{0,1}$	3.2	3.6	–	–	**4.8**	**4.4**	–	–
	χ^2_3	2.0	2.0	–	–	**4.8**	4.0	–	–
2	$N_{0,1}$	**4.8**	3.2	3.2	**4.4**	**4.4**	2.8	2.8	3.2
	t_2	**4.4**	**4.4**	**4.8**	7.6	**4.8**	**4.8**	2.8	**5.2**
	t_3	**4.4**	3.6	3.6	**4.4**	1.2	**4.4**	3.6	**4.4**
	$LN_{0,1}$	6.0	6.4	–	–	3.6	**5.2**	–	–
	χ^2_3	**5.2**	**4.4**	–	–	3.6	4.0	–	–
3	$N_{0,1}$	3.6	1.2	3.2	4.0	3.2	2.8	**4.8**	7.2
	t_2	4.0	**4.8**	3.2	6.4	3.6	3.2	3.2	**5.2**
	t_3	2.8	2.8	3.6	**5.6**	3.2	2.4	3.6	6.8
	$LN_{0,1}$	1.6	2.0	–	–	3.2	4.0	–	–
	χ^2_3	6.0	3.6	–	–	4.0	2.4	–	–
4	$N_{0,1}$	1.6	1.6	2.4	3.6	2.8	2.8	**4.8**	6.4
	t_2	4.0	2.4	2.4	**5.6**	2.8	3.6	1.2	3.6
	t_3	3.6	3.2	1.2	1.6	2.8	2.0	2.8	**5.6**
	$LN_{0,1}$	**4.8**	**4.8**	–	–	3.6	2.0	–	–
	χ^2_3	**5.2**	**5.6**	–	–	2.4	1.6	–	–
5	$N_{0,1}$	8.8	8.4	8.8	9.6	**4.4**	**4.8**	**4.8**	6.0
	t_2	3.6	3.6	**5.6**	6.0	1.2	2.4	2.8	3.6
	t_3	7.6	8.4	6.0	6.0	4.0	3.2	**5.2**	6.0
	$LN_{0,1}$	3.2	3.6	–	–	6.4	8.0	–	–
	χ^2_3	8.4	6.4	–	–	**4.8**	4.0	–	–
6	$N_{0,1}$	7.2	6.4	**4.8**	**5.6**	1.6	2.4	1.6	2.8
	t_2	**4.8**	3.6	1.6	4.0	2.4	3.2	2.0	4.0
	t_3	4.0	**4.4**	3.2	**4.4**	2.8	**4.4**	2.4	6.0
	$LN_{0,1}$	**5.6**	6.8	–	–	3.2	2.8	–	–
	χ^2_3	**5.2**	3.2	–	–	4.0	6.4	–	–

4.1. Type I error

In this subsection, we discuss the type I error control of all procedures in a one-way layout and present further results for a 2×2-design in the supplement. In detail, we considered a multivariate set-up with $k = 2$ groups and $d = \{4, 8\}$ dimensions. Moreover, we restricted to the median $\mathbf{m}_i$, $i \in \{1, 2\}$, ($p = 0.5$), because it is the most relevant quantile for statistical analysis and is comparable to the mean. This lead us to the null hypothesis $\mathcal{H}_0 : \mathbf{m}_1 = \mathbf{m}_2$ for the layout matrix $\mathbf{T} = \mathbf{P}_2 \otimes \mathbf{I}_4$ and in all to 720 different scenarios. In Fig. 1, the different tests are named by a combination of the used test statistic and its covariance estimator. It is apparent that the bootstrap covariance estimator has the best performance regarding the type I error control. Overall, the MATS and the ATS test statistic have a similar type-one-error control. However, the MATS with the bootstrap covariance estimator performs the best as one can observe from the close position of the boxes to the binomial interval [4.4, 5.6]. With the other covariance estimators, ATS and MATS show a quite conservative type I error control. In general, the observed type I error rates comes closer to the 5%-benchmark for larger K. This is in line with the theoretical findings from Section 3.1. All in all, we can only recommend the ATS and MATS combined with the bootstrap covariance estimator.

In the next step, we compare the two favourable QMANOVA methods, from now denoted by *medMATS* and *medATS*, with the mean-based MATS of [18] denoted by *meanMATS-p* and *meanMATS-w* to differentiate between the parametric (p) and wild (w) bootstrap versions. For a fair and appropriate comparison, we restrict ourselves to the symmetric distributions such that the mean-based hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and the median-based hypothesis $\mathcal{H}_0 : \mathbf{m}_1 = \mathbf{m}_2$ coincide. Note that a comparison of mean- and median-based hypotheses is not meaningful otherwise. The type I error rates for all four tests are summarized in Table 1 for $\mathbf{n}^{(1)} = (10, 10)$ and in the supplement for $\mathbf{n}^{(2)}$ and $\mathbf{n}^{(3)}$. For completeness reasons, we also include the error rates for two QMANOVA strategies for the nonsymmetric distributions ($LN_{0,1}$, χ^2_3). The results inside the binomial interval [4.4, 5.6] for the significance level $\alpha = 5\%$ are printed in bold. A detailed study of Table 1 exhibits that the MATS performs slightly better than the ATS since the simulated type I errors are more frequent in [4.4, 5.6] for the MATS (17 times) than for the ATS (14 times). To further judge the performance for

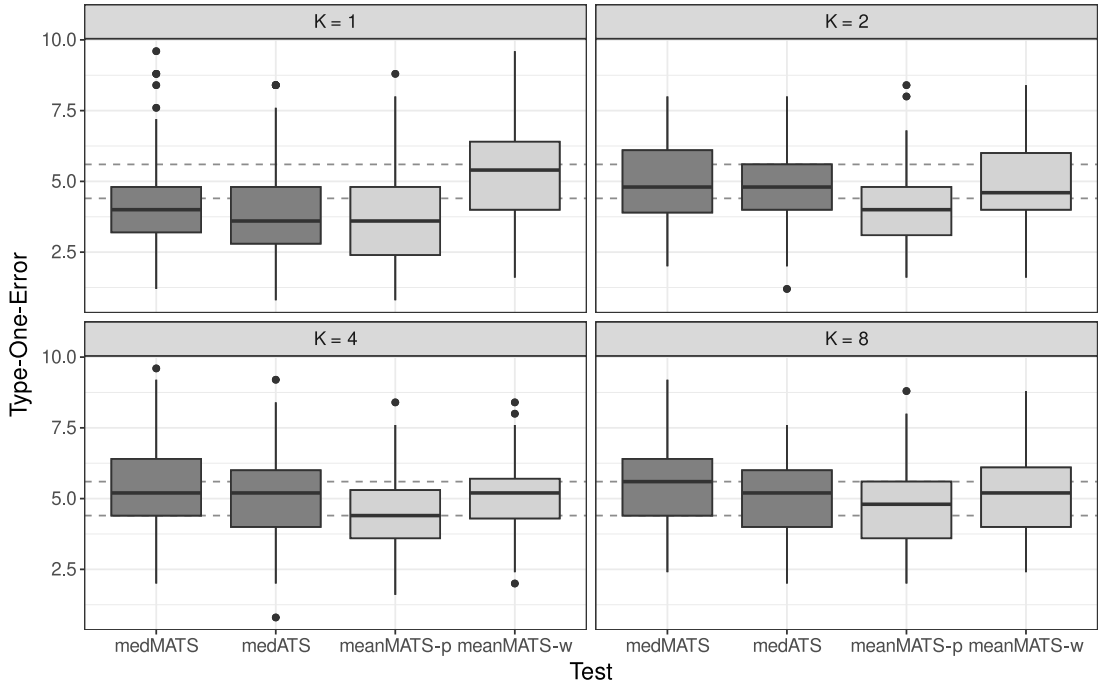

Fig. 2. Type I error in % of the QMANOVA with bootstrap covariance estimator and the mean-based MATS by Friedrich and Pauly [18] with parametric (p) and wild (w) bootstrap for the one-way layout, the symmetric distributions ($N_{0,1}$, t_2, t_3), all covariance settings (1)–(6) as well as for balanced and unbalanced sample sizes $\mathbf{n} = K \cdot \mathbf{n}^{(r)}$ for $K \in \{1, 2, 4, 8\}$.

larger sample sizes, we summarized the type I error rates for all symmetric distributions and all different sample sizes $K\mathbf{n}^{(r)}$, $r \in \{1, 2, 3\}$, in boxplots displayed in Fig. 2 divided into the different choices of the scaling factor $K \in \{1, 2, 4, 8\}$. In Fig. 2, the medMATS tends to be more conservative (empirical type I error smaller than 4.4) and the meanMATS-w tends to be more liberal (empirical type I error larger than 5.6). In case of small sample sizes ($K = 1$), the medATS and both MATS approaches exhibit a conservative type I error control for almost all settings. In settings with $d = 8$ and the covariance choices (4) or (6), a conservative type I error control can even be found for larger samples ($K \in \{2, 4, 8\}$), see Table 1. Additionally, a liberal behaviour most often occurs for the covariance setting (5) in combination with the dimension $d = 4$ (Table 1). For all sample sizes, the MATS tends to be more liberal than the ATS, because more scenarios have a far too liberal behaviour (type I error larger than 7) regarding MATS (74) than ATS (56). The liberal behaviour occurs often in combination with the normal distribution or t-distribution with three degrees of freedom and with $d = 4$ dimensions. For further details on the influence of the simulation scenarios' aspects we refer to the Supplement. There, we explain that the choice of the covariance setting influences the performance most while the QMANOVA method is mostly robust.

4.1.1. More investigations under the null hypothesis

Beyond the presented results, we analysed various another settings and also other methods to calculate critical values in the supplement. We shortly summarize our findings here. All details can be found in the Supplement.

Using Monte Carlo critical values instead of resampling. As suggested by a reviewer it is also tempting to compute critical values from the limiting distribution as in Sattler et al. [43]. In this method, the eigenvalues λ_i in the limiting distribution B in Formula (3) are estimated by the eigenvalues $\hat{\lambda}_i$ of $(\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T})^{\frac{1}{2}}\, \Xi\left(\mathbf{T}, \hat{\mathbf{\Sigma}}\right)(\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T})^{\frac{1}{2}}$ and $\hat{B} = \sum_{i=1}^{dk} \hat{\lambda}_i B_i$ is used for the computation of critical values, i.e. $1-\alpha$-quantiles from $\hat{B}$. We tested the behaviour of this approach for the QMANOVA. To do this we additionally simulated some small-sample scenarios. As with the bootstrap we used `nboot` = 2000 iterations. From our simulation results we can see, that the tests with the Monte-Carlo-resampling has a very conservative behaviour in the sense that the tests rejected almost never. This shows that bootstrapping, though more time-consuming, is the preferred method for computing critical values.

Non continuous distributions. As the assumption of continuous densities is substantial for the proof of the method we simulated some Poisson data and saw that the tests performance is still good in these scenarios. Here, medATS as well as medMATS has a similar performance (median of type I error rate 4.4 for both tests) and tend to be a bit too conservative in their behaviour. From this result it seems that the violation of the assumption has no great influence on the test. The data generation process and detailed results are presented in the Supplement.

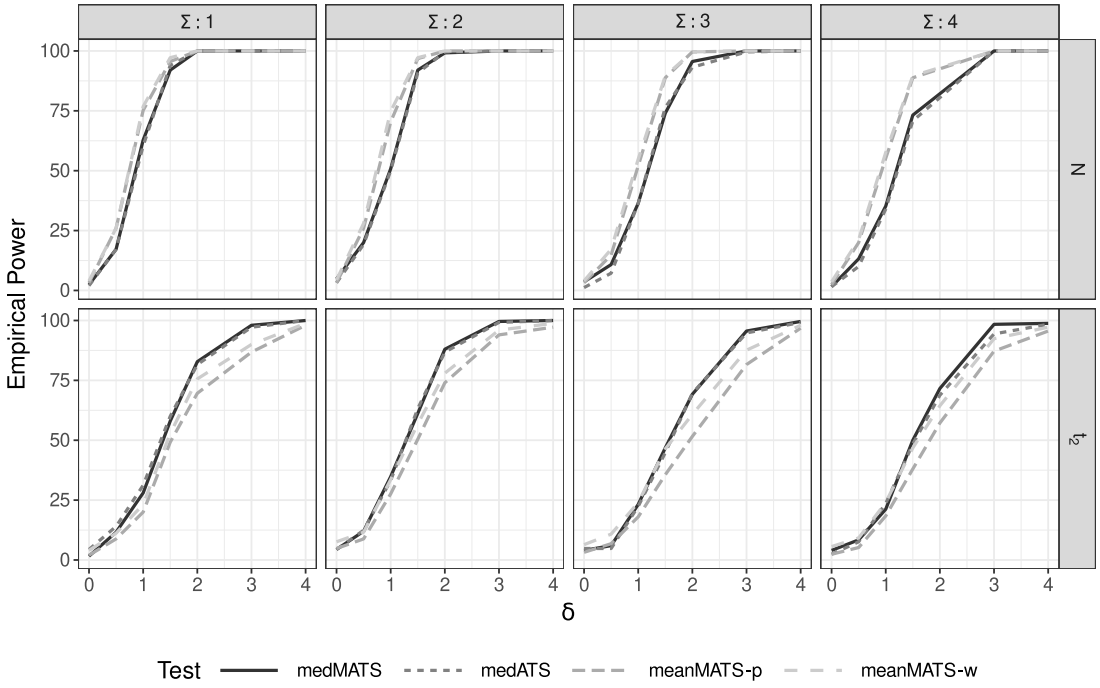

Fig. 3. Empirical Power in Percent of the QMANOVA and the mean-based MATS by Friedrich and Pauly [18] for normal distributed (N) and t_2-distributed (t_2) data, the small sample size (10, 10) and the covariance settings $1-4$.

Other designs. Moreover, we also present type I error simulations study for a two-way layout and for a one-way layout for four groups in the supplement. For the two-way layout, we simulated scenarios for a 2×2 design with dimension $d = 4$ for similar covariance settings. For the one-way layout we used the first four covariance settings, some balanced and unbalanced samples and otherwise the same aspects as in the simulations for two groups. The corresponding findings are similar to the one-layout.

4.2. Power

For the power comparison, we restrict to a representative subset of the settings from the previous section, namely to $d = 4$, the covariance settings (1)–(4), the samples $Kn^{(1)}$ and $Kn^{(3)}$ for $K \in \{2, 4\}$ while we still consider all five distributions. To obtain alternative settings, we shift the data from the first group by $\delta \in \{0.5, 1, 1.5, 2, 3, 4\}$, i.e., $\mathbf{X}_{1j} = \delta + \left((\boldsymbol{\Sigma}^{(1)})^{\frac{1}{2}}\mathbf{e}_{1j} - \mathbf{M}_{1j}\right), j \in \{1, \ldots, n_i\}$ Moreover, we restrict to a representative subset of the settings, namely to $d = 4$, the covariance settings (1)–(4), the samples $Kn^{(1)}$ and $Kn^{(3)}$ for $K \in \{2, 4\}$ while we still consider all five distributions. We again compare the two favourable QMANOVA methods with the mean-based MATS for symmetric distributions. Fig. 3 includes the empirical power of the four methods for the normal and the t_2-distribution and all simulated covariance settings. Studying Fig. 3 one can observe, that the mean-based tests are more powerful in the cases with the normal distribution, but for the t_2-distribution we can see the exact opposite. This observation fits to the power simulation results in Ditzhaus et al. [10]. There, an explanation for this is also given: mean and median are as location estimators asymptotically different efficient in the distributional scenarios. The sample median is the better location estimator in case of heavy-tailed data like the t_2-distribution, but for normal distributed data the situation is reversed. For all distributions the power increases faster for bigger sample sizes and again, the unbalanced designs do not seem to have any influence on this.

5. Illustrative data analysis

To illustrate the new methods on real data, we re-analyse the Egyptian skulls data set from Everitt and Hothorn [13] available in the R-Package `HSAUR` [14]. For 90 skulls, there are four variables ($d = 4$) measured in mm and denoted by mb (maximal breadth), bh (basibregmatic height), bl (basialveolar length) and nh (nasal height). The skulls can be divided into three groups ($k = 3$) which are characterized by time periods in years around 4000 BC ($i = 1$), around 3300 BC ($i = 2$) and around 1850 BC ($i = 3$) [37]. The data is balanced with 30 skulls per group [14]. All four measurements together characterize the skulls in their basic shape. We are interested in inferring whether there are differences between

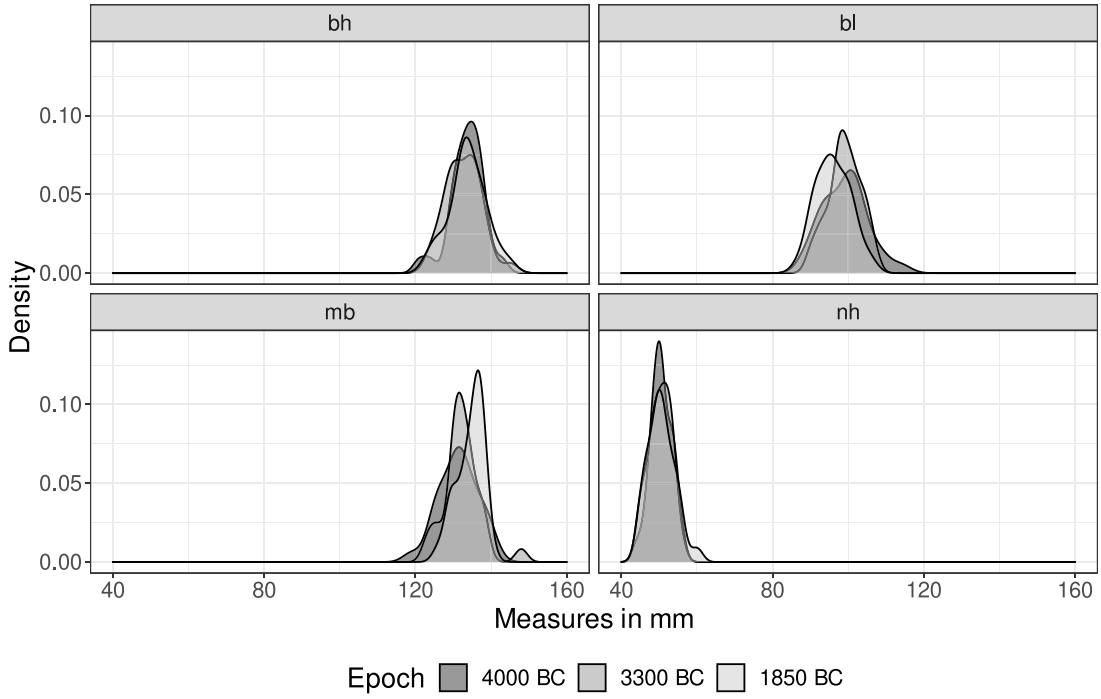

Fig. 4. Kernel density estimates [34] based on Gaussian kernels of the four ($d = 4$) characteristics of the skulls mb (maximal breadth), bh (basibregmatic height), bl (basialveolar length) and nh (nasal height) divided by three epochs ($k = 3$) around 4000 BC, around 3300 BC and around 1850 BC. The bandwidth of the kernel densities is chosen by Silverman's rule-of-thumb [45, Eq. 3.31].

Table 2
p-values for the different null hypotheses and the tests *medMATS* (quantile-based MATS with bootstrap covariance estimator) and *meanMATS-w* (mean-based MATS with wild bootstrap resampling).

Test	Hypothesis			
	$\mathcal{H}_0^{123}$: $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{m}_3$	$\mathcal{H}_0^{12}$: $\mathbf{m}_1 = \mathbf{m}_2$	$\mathcal{H}_0^{23}$: $\mathbf{m}_2 = \mathbf{m}_3$	$\mathcal{H}_0^{13}$: $\mathbf{m}_1 = \mathbf{m}_3$
medMATS	0.038*	0.726	0.042*	0.007*
meanMATS-w	0.062	0.756	0.061	0.038

The *p*-values marked with a * indicate the rejected hypotheses at multiple level 5%.

the three epochs. Fig. 4 shows kernel density plots for each univariate measurement. We observe that the data is rather heavy-tailed and potentially heteroscedastic in each of the four measures. Thus, a median-based approach is reasonable leading to the multivariate null hypothesis $\mathcal{H}_0^{123} : \mathbf{m}_1 = \mathbf{m}_2 = \mathbf{m}_3$. Due to its convincing type I error control in our simulation study, we choose the quantile-based MATS combined with the bootstrap covariance estimator and compare it with the mean-based MATS [18] using wild bootstrap critical values. Similar to the simulation study, all tests are computed based upon 2000 bootstrap iterations. The resulting *p*-values are given in Table 2.

It can be seen, that the *medMATS* rejects the null hypothesis at significance level 0.05 whereas the mean-based *meanMATS-w* does not. A probable reason for this is that the data is nearly symmetric and heavy-tailed in all dimensions (cf. Fig. 4). In such settings the *medMATS* exhibit a better power performance compared to *meanMATS-w*. After rejecting this global null hypothesis, it is intuitive to perform group-wise post hoc analyses. Consequently, one can formulate all pairs hypotheses: $\mathcal{H}_0^{12} : \mathbf{m}_1 = \mathbf{m}_2$, $\mathcal{H}_0^{23} : \mathbf{m}_2 = \mathbf{m}_3$ and $\mathcal{H}_0^{13} : \mathbf{m}_1 = \mathbf{m}_3$. The tests' p-values are displayed in Table 2. Altogether, the considered hypotheses form a closed testing procedure [19] and thus do not need adjustment. That is why the multiple *meanMATS-w* and *medMATS* tests control the family-wise-error-rate without an extra adjustment of the *p*-values. We obtain different test results from both methods: The multiple *medMATS* test rejects the global null and detects a difference in the 3300 BC and the 1850 BC skulls at the 5% level. In comparison, the multiple mean-based *meanMATS-w* test does not reject the global null hypothesis $\mathcal{H}_0^{123}$ at the 5% level. Thus, it would also not detect the difference between the 3300 BC and the 1850 BC skulls as no pairwise posthoc comparisons would have been performed.

6. Conclusion and outlook

We have introduced six statistical tests in a general MANOVA-set-up (QMANOVA) regarding marginal quantiles. These are based on different test statistics: an ANOVA-type statistic (ATS) and a modified ATS (MATS), both in combination with three different covariance estimators (based on a kernel, bootstrap and interval-based approach). All statistics are quadratic forms in the normalized vector of pooled quantiles and can be seen as a generalization of the univariate QANOVA presented in Ditzhaus et al. [10]. As all test statistics are no asymptotic pivots, we propose a non-parametric bootstrap approach to calculate critical values. We analyse the corresponding limit behaviour and prove that the resulting tests are asymptotic exact and consistent. In fact, the methods are asymptotically valid in general factorial designs and do not postulate homoscedasticity or a specific distribution. In an extensive simulation study focusing on the median, it turned out that the MATS with the bootstrap covariance estimator performs the best among the six proposed QMANOVA methods. In particular, it is robust against various aspects of data and performs well on heavy-tailed and skewed data with equal, unequal and singular covariance structures. We additionally compared its performance with the mean-based MATS proposed by Friedrich and Pauly [18] as benchmark method. In-line with theoretical properties of means and medians, our power simulation study showed that the median-based QMANOVA performs better than the corresponding mean-based approach in terms of power. This has also been confirmed in an illustrative data analyses with symmetric and heavy-tailed data in a three-way MANOVA setting. The test results suggest that using the QMANOVA instead of an mean-based method is an added value in this application.

Apart from the illustrative data analyses, the focus of the paper was on deriving global test procedures. Having rejected a global null, post-hoc analyses on the components or factor levels are of interest and multiplicity may become an issue. In the three-way MANOVA setting of the data example this was no issue. But it would be for more complex situations. Thus, we plan to derive multiple contrast tests (MCTs) for contrasts of marginal medians and quantiles in the future. Here, concepts from Gunawardana and Konietschke [21] could be adapted. As the derived methods can directly be inverted in confidence regions we would also like to derive simultaneous confidence regions and intervals that are compatible to the MCTs decisions. This would allow a deeper insight into the behaviour of the estimates. Moreover, similar to mean-based MANCOVA [55], we plan to derive QMANCOVAs that allow for covariate adjustments. Finally, as suggested by a reviewer, studying other quantiles such as equi-coordinate quantiles within a similar MANOVA setting is of future interest.

CRediT authorship contribution statement

Marléne Baumeister: Writing – original draft, Software, Methodology, Formal analysis, Visualization. **Marc Ditzhaus:** Conceptualization, Writing – review & editing, Supervision, Project administration. **Markus Pauly:** Conceptualization, Writing – review & editing, Supervision, Project administration.

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as project 271512359. Furthermore, the authors thank two anonymous reviewers for their comments that substantially improved the paper's quality. We thank Dr. Paavo Sattler for his helpful comments on the statistical inference and his improvement of the simulations code.

Appendix A. Technical details and proofs

In the following Section we present technical details of the test construction and the proofs of the paper.

A.1. Mathematical foundation of Proposition 1 through empirical processes

To get a complete picture and to present consistent proofs we show a central limit theorem for the marginal empirical distribution functions and its foundation with empirical processes. Furthermore, this has the advantage, that we can construct the bootstrap procedure from it. First, we have to consider the set of group-specific marginal empirical distribution functions:

$$\begin{pmatrix} \hat{F}_{i1}(t_1) \\ \vdots \\ \hat{F}_{id}(t_d) \end{pmatrix} = \frac{1}{n_i}\sum_{j=1}^{n_i} \begin{pmatrix} \mathbf{1}_{(-\infty,t_1]}(X_{ij1}) \\ \vdots \\ \mathbf{1}_{(-\infty,t_d]}(X_{ijd}) \end{pmatrix} = \frac{1}{n_i}\sum_{j=1}^{n_i} \begin{pmatrix} \mathbf{1}_{(-\infty,t_1]\times\mathbb{R}^{d-1}}(X_{ij1},\dots,X_{ijd}) \\ \vdots \\ \mathbf{1}_{\mathbb{R}^{d-1}\times(-\infty,t_d]}(X_{ij1},\dots,X_{ijd}) \end{pmatrix}. \tag{A.1}$$

For $t \in \mathbb{R}$ and $\ell, r \in \{1, \dots, d\}$, set $A_{t\ell r} = (-\infty, t]$ if $\ell = r$ and $A_{t\ell r} = \mathbb{R}$ else. Then, the functions

$$\mathbf{1}_{\bigotimes_{r=1}^{d} A_{t\ell r}} : \mathbb{R}^d \to \mathbb{R},$$
$$(x_1, \dots, x_d)^T \mapsto \mathbf{1}_{\bigotimes_{r=1}^{d} A_{t\ell r}}(x_1, \dots, x_d) = \begin{cases} 1, & x_\ell \in A_{t\ell\ell} = (-\infty, t], \\ 0, & x_\ell \notin A_{t\ell\ell} = (-\infty, t], \end{cases}$$

characterize the marginal empirical distribution functions and they form a set:

$$\mathcal{E} := \left\{ \mathbf{1}_{\bigotimes_{r=1}^{d} A_{t\ell r}} | A_{t\ell\ell} = (-\infty, t] \wedge A_{t\ell r} = \mathbb{R} : r \neq \ell;\ r, \ell \in \{1, \dots, d\} \wedge t \in \mathbb{R} \right\}.$$

Lemma 1. *The set of measurable functions $\mathcal{E}$ is a VC-class.*

Proof. Due to Problem 9 in Section 2.6 of van der Vaart and Wellner [49, p. 151] it remains to show that

$$\mathcal{C} = \left\{ \bigotimes_{r=1}^{d} A_{t\ell r} | A_{t\ell\ell} = (-\infty, t] \wedge A_{t\ell r} = \mathbb{R} :\ r \neq \ell \in \{1, \dots, d\} \wedge t \in \mathbb{R} \right\}$$

is a VC-class. We use the fact that the d-dimensional cells

$$\mathcal{D} = \left\{ \bigotimes_{r=1}^{d} A_r | A_r = (-\infty, t_r] : r \in \{1, \dots, d\} \wedge t_r \in \mathbb{R} \right\}$$

form a VC-class with VC-index $d + 1$ [cf. 49, Ex. 2.6.1]. To prove that $\mathcal{C}$ is a VC-class, let us suppose for a moment that $\mathcal{C}$ shatters the subset $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^d$ for some $m \in \mathbb{N}$, e.g. every subset $\mathbf{X}$ of $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ can be written as an intersection between $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and an element $C \in \mathcal{C}$: $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \cap C$. Define $M = \max\{x_{sr} | s \in \{1, \dots, m\};\ r \in \{1, \dots, d\}\} + 1$. Thus, M is larger than any component of $\mathbf{x}_1, \dots, \mathbf{x}_m$. Then it is clear that

$$\mathcal{C}' = \left\{ \bigotimes_{r=1}^{d} A_{t\ell r} | A_{t\ell\ell} = (-\infty, t] \wedge A_{tr\ell} = (-\infty, M] : r \neq \ell \in \{1, \dots, d\} \wedge t \in \mathbb{R} \right\}$$

shatters $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^d$ as well. Since $\mathcal{C}'$ is clearly a subset of $\mathcal{D}$, $\mathcal{C}'$ is for $m \leq d+1$ also a VC-class with VC-index $d+1$ or smaller. This is a contradiction. Thus, $\mathcal{C}$ does not shatter the subset $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^d$ and it can finally be deduced that $\mathcal{C}$ is a VC-class with a VC-index smaller or equal to $d + 1$. □

Categorizes the set $\mathcal{E}$ as a VC-Class amounts to apply the theory of empirical processes on the marginal empirical distribution functions and yields the central limit theorem about them:

Proposition 4. *The Skorokhod Space $D(\mathbb{R})$ contains all right-continuous functions $G : \mathbb{R} \to \mathbb{R}$ with left limits [49, p. 3]. Then, we have convergence in distribution in $D(\mathbb{R})^d$:*

$$\sqrt{n_i} \left(\hat{F}_{i\ell} - F_{i\ell} \right)_{\ell \in \{1, \dots, d\}} \xrightarrow{d} \mathbb{G}_i \text{ in } D(\mathbb{R})^d,\ i \in \{1, \dots, k\}. \tag{A.2}$$

Proof. Applying Theorem 2.6.7 in van der Vaart and Wellner [49], Lemma 1 yields that $\mathcal{E}$ satisfies the uniform entropy condition (2.5.1) in van der Vaart and Wellner [49, p. 127]. The function

$$\mathbf{1}_{\mathbb{R}^d} : \mathbb{R}^d \to \mathbb{R},\ (x_1, \dots, x_d) \mapsto 1,$$

is an envelope function for $\mathcal{E}$ and it holds for every distribution P on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$, that $\|\mathbf{1}_{\mathbb{R}^d}\|_2 < \infty$. From these three conditions it can be concluded that $\mathcal{E}$ is P-Donsker [49, p. 141]. Consequently:

$$\sqrt{n_i} \left(\hat{F}_{i\ell} - F_{i\ell} \right)_{\ell \in \{1, \dots, d\}} \xrightarrow{d} \mathbb{G}_i \text{ in } \ell^\infty(\mathcal{E}),\ i \in \{1, \dots, k\}.$$

By construction the limit process $\mathbb{G}_i : \mathcal{E} \to \mathbb{R}$, $f \mapsto \mathbb{G}_i f$ is a empirical process in $\ell^\infty(\mathcal{E})$ for every $i \in \{1, \dots, k\}$ and it is furthermore a zero-mean Gaussian process [49, p. 81 f.] with covariance

$$\mathrm{Cov}(\mathbb{G}_i f_{t\ell}, \mathbb{G}_i f_{sm}) = \begin{cases} F_{i\ell}(\min(t, s)) - F_{i\ell}(s) F_{i\ell}(t), & \ell = m, \\ \mathbf{F}_{i\ell m}(s, t) - F_{i\ell}(t) F_{im}(s), & \ell \neq m, \end{cases}$$

A calculation of this can be found in the supplement. By Lemma 1.5.3 in van der Vaart and Wellner [49] this characterizes $\mathbb{G}_i$ in $\ell^\infty(\mathcal{E})$ completely. For a fixed $\ell \in \{1, \dots, d\}$ any function $f_{t\ell} \in \mathcal{E}$ can be identified with $t \in \mathbb{R}$, and thus, $\ell^\infty(\mathcal{E})$ with $[\ell^\infty(\mathbb{R})]^d$ [49, Example 2.1.3]. When the space is equipped with the supremum norm $\|G\| = \sup_{t \in \mathbb{R}} |G(t)|$, the weak convergence in $D(\mathbb{R})^d$ follows from the weak convergence in $[\ell^\infty(\mathbb{R})]^d$ [49, Example 2.1.3]. □

A.2.

Proof of Proposition 1. In contrast to the proof of Babu and Rao [1], this one works with empirical processes. Consider A.1 for the technical foundation. Applying the delta-method for metrizable topological vector spaces [cf. 49, Theorem 3.9.4] to Eq. (A.2) yields the assertion. Let $\mathbb{D} = \{G : \mathbb{R} \to \mathbb{R} \,|\text{nondecreasing, right-continuous}\}$, it holds $\mathbb{D} \subset D(\mathbb{R})$ [cf. 10, Supplement, p. 10]. The function applied in the delta-method is the inverse mapping [cf. 49, p. 385]:

$$\phi_p : \mathbb{D} \subset D(\mathbb{R}) \to \mathbb{R},\ \phi_p(G) = G^{-1}(p) = \inf\{t \in \mathbb{R} | G(t) \geq p\},\ p \in (0,1). \tag{A.3}$$

For a fixed $\ell \in \{1, \dots, d\}$ the function $F_{i\ell}$ is in $\mathbb{D}$ and by Assumption 2 of the paper $F_{i\ell}$ is differentiable at $q_{i\ell} = F_{i\ell}^{-1}(p)$ with positive derivative $f_{i\ell}(q_{i\ell})$. Thus, by Lemma 3.9.20 of van der Vaart and Wellner [49, p. 385] the function ϕ_p is for every $p \in (0,1)$ Hadamard differentiable at $F_{i\ell}$ tangentially to the space $\mathbb{D}_{q_{i\ell}} \subset D(\mathbb{R})$, which contains all function $\alpha \in \mathbb{D}(\mathbb{R})$ that are continuous at $q_{i\ell}$. The Hadamard derivative is calculated by

$$\phi'_{p,F_{i\ell}}(\alpha) = -\frac{\alpha(q_{i\ell})}{f_{i\ell}(q_{i\ell})}.$$

It holds $\sqrt{n_i} \xrightarrow{n_i \to \infty} \infty$ and $\mathbb{G}_i$ is separable. Otherwise the separable version of $\mathbb{G}_i$ is chosen as described in van der Vaart and Wellner [49, Section 2.2.3]. Thus, the requirements of Theorem 3.9.4 in van der Vaart and Wellner [49] are fulfilled and it follows that

$$\sqrt{n_i}\left(\phi_p(\hat{F}_{i\ell}) - \phi_p(F_{i\ell})\right) = \sqrt{n_i}\left(\hat{F}_{i\ell}^{-1}(p) - F_{i\ell}^{-1}(p)\right) = \sqrt{n_i}\left(\hat{q}_{i\ell} - q_{i\ell}\right)$$

converges for every $\ell \in \{1, \dots, d\}$ weakly to

$$\phi'_{F_{i\ell}}(\mathbb{G}_i) = -\frac{\mathbb{G}_i(q_{i\ell})}{f_{i\ell}(q_{i\ell})}.$$

For $a, b \in \{1, \dots, d\}$ it holds

$$\mathrm{E}\left(-\frac{\mathbb{G}_i(q_{ia})}{f_{ia}(q_{ia})}\right) = -\frac{1}{f_{ia}(q_{ia})} E\left(\mathbb{G}_i(q_{ia})\right) = 0$$

and

$$\begin{aligned}
\mathrm{Cov}\left(-\frac{\mathbb{G}_i(q_{ia})}{f_{ia}(q_{ia})}, -\frac{\mathbb{G}_i(q_{ib})}{f_{ib}(q_{ib})}\right) &= \frac{1}{f_{ia}(q_{ia}) f_{ib}(q_{ib})} \mathrm{Cov}\left(\mathbb{G}_i(q_{ia}), \mathbb{G}_i(q_{ib})\right) \\
&= \begin{cases} \dfrac{1}{f_{ia}^2(q_{ia})}\left[F_{ia}\left(\min(q_{ia}, q_{ia})\right) - F_{ia}(q_{ia}) F_{ia}(q_{ia})\right], & a = b, \\ \dfrac{1}{f_{ia}(q_{ia}) f_{ib}(q_{ib})}\left[\mathbf{F}_{iab}(q_{ia}, q_{ib}) - F_{ia}(q_{ia}) F_{ib}(q_{ib})\right], & a \neq b, \end{cases} \\
&= \begin{cases} \dfrac{1}{f_{ia}^2(q_{ia})}\left[F_{ia}\left(F_{ia}^{-1}(p)\right) - F_{ia}\left(F_{ia}^{-1}(p)\right) F_{ia}\left(F_{ia}^{-1}(p)\right)\right], & a = b, \\ \dfrac{1}{f_{ia}(q_{ia}) f_{ib}(q_{ib})}\left[\mathbf{F}_{iab}\left(F_{ia}^{-1}(p), F_{ib}^{-1}(p)\right) - F_{ia}\left(F_{ia}^{-1}(p)\right) F_{ib}\left(F_{ib}^{-1}(p)\right)\right], & a \neq b, \end{cases} \\
&= \begin{cases} \dfrac{1}{f_{ia}^2(q_{ia})}\left[p - p^2\right], & a = b, \\ \dfrac{1}{f_{ia}(q_{ia}) f_{ib}(q_{ib})}\left[F_{iab}\left(F_{ia}^{-1}(p), F_{ib}^{-1}(p)\right) - p^2\right], & a \neq b. \end{cases}
\end{aligned} \tag{A.4}$$

Consequently, this means $\frac{\sqrt{n_i}}{\sqrt{n}} \sqrt{n}\left(\hat{q}_{i\ell} - q_{i\ell}\right) = \sqrt{n_i}\left(\hat{q}_{i\ell} - q_{i\ell}\right) \xrightarrow{d} \sqrt{\kappa_i}\, \mathbf{Z}_i$ and with the assumption of non-vanishing groups the assertion follows. □

A.3.

Proof of Proposition 2. The estimator $\Xi(\mathbf{T}, \hat{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{T}\hat{\mathbf{\Sigma}}\mathbf{T})^{-1}\mathbf{I}_{dk}$ is consistent for $\mathrm{tr}\left(\mathbf{T}\mathbf{\Sigma}\mathbf{T}\right)^{-1}\mathbf{I}_{dk}$ as a continuous function of the consistent estimator $\hat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$. Instead of the classical inverse, the Moore–Penrose inverse is not a continuous function. That is why there is more to do to prove the consistency of $\Xi(\mathbf{T}, \hat{\mathbf{\Sigma}}) = (\mathbf{T}\hat{\mathbf{\Sigma}}_0\mathbf{T})^{+}$. The consistency of $\hat{\mathbf{\Sigma}}_0$ follows from the consistency of $\hat{\mathbf{\Sigma}}$. And by the Continuous Mapping Theorem [49, Thm. 1.11.1] it holds $\mathbf{T}\hat{\mathbf{\Sigma}}_0\mathbf{T} \xrightarrow{P} \mathbf{T}\mathbf{\Sigma}_0\mathbf{T}$ and $\hat{\mathbf{\Sigma}}_0$ has full rank. Consequently, there is no rank jump in $\mathbf{T}\hat{\mathbf{\Sigma}}_0\mathbf{T}$ and from Theorem 4.2 in Rakočević [41] it follows the assertion. □

A.4.

Proof of Theorem 1.

(i) The statement from Proposition 1 is for every $i \in \{1, \dots, k\}$:

$$\sqrt{n}\left(\hat{\mathbf{q}}_i - \mathbf{q}_i\right) \xrightarrow{d} \mathbf{Z}_i \sim \mathcal{N}\left(\mathbf{0}_d, \mathbf{\Sigma}_i\right)$$

and from the independent distributed groups $i \in \{1, \dots, k\}$ it follows

$$\sqrt{n}\left(\hat{\mathbf{q}} - \mathbf{q}\right) \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}\left(\mathbf{0}_{dk}, \mathbf{\Sigma}\right).$$

Under $\mathcal{H}_0$ and by the Continuous Mapping Theorem [49, Thm. 1.11.1] the following is also true:

$$\sqrt{n}\mathbf{T}\hat{\mathbf{q}} = \sqrt{n}\left(\mathbf{T}\hat{\mathbf{q}} - \mathbf{T}\mathbf{q}\right) = \mathbf{T}\sqrt{n}\left(\hat{\mathbf{q}} - \mathbf{q}\right) \xrightarrow{d} \mathbf{T}\mathbf{Z} \sim \mathcal{N}\left(\mathbf{0}_{dk}, \mathbf{T}\mathbf{\Sigma}\mathbf{T}\right).$$

Therefore, due to the consistency of $\Xi(\mathbf{T}, \hat{\mathbf{\Sigma}})$ and by Slutzky's theorem [cf. 49, Example 1.4.7] it follows

$$S_n(\mathbf{T}) = n(\mathbf{T}\hat{\mathbf{q}})' \Xi\left(\mathbf{T}, \hat{\mathbf{\Sigma}}\right) \mathbf{T}\hat{\mathbf{q}} = \left(\sqrt{n}\mathbf{T}\hat{\mathbf{q}}\right)' \Xi\left(\mathbf{T}, \hat{\mathbf{\Sigma}}\right) \left(\sqrt{n}\mathbf{T}\hat{\mathbf{q}}\right) \xrightarrow{d} (\mathbf{T}\mathbf{Z})' \Xi\left(\mathbf{T}, \mathbf{\Sigma}\right) (\mathbf{T}\mathbf{Z}).$$

By Mathai and Provost [32, S. 90] this has the same distribution as the random variable $\sum_{i=1}^{dk} \lambda_i B_i$ with $B_i \overset{iid}{\sim} \chi^2_1$, $i \in \{1, \dots, dk\}$, and λ_i are the eigenvalues of $(\mathbf{T}\mathbf{\Sigma}\mathbf{T})^{\frac{1}{2}} \Xi\left(\mathbf{T}, \mathbf{\Sigma}\right) (\mathbf{T}\mathbf{\Sigma}\mathbf{T})^{\frac{1}{2}}$.

(ii) Proposition 1 is not restricted to the null hypothesis. Thus, it can be concluded from the Continuous Mapping Theorem [49, Thm. 1.11.1] that $n^{-1}S_n(\mathbf{T})$ always converges in probability to $(\mathbf{T}\mathbf{q})' \Xi\left(\mathbf{T}, \mathbf{\Sigma}\right) \mathbf{T}\mathbf{q}$. That is why it remains to prove that from $\mathcal{H}_1 : \mathbf{T}\mathbf{q} \neq \mathbf{0}_{dk}$ it always follows that $(\mathbf{T}\mathbf{q})' \Xi\left(\mathbf{T}, \mathbf{\Sigma}\right) \mathbf{T}\mathbf{q} > 0$. Let $\mathbf{T}\mathbf{q} \neq \mathbf{0}_{dk}$. We need to consider the two versions of $\Xi\left(\mathbf{T}, \mathbf{\Sigma}\right)$ separately. The proof of the ATS follows immediately with

$$\mathbf{T}\mathbf{q} \neq \mathbf{0}_{dk} \Rightarrow (\mathbf{T}\mathbf{q})'\mathbf{T}\mathbf{q} > 0 \Rightarrow \frac{(\mathbf{T}\mathbf{q})'\mathbf{T}\mathbf{q}}{\operatorname{tr}(\mathbf{T}\mathbf{\Sigma}\mathbf{T})} > 0$$

The proof of the MATS follows analogously to the proof of Theorem 2 in Ditzhaus et al. [10]. The covariance matrix $\mathbf{\Sigma}$ is positive semidefinite and symmetric by definition. Thus, the square root $\mathbf{\Sigma}^{\frac{1}{2}}$ exists and is also positive semidefinite and symmetric [22, Thm. 7.2.6]. As a consequence, the square root $\mathbf{\Sigma}_0^{\frac{1}{2}}$ exists as well. Moreover, there is some $\tilde{\mathbf{q}} \in \mathbb{R}^{dk}$ such that $\mathbf{q} = \mathbf{\Sigma}_0^{\frac{1}{2}} \tilde{\mathbf{q}}$. From the preceding and the non-singularity of $\mathbf{\Sigma}_0$ it follows

$$\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}} \left[\left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{T}\mathbf{q}\right] = \mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}} \left[\left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}} \tilde{\mathbf{q}}\right] = \mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}} \tilde{\mathbf{q}} = \mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}} \left(\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{q} = \mathbf{T}\mathbf{q} \neq \mathbf{0}_{dk}.$$

And with the well known properties $\left(\mathbf{A}'\right)^+ = \left(\mathbf{A}^+\right)'$ and $\left(\mathbf{A}'\mathbf{A}\right)^+ = \mathbf{A}^+\left(\mathbf{A}'\right)^+$ of Moore–Penrose inverses [cf. 42, p. 67] we conclude

$$\begin{aligned}(\mathbf{T}\mathbf{q})' \left(\mathbf{T}\mathbf{\Sigma}_0\mathbf{T}\right)^+ \mathbf{T}\mathbf{q} &= (\mathbf{T}\mathbf{q})' \left[\left(\mathbf{\Sigma}_0^{\frac{1}{2}}\mathbf{T}\right)' \mathbf{\Sigma}_0^{\frac{1}{2}}\mathbf{T}\right]^+ \mathbf{T}\mathbf{q} = (\mathbf{T}\mathbf{q})' \left(\mathbf{\Sigma}_0^{\frac{1}{2}}\mathbf{T}\right)^+ \left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{T}\mathbf{q} = (\mathbf{T}\mathbf{q})' \left[\left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+\right]' \left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{T}\mathbf{q} \\ &= \left[\left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{T}\mathbf{q}\right]' \left[\left(\mathbf{T}\mathbf{\Sigma}_0^{\frac{1}{2}}\right)^+ \mathbf{T}\mathbf{q}\right] > 0. \quad \square\end{aligned}$$

A.5.

Proof of Proposition 3. Even if Babu and Rao [1] proved this, the result can be easily shown with the theory of empirical processes. By Example 2.1.3 in van der Vaart and Wellner [49] the 2-dimensional empirical distribution functions $\mathbf{F}_{i\ell m}$ can be identified with the empirical measure indexed by $\mathcal{F} = \{\mathbf{1}_{(-\infty,t_\ell]\times(-\infty,t_m]} | (t_\ell, t_m) \in \mathbb{R}^2\}$, which forms a Donsker-Class for $i \in \{1, \dots, k\}$, $\ell, m \in \{1, \dots, d\}$. Thus, the set $\mathcal{F}$ is also a Glivenko–Cantelli-Class [cf. 49, Lemma 2.4.5]:

$$\sup_{(t_1,t_2)\in\mathbb{R}^2} \left|\hat{\mathbf{F}}_{i\ell m}(t_1, t_2) - \mathbf{F}_{i\ell m}(t_1, t_2)\right| \xrightarrow{a.e.} 0.$$

From the continuity of $\mathbf{F}_{im\ell}$ at $(q_{i\ell}, q_{im})$ and $(\hat{q}_{i\ell}, \hat{q}_{im}) \xrightarrow{a.e.} (q_{i\ell}, q_{im})$ it follows from this analogous to Babu and Rao [cf. 1, p. 18]:

$$\hat{\mathbf{F}}_{i\ell m}\left(\hat{q}_{i\ell}, \hat{q}_{im}\right) \xrightarrow{a.e.} \mathbf{F}_{i\ell m}(q_{i\ell}, q_{im}).$$

This yields the assertion. $\square$

A.6.

Proof of Theorem 2. The proof is analogous to the proof of Theorem 1. The bootstrap version of (A.2) follows from Theorem 3.6.2, Chapter 1.12 in van der Vaart and Wellner [49] and from Lemma 1:

$$\sqrt{n_i}\left(\hat{F}^*_{i\ell}-\hat{F}_{i\ell}\right)_{\ell\in\{1,\dots,d\}}\xrightarrow{d}\mathbb{G}_i \text{ in } D(\mathbb{R})^d,\ i\in\{1,\dots,k\}, \tag{A.5}$$

given the data in probability. Here, $\hat{F}^*_{i\ell}$ describes the empirical distribution function calculated with the bootstrap sample, e.g. $\hat{F}^*_{i\ell}(t)=\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{1}_{\{X^*_{ij\ell}\le t\}}$. The delta-method for bootstrapping [49, Theorem 3.9.11] yields analogous to Proposition 1 under Assumption 1:

$$\sqrt{n}\left(\hat{q}^*_{i\ell}-\hat{q}_{i\ell}\right)_{\ell\in\{1,\dots,d\}}\xrightarrow{d}\mathbf{Z}_i,\ i\in\{1,\dots,k\}, \tag{A.6}$$

given the data in probability, where $\hat{q}^*_{i\ell}=F^{*-1}_{i\ell}(p)$ is the empirical bootstrap quantile and $\mathbf{Z}_i$ is as in Proposition 1. As a result, the bootstrap version of the central limit theorem gives us the same limit process $\mathbf{Z}_i$, $i\in\{1,\dots,k\}$ as the regular one. That is why the covariance of the limit process is again $\mathbf{\Sigma}=\bigoplus_{i=1}^k\mathbf{\Sigma}^{(i)}$ and it is needed to estimate $\mathbf{\Sigma}$ with the bootstrap sample. By (A.6) it holds:

$$\sqrt{n}\left(\hat{\mathbf{q}}^*-\hat{\mathbf{q}}\right)\xrightarrow{d}\mathbf{Z}\sim\mathcal{N}\left(\mathbf{0}_{dk},\mathbf{\Sigma}\right).$$

From the Continuous Mapping Theorem [49, Thm. 1.11.1] it follows

$$\sqrt{n}\mathbf{T}\left(\hat{\mathbf{q}}^*-\hat{\mathbf{q}}\right)\xrightarrow{d}\mathbf{T}\mathbf{Z}\sim\mathcal{N}\left(\mathbf{0}_{dk},\mathbf{T}\mathbf{\Sigma}\mathbf{T}\right).$$

The consistency of $\Xi(\mathbf{T},\hat{\mathbf{\Sigma}}^*)$ follows from Proposition 2. Identical to the proof of Theorem 1, this yields the assertion. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2023.105246.

References

[1] G. Babu, C. Rao, Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population, J. Multivariate Anal. 27 (1) (1988) 15–23.
[2] M.S. Bartlett, A note on tests of significance in multivariate analysis, Math. Proc. Camb. Phil. Soc. 35 (2) (1939) 180–185.
[3] A.C. Bathke, S. Friedrich, M. Pauly, F. Konietschke, W. Staffen, N. Strobl, Y. Höller, Testing mean differences among groups: Multivariate and repeated measures analysis with minimal assumptions, Multivar. Behav. Res. 53 (3) (2018) 348–359.
[4] C. Becker, R. Fried, S. Kuhnt (Eds.), Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather, first ed., Springer, New York, 2013.
[5] M. Boerner, O. Krüger, Aggression and fitness differences between plumage morphs in the common buzzard (Buteo Buteo), Behav. Ecol. 20 (1) (2009) 180–185.
[6] D.G. Bonett, Confidence interval for a coefficient of quartile variation, Comput. Statist. Data Anal. 50 (11) (2006) 2953–2957.
[7] E. Brunner, H. Dette, A. Munk, Box-Type approximations in nonparametric factorial designs, J. Amer. Statist. Assoc. 92 (440) (1997) 1494–1502.
[8] E. Chung, J.P. Romano, Exact and asymptotically robust permutation tests, Ann. Statist. 41 (2) (2013) 484–507.
[9] A.P. Dempster, A significance test for the separation of two highly multivariate small samples, Biometrics 16 (1) (1960) 41–50.
[10] M. Ditzhaus, R. Fried, M. Pauly, QANOVA: Quantile-based permutation methods for general factorial designs, TEST 30 (4) (2021) 960–979.
[11] D. Dobler, S. Friedrich, M. Pauly, Nonparametric MANOVA in meaningful effects, Ann. Inst. Statist. Math. 72 (4) (2020) 997–1022.
[12] B. Efron, Bootstrap methods: Another look at the Jackknife, Ann. Statist. 7 (1) (1979) 1–26.
[13] B. Everitt, T. Hothorn, A Handbook of Statistical Analyses using R, Chapman & Hall/CRC, Boca Raton, 2006.
[14] B.S. Everitt, T. Hothorn, HSAUR: A Handbook of Statistical Analyses Using R (1st Edition), 2022.
[15] S. Friedrich, F. Konietschke, M. Pauly, A wild bootstrap approach for nonparametric repeated measurements, Comput. Statist. Data Anal. 113 (2017) 38–52.
[16] S. Friedrich, F. Konietschke, M. Pauly, Resampling-based analysis of multivariate data and repeated measures designs with the R package MANOVA.RM, R J. 11 (2) (2019) 380.
[17] S. Friedrich, F. Konietschke, M. Pauly, MANOVA.RM: Resampling-based analysis of multivariate data and repeated measures designs, 2022, R package version 0.5.3.
[18] S. Friedrich, M. Pauly, MATS: Inference for potentially singular and heteroscedastic MANOVA, J. Multivariate Anal. 165 (2018) 166–179.
[19] K.R. Gabriel, Simultaneous test procedures–some theory of multiple comparisons, Ann. Math. Stat. 40 (1) (1969) 224–250.
[20] M. Ghosh, W.C. Parr, K. Singh, G.J. Babu, A note on bootstrapping the sample median, Ann. Statist. 12 (3) (1984) 1130–1135.
[21] A. Gunawardana, F. Konietschke, Nonparametric multiple contrast tests for general multivariate factorial designs, J. Multivariate Anal. 173 (2019) 165–180.
[22] R.A. Horn, C.R. Johnson, Matrix Analysis, second ed., Cambridge University Press, New York, NY, 2013.
[23] A. Janssen, T. Pauls, How do bootstrap and permutation tests work? Ann. Statist. 31 (3) (2003) 768–806.
[24] R. Koenker, K.F. Hallock, Quantile regression, J. Econ. Perspect. 15 (4) (2001) 143–156.
[25] R. Koenker, J.A.F. Machado, Goodness of fit and related inference processes for quantile regression, J. Amer. Statist. Assoc. 94 (448) (1999) 1296–1310.
[26] F. Konietschke, A.C. Bathke, S.W. Harrar, M. Pauly, Parametric and nonparametric bootstrap methods for general MANOVA, J. Multivariate Anal. 140 (2015) 291–301.

[27] K. Krishnamoorthy, F. Lu, A parametric bootstrap solution to the MANOVA under heteroscedasticity, J. Stat. Comput. Simul. 80 (8) (2010) 873–887.
[28] D.N. Lawley, A generalization of Fisher's z test, Biometrika 30 (1/2) (1938) 180–187.
[29] T. Liu, M. Ditzhaus, J. Xu, A resampling-based test for two crossing survival curves, Pharmac. Stat. 19 (4) (2020) 399–409.
[30] J.S. Maritz, R.G. Jarrett, A note on estimating the variance of the sample median, J. Amer. Statist. Assoc. 73 (361) (1978) 194–196.
[31] R.A. Maronna, R.D. Martin, V.J. Yohai, D. Martin, Robust statistics: Theory and methods, Reprinted with corr, in: Wiley Series in Probability and Statistics, Wiley, Chichester Weinheim, 2006.
[32] A.M. Mathai, S.B. Provost, Quadratic Forms in Random Variables: Theory and Applications, in: Statistics, Textbooks and Monographs, (v. 126) M. Dekker, New York, 1992.
[33] J.W. McKean, R.M. Schrader, A comparison of methods for studentizing the sample median, Comm. Statist. Simulation Comput. 13 (6) (1984) 751–773.
[34] É.A. Nadaraya, On non-parametric estimates of density functions and regression curves, Theory Probab. Appl. 10 (1) (1965) 186–190.
[35] K. Nordhausen, H. Oja, Multivariate L1 methods: The package MNM, J. Stat. Softw. 43 (5) (2011) 1–28.
[36] H. Oja, Descriptive statistics for multivariate distributions, Statist. Probab. Lett. 1 (6) (1983) 327–332.
[37] H. Oja, Multivariate nonparametric methods with R, in: Lecture Notes in Statistics, vol. 199, Springer New York, New York, NY, 2010.
[38] M. Pauly, E. Brunner, F. Konietschke, Asymptotic permutation tests in general factorial designs, J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (2) (2015) 461–473.
[39] K.C.S. Pillai, Some new test criteria in multivariate analysis, Ann. Math. Stat. 26 (1) (1955) 117–121.
[40] R.M. Price, D.G. Bonett, Estimating the variance of the sample median, J. Stat. Comput. Simul. 68 (3) (2001) 295–305.
[41] V. Rakočević, On continuity of the Moore-Penrose and Drazin inverses, Mat. Vesnik 49 (3–4) (1997) 163–172.
[42] C.R. Rao, S.K. Mitra, Generalized inverse of matrices and its applications, in: Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1971.
[43] P. Sattler, A.C. Bathke, M. Pauly, Testing hypotheses about covariance matrices in general MANOVA designs, J. Statist. Plann. Inference 219 (2022) 134–146.
[44] R. Serfling, Quantile functions for multivariate analysis: Approaches and applications, Stat. Neerl. 56 (2) (2002) 214–232.
[45] B.W. Silverman, Density Estimation for Statistics and Data Analysis, in: Monographs on Statistics and Applied Probability, (no. 26) Chapman & Hall/CRC, Boca Raton, 1998.
[46] Ł. Smaga, Inference for general MANOVA based on ANOVA-Type statistic, in: T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, M. Vichi (Eds.), Advanced Studies in Classification and Data Science, in: Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Singapore, 2020, pp. 241–254.
[47] C.G. Small, A survey of multidimensional medians, Int. Stat. Rev. / Rev. Int. de Stat. 58 (3) (1990) 263.
[48] R.C. Team, R: A Language and Environment for Statistical Computing, Technical Report, R Foundation for Statistical Computing, Vienna, Austria, 2021.
[49] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer, New York, 2000.
[50] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Amer. Math. Soc. 54 (3) (1943) 426–482.
[51] R. Warne, A primer on multivariate analysis of variance (MANOVA) for behavioral scientists, Pract. Assess., Res. Eval. 19 (2014) 17.
[52] S.S. Wilks, Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution, Ann. Math. Stat. 17 (3) (1946) 257–281.
[53] J.-T. Zhang, Two-Way MANOVA with unequal cell sizes and unequal cell covariance matrices, Technometrics 53 (4) (2011) 426–439.
[54] J.-T. Zhang, B. Zhou, J. Guo, Linear hypothesis testing in high-dimensional heteroscedastic one-way MANOVA: A normal reference L2-norm based test, J. Multivariate Anal. 187 (2022) 104816.
[55] G. Zimmermann, M. Pauly, A.C. Bathke, Multivariate analysis of covariance with potentially singular covariance matrices and non-normal responses, J. Multivariate Anal. 177 (2020) 104594.

Article 2

Baumeister, M., Munko, M., Gladow, K.-P.,
Ditzhaus, M., Chakarov, N., & Pauly, M. (2025).

Early and Late Buzzards: Comparing Different Approaches for Quantile-Based Multiple Testing in Heavy-Tailed Wildlife Research Data.

Biometrical Journal, 67(4), e70065.
https://doi.org/10.1002/bimj.70065

Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Early and Late Buzzards: Comparing Different Approaches for Quantile-Based Multiple Testing in Heavy-Tailed Wildlife Research Data

Marléne Baumeister[1,2] | Merle Munko[3] | Kai-Philipp Gladow[4] | Marc Ditzhaus[3,†] | Nayden Chakarov[4,5] | Markus Pauly[1,2]

[1]Department of Statistics, TU Dortmund University, Dortmund, Germany | [2]Research Center Trustworthy Data Science and Security, UA Ruhr, Germany | [3]Department of Mathematics, Otto-von-Guericke University Magdeburg, Magdeburg, Germany | [4]Department of Animal Behaviour, Bielefeld University, Bielefeld, Germany | [5]Joint Institute for Individualisation in a Changing Environment (JICE), Bielefeld University and University of Münster, Bielefeld, Germany

Correspondence: Marléne Baumeister (baumeister@statistik.tu-dortmund.de)

Received: 15 October 2024 | **Revised:** 23 April 2025 | **Accepted:** 15 May 2025

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (Numbers 314838170 and 396780709), the Research Center Trustworthy Data Science and Security, and Friedrich–Ebert–Stiftung (FES).

Keywords: Bonferroni adjustment | factorial designs | multiple contrast tests | quantiles | resampling

ABSTRACT

In medical, ecological, and psychological research, there is a need for methods to handle multiple testing, for example, to consider group comparisons with more than two groups. Typical approaches that deal with multiple testing are mean- or variance-based which can be less effective in the context of heavy-tailed and skewed data. Here, the median is the preferred measure of location and the interquartile range (IQR) is an adequate alternative to the variance. Therefore, it may be fruitful to formulate research questions of interest in terms of the median or the IQR. For this reason, we compare different inference approaches for two-sided and noninferiority hypotheses formulated in terms of medians or IQRs in an extensive simulation study. We consider multiple contrast testing procedures combined with a bootstrap method as well as testing procedures with Bonferroni correction. As an example of a multiple testing problem based on heavy-tailed data, we analyze an ecological trait variation in early and late breeding in a medium-sized bird of prey.

1 | Introduction

In this paper, we systematically compare possibilities to handle quantile-based multiple testing procedures in general factorial designs. This comparison is motivated by a testing problem involving the hatching dates of a population of common buzzards (*Buteo buteo*) (cf. Figure 6). In the context of species protection, it is necessary to analyze the behavior of animals to understand how animals deal with environmental change and to develop strategies to protect them effectively (Halupka and Halupka 2017). It is well-known that the hatching dates are influenced by the weather in general (Lehikoinen et al. 2009), but in context of increasing temperatures and weather extreme events due to climate change it is interesting to understand which aspects of

[†]We would like to thank the late Marc Ditzhaus, a bright mind and wonderful collaboration partner, for his contributions to that paper and for bringing us together which has made us the scientists we are today.

Marléne Baumeister and Merle Munko contributed equally to this work.

weather influence the hatching dates in detail. Hence, we want to identify years between 2006 and 2022 with an earlier and a later hatching phenology, which could be used in future studies to compare weather conditions and population characteristics between years with early and late breedings. In the end, this ecological question leads to a noninferiority multiple testing procedure. There is the often observed phenomenon in the context of human or animal behavior that data is skewed and heavy-tailed and therefore substantially deviates from normality. Established multiple testing procedures are mean-based and reach their limitations in case of skewed or heavy-tailed data, because they are sensitive to outliers. Bonett and Price (2002) pointed out: "Every student of introductory statistics is taught that the population median may be more meaningful than the population mean when the distribution is skewed." That is why it can be fruitful to consider quantile-based statistical concepts such as the median or the interquartile range (IQR) instead of mean- or variance-based approaches. Another issue is the consideration of multiplicity, because it is natural to formulate further hypotheses regarding post hoc comparisons after rejecting a global hypothesis (Ruxton and Beauchamp 2008). In factorial designs like ours, multiple hypotheses are of potential interest and inferring all of them lead to the problem of the type I error cumulation.

There are some quantile-based methods for statistical inference. A quantile-based regression was already introduced by Koenker and Bassett (1978) and these methods are available in the R-Package `quantreg` (Koenker 2024). Quantile-based testing of global hypotheses in factorial designs has been successfully developed by Chung and Romano (2013) and Ditzhaus et al. (2021) (univariate), and Chung and Romano (2016) and Baumeister et al. (2024) (multivariate). Moreover, Segbehoe et al. (2022) tackle the multiple testing problem regarding quantiles with the development of quantile-based multiple contrast testing procedures (MCTPs). In general, MCTPs are useful in many situations because they overcome the multiple testing problem by redefining the rejection of the global hypothesis: it is simultaneously rejected if any of the individual comparisons are rejected. The general concept of MCTPs was introduced, for example, in Mukerjee et al. (1987). Furthermore, MCTPs are known to be often more powerful than methods with classical p-value adjustment like the Bonferroni procedure (Bretz et al. 2011; Konietschke et al. 2013). Because of these advantages, there are many different adaptations of MCTPs (e.g., Bretz et al. 2001; Hasler and Hothorn 2008; Konietschke et al. 2013; Hasler 2014; Umlauft et al. 2019; Noguchi et al. 2020; Rubarth et al. 2022).

For our approach, the method of Segbehoe et al. (2022) appears to be most suitable. Among others, they introduced MCTPs regarding differences of quantiles for two-sided statistical hypotheses, but not yet for the noninferiority testing problem. Their method is based on an asymptotic approach, which means that the test statistic is compared with a theoretical quantile of a multivariate distribution, and a bootstrap approach, where the test statistic is compared to an empirical bootstrap quantile. The method is included in the R-Package `mratios` (Djira et al. 2020). Segbehoe et al. (2022) compare the type I error performance of these two approaches in a Monte-Carlo simulation study, which takes into account only scenarios with balanced, predominantly large samples and, most importantly, only three groups. In a multiple pairwise comparison with three groups, however it follows from the closing testing procedure (Marcus et al. 1976) that there is no need to adjust the levels for the local hypotheses of pairwise comparisons to control the family-wise error rate (FWER) if the global hypothesis can be rejected. Because of the simplicity of the closing test for three groups, it is to be expected that every testing procedure that controls the FWER will perform relatively well in this setting. The reason for this is that the closing testing procedure works in principle and set-theoretically for every type of test. See Goeman and Solari (2022) for a discussion of the closing test procedure especially for three groups. Furthermore, the simulation of Segbehoe et al. (2022) does not include a comparison with other multiple testing procedures, for example, with Bonferroni-adjusted multiple tests (Dunn 1961). It is therefore impossible to get a broader overview about the performance of the tests. Because our ecological problem regarding the hatch data contains a much larger number of comparisons the simulation study of Segbehoe et al. (2022) cannot help us to decide if this method is adequate for our problem. More generally, it is not possible to decide for one multiple testing procedure to handle multiple testing problems in skewed and heavy-tailed data. This is our motivation to consider a more comprehensive and competitive simulation study.

In particular, our aim is to compare the performance of different statistical testing procedures that deal with quantile-based multiple hypotheses. Beyond the multiple testing problem, statistical questions do not only arise with two-sided hypotheses, as we have seen in our example with regard to buzzards. To give a broader overview of the methodological possibilities and capabilities of different multiple testing procedures, we study three commonly important versions of hypotheses: noninferiority, two-sided, and equivalence hypotheses. An intuitive way to deal with the multiple testing problem is to define permutation tests in the framework of the QANOVA by Ditzhaus et al. (2021) and to adjust them with the well-known Bonferroni correction (Dunn 1961). It is a new approach to define tests in this framework that can be applied in one- and two-sided testing problems. We also extend the method of Segbehoe et al. (2022) to noninferiority testing. Similar to their work, we consider two ways of deriving critical values: from the asymptotic distribution or via groupwise bootstrapping. Besides, we explain that the considered methods are theoretically valid and their inference works without any restrictive distributional assumption and allows for potential heteroskedasticity. We compare all these quantile-based multiple testing procedures for one- and two-sided hypotheses through an extensive simulation study regarding type I error and power. In particular, we consider various testing problems (Dunnett, Tukey, and Grand Mean), varying sample sizes, distributions, as well as homo- and heteroskedastic settings. Through our comparison, we come to the result that the MCTP methods are not in general superior to the QANOVA permutation approaches with Bonferroni correction regarding empirical FWER control and power.

The paper is structured as follows. We state models and hypotheses in Section 2. Afterward, we introduce different statistical testing procedures (Section 3) including explanations of the asymptotic Bonferroni-adjusted QANOVA tests and their permutational version (Section 3.1, details of the permutation approach in Section A.2 in the Appendix), as well as quantile-based MCTPs and a bootstrap version (Section 3.2). An extensive simulation

study (Section 4) gives an overview of the performances of all methods for various scenarios. Section 5 analyzes the data example with these methods. There, we also explain our motivational data example on buzzards and how it fits to our statistical model. Section 6 concludes the results.

2 | Model and Hypotheses

Suppose we have $k \in \mathbb{N}$ mutually independent samples $X_{i1}, \dots, X_{in_i} \sim F_i, i \in \{1, \dots, k\}$, where F_i are distribution functions. Here, n_i represent the sample sizes per group and $n := \sum_{i=1}^{k} n_i$ denotes the total sample size. To define the quantity of interest, let $0 < p_1 < \dots < p_m < 1$ denote $m \in \mathbb{N}$ probabilities of interest with corresponding quantiles

$$q_{ij} := F_i^{-1}(p_j) = \inf\{u \in \mathbb{R} \mid F_i(u) \geq p_j\},$$
$$i \in \{1, \dots, k\}, j \in \{1, \dots, m\}.$$

The pooled quantile vector is denoted by $\mathbf{q} := (q_{11}, \dots, q_{1m}, q_{21}, \dots, q_{km})'$. For our asymptotic derivations, we need the following assumption throughout this paper.

Assumption 2.1. We assume that F_i is continuously differentiable at $q_{i1}, \dots, q_{im}$ with positive derivatives $f_i(q_{ij}) > 0$ for all $i \in \{1, \dots, k\}, j \in \{1, \dots, m\}$. Moreover, we assume $n_i/n \to \kappa_i > 0$ as $n \to \infty$ for all $i \in \{1, \dots, k\}$.

In practice, the assumption of a continuous derivative at $q_{i1}, \dots, q_{im}$ cannot really be checked because usually neither F_i nor $q_{i1}, \dots, q_{im}$ are known. However, if there are (many) ties in the data, this is at least an indicator that F_i is not continuous and, thus, not differentiable at the tie points making the previous assumption less plausible. Let $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_r]' \in \mathbb{R}^{r \times km}$ denote a matrix of vectors $\mathbf{h}_\ell = (h_{\ell 11}, \dots, h_{\ell 1m}, h_{\ell 21}, \dots, h_{\ell km})' \in \mathbb{R}^{km}$, $\ell \in \{1, \dots, r\}$ with the contrast property $\sum_{i=1}^{k} h_{\ell ij} = 0$ for all $j \in \{1, \dots, m\}$. This contrast property means that only contrasts over the different groups may be considered and is actually also needed in Ditzhaus et al. (2021). The property can easily be checked for a known matrix $\mathbf{H}$ and all examples given below fulfill the contrast property. Moreover, let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_r)' \in \mathbb{R}^r$ denote a vector of constants. Then, we aim to infer the multiple testing problem

$$\mathcal{H}_{0,\ell} : \mathbf{h}_\ell' \mathbf{q} = \epsilon_\ell \text{ vs. } \mathcal{H}_{1,\ell} : \mathbf{h}_\ell' \mathbf{q} \neq \epsilon_\ell, \quad \text{for } \ell \in \{1, \dots, r\}. \tag{1}$$

These hypotheses follow the usual definition of hypotheses that can be answered through multiple contrast tests, see, for example, Hothorn et al. (2008) and Konietschke et al. (2013). In addition, we consider a multiple one-sided noninferiority problem:

$$\mathcal{H}_{0,\ell}^I : \mathbf{h}_\ell' \mathbf{q} \leq \epsilon_\ell \text{ vs. } \mathcal{H}_{1,\ell}^I : \mathbf{h}_\ell' \mathbf{q} > \epsilon_\ell, \quad \text{for } \ell \in \{1, \dots, r\}. \tag{2}$$

The corresponding global hypotheses are given by $\mathcal{H}_0 : \mathbf{Hq} = \boldsymbol{\epsilon}$ and $\mathcal{H}_0^I : \mathbf{Hq} \leq \boldsymbol{\epsilon}$, respectively. Our motivation to consider both types of hypotheses is that they have widely different interpretations despite the methodological similarity. Interpreting noninferiority tests is grounded in another research question than the approach of two-sided tests. Testing noninferiority means that someone has the aim to show that one treatment/group is not unacceptably worse compared to one other group (Schumi and Wittes 2011). What *unacceptably worse* means, is characterized in the vector of constants $\boldsymbol{\epsilon}$. If the directed deviation in $\mathcal{H}_{0,\ell}^I$ is smaller than ϵ_ℓ something seems to be unacceptably worse and $\mathcal{H}_{0,\ell}^I$ is not rejected. The constant can be identified with the maximal directed deviation in which a significant difference or improvement is not indicated. Regarding the testing problem and the question of interest, differences smaller than the constant ϵ_ℓ are not indicated as differences. We refer to Scott (2009) and Schumi and Wittes (2011) for the idea of noninferiority tests and the interpretation and meaning of $\boldsymbol{\epsilon}$. Moreover, within this framework it is possible to infer equivalence hypotheses. To this end, we can adapt the equivalence testing approach of Hauck and Anderson (1984) for quantiles. Let $[-\delta_\ell, \delta_\ell]$ be equivalence intervals for every $\ell \in \{1, \dots, r\}$. Then, the multiple equivalence hypotheses problem has the form:

$$\mathcal{H}_{0,\ell}^E : |\mathbf{h}_\ell' \mathbf{q}| \geq \delta_\ell \text{ vs. } \mathcal{H}_{1,\ell}^E : |\mathbf{h}_\ell' \mathbf{q}| < \delta_\ell, \quad \text{for } \ell \in \{1, \dots, r\}. \tag{3}$$

This hypotheses lead to a TOST procedure (Schuirmann 1987) for quantiles, where the statistical question is answered by two one-sided tests with the halved level of significance. Thus, the methodological treatment of (3) follows from that in (2). We want to point out that it is possible to consider far different statistical questions with similar methodology, but in the following we focus on the two-sided and the noninferiority hypotheses only. Below we give some concrete examples of covered multiple testing problems.

Examples of covered hypotheses.

The hypotheses $\mathcal{H}_0$ and $\mathcal{H}_0^I$ cover various local and multiple testing problems of interest. For a single quantile $\mathbf{q} = (q_1, \dots, q_k)'$, $m = 1$, we can formulate hypotheses that are well-known for vectors of means (cf. Bretz et al. 2011; Konietschke et al. 2013) in terms of medians, quantiles, or more general quantile contrasts. This explicitly includes

i. **All-pairs comparisons for medians**.Choosing $p_1 = 0.5$, $m = 1$, and the Tukey-type (Tukey 1994) matrix as contrast matrix $\mathbf{H}$ leads to the one- and two-sided hypotheses $\mathcal{H}_{0,\ell_1\ell_2} : m_{\ell_1} - m_{\ell_2} = \epsilon_{\ell_1\ell_2}$ and $\mathcal{H}_{0,\ell_1\ell_2}^I : m_{\ell_1} - m_{\ell_2} \leq \epsilon_{\ell_1\ell_2}$, where $\ell_1, \ell_2 \in \{1, \dots, k\}, \ell_1 > \ell_2$ of all-pairs comparisons for medians $m_i := F_i^{-1}(0.5), i \in \{1, \dots, k\}$, in one-way layouts.

ii. **Many-to-one comparisons for medians**. Similarly, choosing the Dunnett-type (Dunnett 1955) matrix gives the one- and two-sided hypotheses $\mathcal{H}_{0,\ell} : m_\ell - m_1 = \epsilon_\ell$ and $\mathcal{H}_{0,\ell}^I : m_\ell - m_1 \leq \epsilon_\ell, \ell \in \{2, \dots, k\}$, of many-to-one comparisons for medians.

iii. **Grand-mean comparisons**. Choosing the Grand-mean-type matrix (Djira and Hothorn 2009) instead leads to the one- and two-sided hypotheses $\mathcal{H}_{0,\ell} : m_\ell - \bar{m} = \epsilon_\ell$ and $\mathcal{H}_{0,\ell}^I : m_\ell - \bar{m} \leq \epsilon_\ell, \ell \in \{1, \dots, k\}$, of median comparisons to the mean $\bar{m} := k^{-1} \sum_{i=1}^{k} m_i$ of all groupwise medians in one-way layouts.

iv. **Multiple testing problems in general quantiles or IQR**. In the above hypotheses, the medians $m_1, \dots, m_k$ can be substituted by any quantile or linear contrast of interest. Thus, we can even infer multiple hypotheses about IQRs $IQR_i := F_i^{-1}(0.75) - F_i^{-1}(0.25)$ leading to hypotheses

of the form $\mathcal{H}_{0,\ell_1\ell_2} : IQR_{\ell_1} - IQR_{\ell_2} = \epsilon_{\ell_1\ell_2}$ and $\mathcal{H}^I_{0,\ell_1\ell_2} : IQR_{\ell_1} - IQR_{\ell_2} \leq \epsilon_{\ell_1\ell_2}, \ell_1, \ell_2 \in \{1,\dots,k\}, \ell_1 > \ell_2$ in the all-pairs comparison setting and similar for the Dunnett- or the Grand-mean-type matrix.

v. **Simultaneous inference for medians and IQRs**. Our test scenario is even more general and also allows for the simultaneous treatment of more than one effect parameter of interest. For example, it would be possible to compare the medians and IQRs simultaneously across the groups by setting $p_1 = 0.25, p_2 = 0.5, p_3 = 0.75, m = 3$ and choosing a hypothesis matrix

$$\mathbf{H} \otimes \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

with $\mathbf{H}$ being one of the Tukey-type, Dunnett-type, or Grand-mean-type matrix, respectively, and $\otimes$ denoting the Kronecker product. Here, the Tukey-type matrix leads to all-pairs comparisons of the medians and IQRs, respectively, with local null hypotheses $\mathcal{H}_{0,\ell_1\ell_2,med} : m_{\ell_1} - m_{\ell_2} = \epsilon_{\ell_1\ell_2,med}$, $\mathcal{H}_{0,\ell_1\ell_2,IQR} : IQR_{\ell_1} - IQR_{\ell_2} = \epsilon_{\ell_1\ell_2,IQR}, \ell_1, \ell_2 \in \{1,\dots,k\}, \ell_1 > \ell_2$ for the two-sided testing problem. If $\boldsymbol{\epsilon}$ is the zero vector, the global null hypothesis that all medians and IQRs are equal is $\mathcal{H}_0 : m_1 = \cdots = m_k, IQR_1 = \cdots = IQR_k$. Analogously, the hypotheses can be formulated for the one-sided testing problem as well as for the Dunnett-type matrix for many-to-one comparisons and the Grand-mean-type matrix for comparisons of the medians and IQRs to the mean of medians and IQRs, respectively.

Even multiple hypotheses on quantiles in more general factorial designs are covered by splitting up indices as in classical ANOVA (Pauly et al. 2015).

3 | Statistical Methods

In the following section, we present four testing procedures that all correspond to the hypotheses in Equations (1) and (2), respectively, and are compared in Section 4 by using simulations. An estimator for the quantile q_{ij} is given by the empirical quantile, that is,

$$\widehat{q}_{ij} := \hat{F}_i^{-1}(p_j),$$

where $\hat{F}_i$ denotes the empirical distribution function. Under Assumption 2.1, Serfling (1980) proved convergence in distribution

$$\sqrt{n}\left(\widehat{q}_{ij} - q_{ij}\right)_{j\in\{1,\dots,m\}} \xrightarrow{d} \mathbf{Z}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(i)}\right) \tag{4}$$

as $n \to \infty$ for all $i \in \{1,\dots,k\}$, where

$$\boldsymbol{\Sigma}^{(i)}_{ab} := \kappa_i^{-1} \frac{1}{f_i(q_{ia}) f_i(q_{ib})} (\min\{p_a, p_b\} - p_a p_b) \tag{5}$$

for all $a, b \in \{1,\dots,m\}$. Let $\boldsymbol{\Sigma} := \bigoplus_{i=1}^k \boldsymbol{\Sigma}^{(i)}$ denote the direct sum (i.e., block diagonal matrix) of the covariance matrices. Since we are also interested in directional hypotheses, we consider the family of test statistics

$$T_n(\mathbf{h}_\ell, \epsilon_\ell) := \sqrt{n} \frac{\mathbf{h}_\ell' \widehat{\mathbf{q}} - \epsilon_\ell}{\sqrt{\mathbf{h}_\ell' \widehat{\boldsymbol{\Sigma}} \mathbf{h}_\ell}}, \quad \ell \in \{1,\dots,r\}, \tag{6}$$

instead of the two-sided QANOVA Wald-type test statistic that was discussed in Ditzhaus et al. (2021). We note that for a single contrast $\mathbf{h}_\ell$, we obtain the QANOVA Wald-type test statistic of Ditzhaus et al. (2021) as $T_n^2(\mathbf{h}_\ell, 0)$.

For appropriate critical values $\widetilde{q}_\ell$, we receive the following test decisions for the *two-sided multiple testing problem*:

i. for each $\ell \in \{1,\dots,r\}$, $\mathcal{H}_{0,\ell}$ is rejected if and only if $|T_n(\mathbf{h}_\ell, \epsilon_\ell)| > \widetilde{q}_\ell$,

ii. the global hypothesis $\mathcal{H}_0 = \bigcap_{\ell=1}^r \mathcal{H}_{0,\ell}$ is rejected if and only if at least one $\mathcal{H}_{0,\ell}$ is rejected, that is, if $\max_{\ell\in\{1,\dots,r\}} |T_n(\mathbf{h}_\ell, \epsilon_\ell)| > \widetilde{q}_\ell$.

Corresponding simultaneous two-sided confidence intervals for $\mathbf{h}_\ell'\mathbf{q}, \ell \in \{1,\dots,r\}$, can be obtained as

$$\left[\mathbf{h}_\ell'\widehat{\mathbf{q}} - \sqrt{\mathbf{h}_\ell'\widehat{\boldsymbol{\Sigma}}\mathbf{h}_\ell}\frac{\widetilde{q}_\ell}{\sqrt{n}}, \mathbf{h}_\ell'\widehat{\mathbf{q}} + \sqrt{\mathbf{h}_\ell'\widehat{\boldsymbol{\Sigma}}\mathbf{h}_\ell}\frac{\widetilde{q}_\ell}{\sqrt{n}}\right], \quad \ell \in \{1,\dots,r\}.$$

Alternatively, there is the ability to formulate simultaneous tests $\mathbf{1}\{|T_n(\mathbf{h}_\ell, \epsilon_\ell)| > \widetilde{q}_\ell\}$ for every Hypothesis $\mathcal{H}_{0,\ell}$, $\ell \in \{1,\dots,r\}$ and a test $\mathbf{1}\{\max_{\ell\in\{1,\dots,r\}} |T_n(\mathbf{h}_\ell, \epsilon_\ell)| > \widetilde{q}_\ell\}$ for the global Hypothesis $\mathcal{H}_0$. Analogously the test decisions for the *noninferiority multiple testing problem* are

i. for each $\ell \in \{1,\dots,r\}$, $\mathcal{H}^I_{0,\ell}$ is rejected if and only if $T_n(\mathbf{h}_\ell, \epsilon_\ell) > q_\ell$,

ii. the global hypothesis $\mathcal{H}^I_0 = \bigcap_{\ell=1}^r \mathcal{H}_{0,\ell}$ is rejected if and only if at least one $\mathcal{H}^I_{0,\ell}$ is rejected, that is, if $\max_{\ell\in\{1,\dots,r\}} (T_n(\mathbf{h}_\ell, \epsilon_\ell)) > q_\ell$

with appropriate critical values q_ℓ and the corresponding simultaneous one-sided confidence intervals for $\mathbf{h}_\ell'\mathbf{q}, \ell \in \{1,\dots,r\}$, are given by

$$\left[\mathbf{h}_\ell'\widehat{\mathbf{q}} - \sqrt{\mathbf{h}_\ell'\widehat{\boldsymbol{\Sigma}}\mathbf{h}_\ell}\frac{q_\ell}{\sqrt{n}}, \infty\right), \quad \ell \in \{1,\dots,r\}. \tag{7}$$

This testing problem can also be formulated in short test notation as $\mathbf{1}\{T_n(\mathbf{h}_\ell, \epsilon_\ell) > q_\ell\}$ for the simultaneous hypotheses $\mathcal{H}^I_{0,\ell}$, $\ell \in \{1,\dots,r\}$, and for the global Hypothesis $\mathcal{H}^I_0$ as $\mathbf{1}\{\max_{\ell\in\{1,\dots,r\}} (T_n(\mathbf{h}_\ell, \epsilon_\ell)) > q_\ell\}$. Note that both testing problems comply with the union-intersection principle introduced by Roy (1953) and that they are in fact quantile-based versions of so-called max-t tests (Bretz et al. 2001).

Of note, an application of a stepwise procedure as the closed-testing procedure (Gabriel 1969), the well-known Holm procedure (Holm 1979), or Shaffer's method (Shaffer 1986) may increase the power of the proposed multiple tests but lacks the obtainment of corresponding simultaneous confidence regions. However, we

focus on multiple testing procedures that come along with corresponding simultaneous confidence intervals in the following. See Pigeot (2000) for a methodologically overview about multiple testing and Gabriel (1969) for the foundation of simultaneous testing procedures.

In order to determine appropriate critical values, we first need to investigate the joint asymptotic behavior of the test statistics. Due to (4), it follows that we have convergence in distribution

$$(T_n(\mathbf{h}_1, \epsilon_1), \dots, T_n(\mathbf{h}_r, \epsilon_r))' \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{DH\Sigma H'D}) \tag{8}$$

as $n \to \infty$ under the null hypotheses in (1), where

$$\mathbf{D} := \operatorname{diag}\Big(\left(\mathbf{h}_1' \mathbf{\Sigma} \mathbf{h}_1\right)^{-1/2}, \dots, \left(\mathbf{h}_r' \mathbf{\Sigma} \mathbf{h}_r\right)^{-1/2} \Big). \tag{9}$$

Note that the covariance matrix in (9) in the limit is a correlation matrix, that is, has a diagonal of ones, and, thus, each test statistic $T_n(\mathbf{h}_\ell, \epsilon_\ell)$ is asymptotically standard normally distributed. Since $\mathbf{\Sigma}$ is usually unknown, the joint limiting distribution is unknown. To get a consistent estimator $\widehat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$, we use three different approaches as discussed in Ditzhaus et al. (2021): a kernel estimator, a bootstrap estimator, and an interval-based approach. In Ditzhaus et al. (2021), there was no clear recommendation for one of them. We thus analyze all of them. The concrete forms are given in Section A.1 in the Appendix. It should be noted that further technical assumptions are needed for the consistency of the covariance estimator $\widehat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$ (see Ditzhaus et al. 2021 for details). With each of the three consistent estimators, we are able to obtain an approximation for the critical values. In the following subsections, we elaborate on different asymptotic- and resampling-based choices of $\widetilde{q}_\ell$ and q_ℓ.

3.1 | Bonferroni-Adjusted QANOVA

Let $\alpha \in (0,1)$ represent the level of significance. An intuitive and well-known method to deal with multiple testing problems is the Bonferroni correction (Dunn 1961), where each individual hypothesis is tested at a smaller local level of α/r. To realize this for our statistical question, recall that $T_n(\mathbf{h}_\ell, \epsilon_\ell)$ is asymptotically standard normal distributed. This motivates to consider standard normal quantiles as critical values. Let z_β denote the β-quantile of the standard normal distribution. Then, choosing $\widetilde{q}_\ell = z_{1-\alpha/(2r)}$ for the two-sided multiple testing problem or $q_\ell = z_{1-\alpha/r}$ for the noninferiority multiple testing problem, respectively, yield the *Bonferroni-adjusted asymptotic testing procedures*.

Regarding (8), this method is expected to work well for large sample sizes. However, resampling methods have proven useful in several different statistical fields if the sample sizes are small (Pauly et al. 2015; Dobler and Pauly 2018; Dobler et al. 2020; Sattler et al. 2022; Ditzhaus et al. 2023; Munko et al. 2024; Baumeister et al. 2024). This particularly holds for permutation tests that even are finitely exact under exchangeability (Hemerik and Goeman 2018; Lehmann and Romano 2022). Ditzhaus et al. (2021) already proposed permutation tests for the QANOVA. In our model, exchangeability means that the distribution functions are equal across the groups, that is, $F_1 = \dots = F_k$. The idea of the permutation approach is to draw the permuted samples $X_{i1}^\pi, \dots, X_{in_i}^\pi, i \in \{1, \dots, k\}$, without replacement from the pooled sample $X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{kn_k}$. Statistics and estimators based on the permuted data $X_{i1}^\pi, \dots, X_{in_i}^\pi, i \in \{1, \dots, k\}$, are denoted here and throughout with a π in the superscript. The permutation QANOVA approach is derived by using permutation-based critical values instead of the standard normal quantiles. Therefore, let $q_{\ell,\beta}^\pi$ and $\widetilde{q}_{\ell,\beta}^\pi$ denote the β-quantiles of the conditional distribution of the permutation test statistics given the data for all $\ell \in \{1, \dots, r\}$. By Equation (A2) in the Appendix A.2, the quantiles are converging in probability to quantiles of the standard normal distribution or its absolute value, respectively. That is why we set $\widetilde{q}_\ell = \widetilde{q}_{\ell,1-\alpha/2r}^\pi$ and $q_\ell = q_{\ell,1-\alpha/r}^\pi$, respectively, for the *Bonferroni-adjusted permutation testing procedure*. The concrete computation of these critical values and necessary assumptions for the asymptotic validity can be found in Section A.2 (see also Ditzhaus et al. 2021). Note that if exchangeability is given, that is, if $F_1 = \dots = F_k$ holds, the permutation test is finitely exact. However, we do not need the exchangeability assumption for proving the asymptotic validity of the permutation test. Hence, the permutation approach also works asymptotically under nonexchangeable data.

3.2 | Multiple Contrast Test Procedures

In this section, we first extend the asymptotic approach of Segbehoe et al. (2022) to inference settings with more than one quantile of interest and to allow for one-sided testing problems. For the asymptotic MCTP, the main ideas are to replace $\mathbf{\Sigma}$ by $\widehat{\mathbf{\Sigma}}$ in the limit distribution in (8) to consider the asymptotic multivariate distribution of the test statistics. Since the local test statistics $T_n(\mathbf{h}_1, \epsilon_1), \dots, T_n(\mathbf{h}_r, \epsilon_r)$ all have the same marginal limit distribution, we may choose the same critical value for all local hypotheses. Then, rejecting the global null hypothesis whenever a local hypothesis is rejected translates into comparing the maximum of the test statistics to the critical value. Hence, in order to determine the critical value for the asymptotic approach, let $(Y_1, \dots, Y_r)' \sim \mathcal{N}(\mathbf{0}, \widehat{\mathbf{D}}\mathbf{H}\widehat{\mathbf{\Sigma}}\mathbf{H}'\widehat{\mathbf{D}})$ given the data with

$$\widehat{\mathbf{D}} := \operatorname{diag}\Big(\left(\mathbf{h}_1' \widehat{\mathbf{\Sigma}} \mathbf{h}_1\right)^{-1/2}, \dots, \left(\mathbf{h}_r' \widehat{\mathbf{\Sigma}} \mathbf{h}_r\right)^{-1/2} \Big).$$

Moreover, denote by $q_{1-\alpha}$ the $(1-\alpha)$-quantile of the conditional distribution of $\max_{\ell \in \{1,\dots,r\}} Y_\ell$ and by $\widetilde{q}_{1-\alpha}$ the $(1-\alpha)$-quantile of the conditional distribution of $\max_{\ell \in \{1,\dots,r\}} |Y_\ell|$ given the data. Due to the consistency of the covariance estimators, $q_{1-\alpha}$ and $\widetilde{q}_{1-\alpha}$ are converging in probability to the $(1-\alpha)$-quantiles of $\max_{\ell \in \{1,\dots,r\}} Z_\ell$ and $\max_{\ell \in \{1,\dots,r\}} |Z_\ell|$, respectively, under Assumption 2.1, see Section A.3 in the Appendix for details. This ensures the asymptotic control of the FWER under Assumption 2.1 by using $\widetilde{q}_\ell = \widetilde{q}_{1-\alpha}$ for the *Asymptotic MCTP for the two-sided problem* and $q_\ell = q_{1-\alpha}$ for the *Asymptotic MCTP for the noninferiority testing problem*, respectively.

For a better small sample performance in the MCTP approach, we also consider a groupwise bootstrap similar to the bootstrap proposed by Segbehoe et al. (2022) to approximate the limiting distribution. This approach is identical to the bootstrap approach in Baumeister et al. (2024). To realize this, we draw a nonparametric bootstrap sample $X_{i1}^*, \dots, X_{in_i}^*$ with replacement from the original i-th sample $X_{i1}, \dots, X_{in_i}$ as in Section 2.4 of Segbehoe et al.

(2022). In detail, $X_{i1}^*, \ldots, X_{in_i}^* \sim \hat{F}_i$ are independent identically distributed given the data $X_{i1}, \ldots, X_{in_i}$. Note that this is simply the adoption of Efron's bootstrap in the context of MCTPs. In the following, the estimators based on the bootstrap samples are denoted with a superscript $*$, respectively. Then, we define the groupwise bootstrap counterpart of the test statistics by

$$T_n^*(\mathbf{h}_\ell) := \sqrt{n}\frac{\mathbf{h}_\ell'(\widehat{\mathbf{q}}^* - \widehat{\mathbf{q}})}{\sqrt{\mathbf{h}_\ell'\widehat{\mathbf{\Sigma}}^*\mathbf{h}_\ell}}, \quad \ell \in \{1, \ldots, r\}.$$

Note that in comparison to Segbehoe et al. (2022), we consider the counterpart of our studentized test statistics (6). Let $q_{1-\alpha}^*$ and $\tilde{q}_{1-\alpha}^*$ denote the $(1-\alpha)$-quantiles of the conditional distribution of the max-test statistics $\max_{\ell\in\{1,\ldots,r\}} T_n^*(\mathbf{h}_\ell)$ and $\max_{\ell\in\{1,\ldots,r\}} |T_n^*(\mathbf{h}_\ell)|$, respectively, given the data. In Section A.4, we prove that choosing $\tilde{q}_\ell = \tilde{q}_{1-\alpha}^*$ and $q_\ell = q_{1-\alpha}^*$ results in asymptotically valid *groupwise bootstrap MCTPs* under Assumption 2.1 whenever the kernel- or interval-based covariance estimator is used. Explicit algorithms for the bootstrap MCTP can be found in Section A.5.

4 | Simulations

Having discussed some asymptotic properties of the different multiple testing approaches, we now evaluate their finite sample performance in various settings. To this end, we did an intensive simulation study using the statistical software R version 4.2.1 (R Core Team 2024). The complete material of the simulation study can be found in the Supporting Information.

4.1 | Simulation for Small Sample Sizes

In this section, we consider $k = 4$ groups and compare the medians, that is, $p_1 = 0.5, m = 1$. Therefore, we use the Dunnett-type, Tukey-type, and Grand-mean-type hypothesis matrix as $\mathbf{H}$, respectively, and $\epsilon_1 = \cdots = \epsilon_r = 0$ for the two-sided and noninferiority hypotheses. Further simulations that focused on the comparison of medians and IQRs simultaneously can be found in the Supporting Information and are summarized at the end of this section. For the data generation, we consider the same setup as in Ditzhaus et al. (2021), that is, we simulate groupwise data from the model

$$X_{is} = \sigma_i(\eta_{is} - m_i) + \mu_i \sim F_i, \quad i \in \{1, \ldots, k\}, s \in \{1, \ldots, n_i\}. \quad (10)$$

Here, we consider different variance settings given by $\boldsymbol{\sigma}_1 = (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 1, 1, 1)$, $\boldsymbol{\sigma}_2 = (1, 1.25, 1.5, 1.75)$, and $\boldsymbol{\sigma}_3 = (1.75, 1.5, 1.25, 1)$ and two different sample size allocations given by $\mathbf{n}_1 = (n_1, n_2, n_3, n_4) = (15, 15, 15, 15)$, $\mathbf{n}_2 = (10, 10, 20, 20)$. This leads to balanced ($\mathbf{n}_1$) and unbalanced ($\mathbf{n}_2$) homo- ($\boldsymbol{\sigma}_1$) and heteroskedastic ($\boldsymbol{\sigma}_2$ and $\boldsymbol{\sigma}_3$) scenarios. In the case of $\mathbf{n}_2$, these can be further divided into heteroskedastic settings with positive ($\boldsymbol{\sigma}_2$) and negative ($\boldsymbol{\sigma}_3$) pairing similar to Pauly et al. (2015). The random variables $\eta_{11}, \ldots, \eta_{1n_1}, \eta_{21}, \ldots, \eta_{kn_k}$ are drawn independently from five different distributions: $\mathcal{N}(0,1)$, $\mathcal{LN}(0,1)$, χ_3^2, t_2, and t_3. Here, $\mathcal{LN}(0,1)$ denotes the log-normal distribution with parameters 0 and 1, χ_3^2 denotes the χ^2-distribution with 3 degrees of freedom, and t_m denotes the t-distribution with m degrees of freedom. The constants m_i in Equation (10) represent the medians of the corresponding distribution. We set $\mu_1 = \cdots = \mu_k = 0$ under the null hypothesis. For power simulations, a shift parameter δ is added to the fourth group as in Ditzhaus et al. (2021), that is, $\mu_4 = \delta \in \{0.5, 1, 1.5\}$. We run $N_{sim} = 5000$ simulation runs for each setting and use $B = 2000$ resampling (permutation, respectively, bootstrap) iterations. The global level of significance was set to $\alpha = 5\%$. Furthermore, the three different covariance matrix estimators as described in Ditzhaus et al. (2021) are considered for all approaches. For the kernel estimator, we used the gaussian kernel and determined the bandwidth by using the following `nrd0` method implemented in the R (R Core Team 2024) function `bw.nrd0`, which is a version of Silverman's rule-of-thumb (Silverman 1998, p. 48): The bandwidth is chosen as $0.9n^{-1/5}\min\{SD, IQR/1.34\}$, where SD denotes the standard deviation, IQR the interquartile range, and n the sample size, if $IQR > 0$. This ensures that the densities are well estimated which in turn ensures that the kernel estimator for the covariance works well. The multiple testing procedures that we compare are the asymptotic MCTP, the bootstrap MCTP, and the Bonferroni-adjusted (abbreviated as B.) asymptotic and permutation-based QANOVA tests of Ditzhaus et al. (2021), as explained in Section 3. This leaves us with 12 different methods which are compared in 120 simulation scenarios. We first discuss their performance in terms of FWER control.

Control of the FWER. In Figures 1–3, the empirical FWERs across all different scenarios are illustrated. The empirical FWERs for the asymptotic MCTP and the asymptotic Bonferroni-adjusted test vary more across the different settings. These tests tend to be too conservative for the bootstrap and interval-based variance estimator, where the Bonferroni adjustment leads to slightly more conservative results than the asymptotic MCTP of Section 3.2. Such a conservative behavior can also be observed in many scenarios for the bootstrap MCTP with interval-based or kernel variance estimator. However, by using the bootstrap MCTP in combination with the bootstrap variance estimator, the type I error of the tests seem to increase and exceeds the desired level of 5% in most of the scenarios for the Dunnett- and Tukey-type contrasts. In contrast, the Bonferroni-adjusted permutation test has a most accurate FWER control across all scenarios. It only exhibits a slight liberality in case of the noninferiority testing for the Grand mean multiple testing family.

Power results. The simulation results for the empirical global and local power can mainly be found in the Supporting Information (Figures 2–10). Here, the empirical global power denotes the rejection rate for a false global hypothesis, while the empirical local power is the rejection rate for a false local hypothesis. It is observable that tests that performed too liberal in terms of type I error control generally also lead to a higher empirical global and local power (as expected). Moreover, the empirical global and local power is always comparable between the asymptotic MCTP and the Bonferroni-adjusted asymptotic and permutation test. In Figure 4, exemplary empirical global power curves are shown for noninferiority Dunnett-type contrast tests in the unbalanced heteroskedastic design with positive pairing ($\mathbf{n}_2$ and $\boldsymbol{\sigma}_2$). It can be seen that the bootstrap MCTP with the bootstrap covariance estimator has generally the highest empirical global power. However, the procedure is also often too liberal as we have seen before. By considering the other methods, we observe that

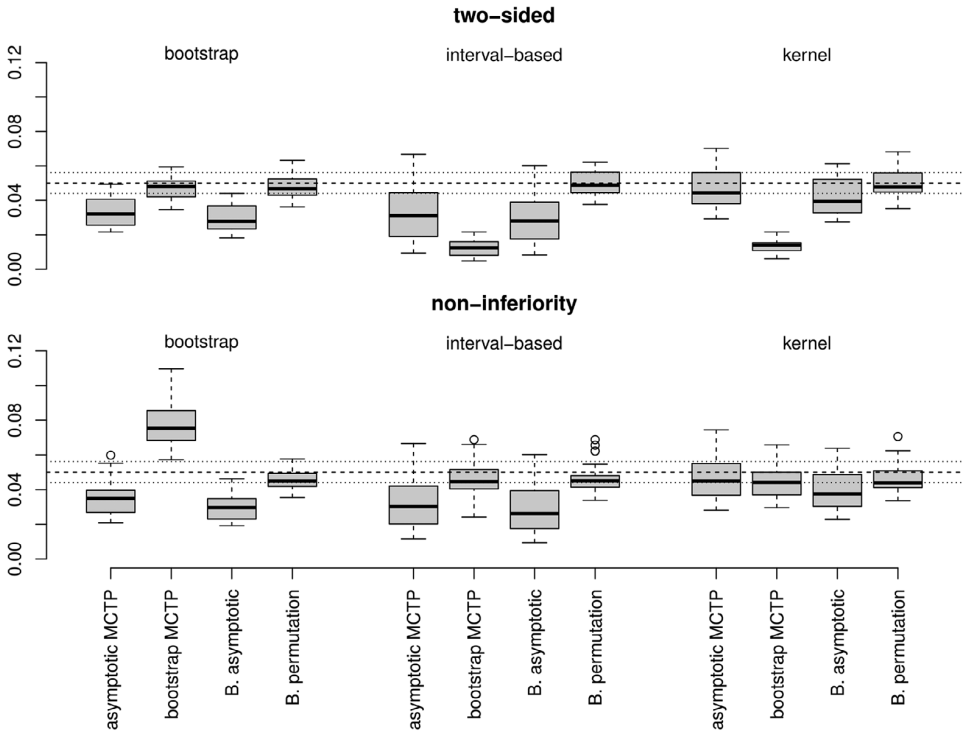

FIGURE 1 | Empirical FWERs for *Dunnett-type* contrasts with different hypotheses (top: two-sided and bottom: noninferiority) and variance estimators (from left to right: bootstrap, interval-based, or kernel). The dashed line represents the desired level of $\alpha = 5\%$ and the dotted lines represent the Binomial interval [0.044,0.0562] for $N_{sim} = 5000$ repetitions.

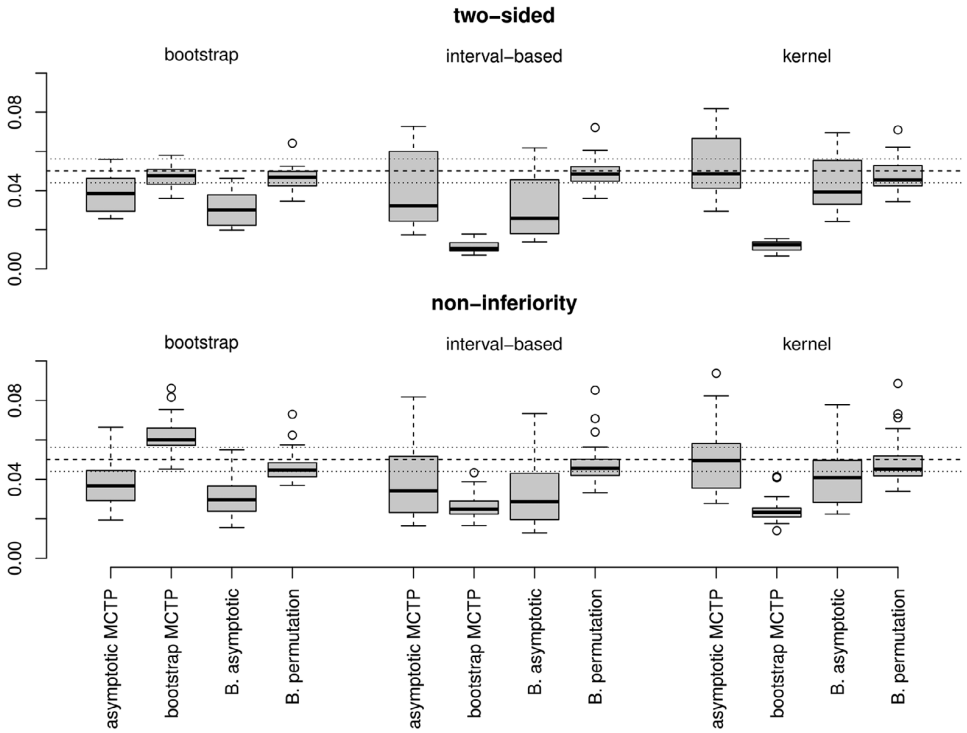

FIGURE 2 | Empirical FWERs for *Tukey-type* contrasts with different hypotheses (top: two-sided and bottom: noninferiority) and variance estimators (from left to right: bootstrap, interval-based, or kernel). The dashed line represents the desired level of significance of $\alpha = 5\%$ and the dotted lines represent the Binomial interval [0.044,0.0562] for $N_{sim} = 5000$ repetitions.

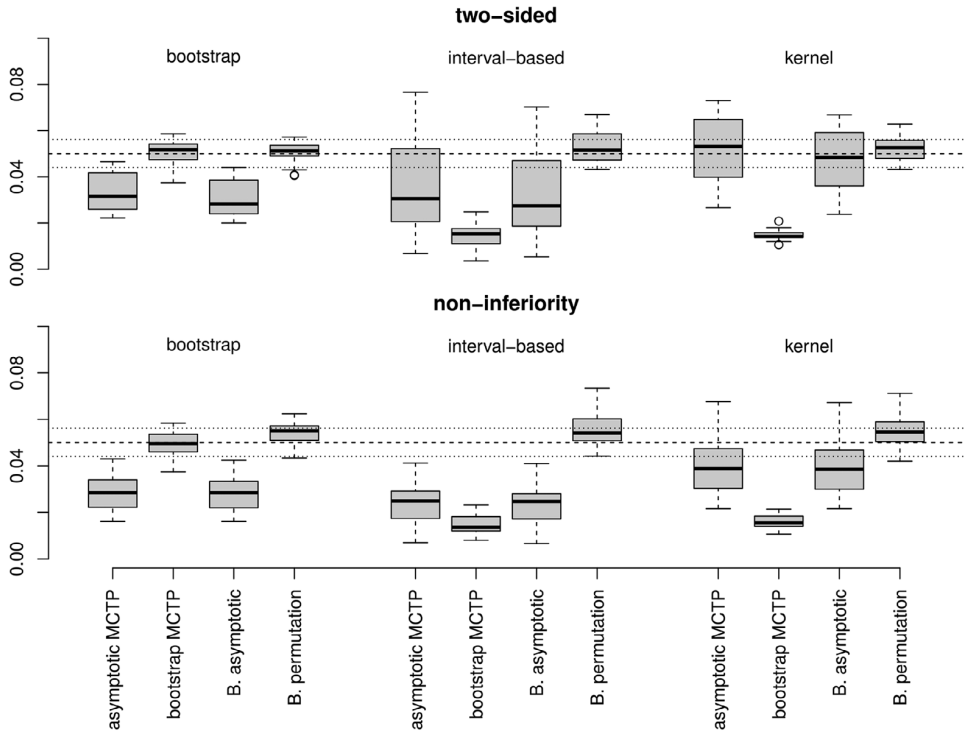

FIGURE 3 | Empirical FWERs for *Grand-mean-type* contrasts with different hypotheses (top: two-sided and bottom: noninferiority) and variance estimators (from left to right: bootstrap, interval-based, or kernel). The dashed line represents the desired level of significance of $\alpha = 5\%$ and the dotted lines represent the Binomial interval [0.044,0.0562] for $N_{sim} = 5000$ repetitions.

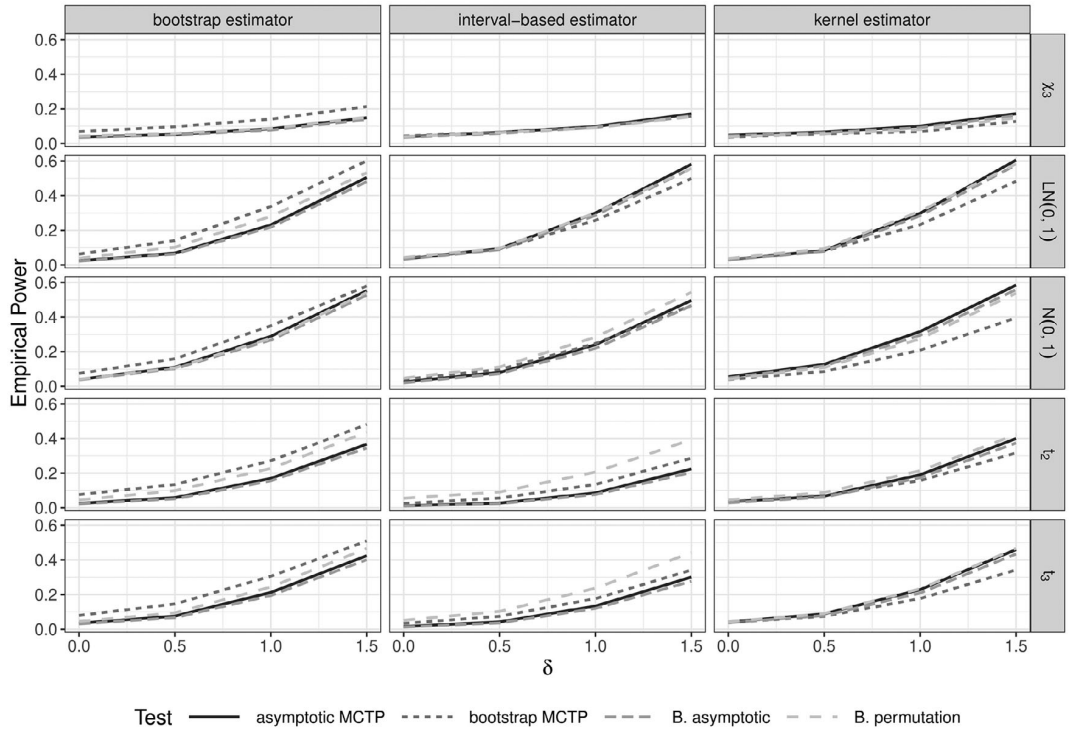

FIGURE 4 | Empirical power for noninferiority Dunnett-type contrast tests in the unbalanced ($\mathbf{n}_2$) heteroskedastic ($\boldsymbol{\sigma}_2$) design with positive pairing under different distributions and variance estimators (from left to right: bootstrap, interval-based, or kernel).

the Bonferroni-adjusted permutation test is usually one of the methods with the highest global and local power or at least with a comparable power to the method with the highest global and local power, respectively. This is also the case for the other variance estimators. Especially for the interval-based estimator and the standard normal and t-distributions, the Bonferroni-adjusted permutation test clearly outperforms the other methods in terms of empirical global power. Furthermore, it is observable that the Bonferroni-adjusted asymptotic test is slightly less powerful than the asymptotic MCTP in all scenarios. All in all, regarding the local and global power one cannot make a clear recommendation, but the power of the Bonferroni-adjusted methods is in general not worse than the MCTP methods.

Other effect parameters. The results of the additional simulation study in the Supporting Information, where medians and IQRs are inferred simultaneously, are similar: The Bonferroni-adjusted permutation test performs quite accurate in terms of FWER control while the asymptotic approaches and the bootstrap MCTP tend to be conservative in most scenarios. Regarding the empirical power, the Bonferroni-adjusted permutation test is comparable and in some scenarios even more powerful than the other approaches.

4.2 | Simulation Motivated by the Data Example

We also conducted an additional simulation study with $r = 16$ tests and larger sample sizes of 58–549 individuals per group as in the data example in Section 5. We consider a modification of the simulation study in Section 4.1 to analyze the performance of the methods in a framework that is closer to the considered data example in Section 5. Therefore, we considered $k = 17$ groups and used the Dunnett-type contrast matrix as $\mathbf{H}$ with group 17 as base. Furthermore, the constants are set to $\epsilon_1 = \cdots = \epsilon_{16} = 0$. For the data generation, we used the model as in Section 4.1. The sample sizes are set to $n = (59, 175, 98, 78, 280, 176, 351, 128, 368, 403, 240, 376, 278, 549, 428, 379, 250)$, which is similarly heterogeneous as the number of individuals in the 17 groups of the hatch data in Section 5 (see Figure 7). The variance setting is motivated by the data example as the parameters $\sigma_1, \dots, \sigma_{17}$ are chosen such that the variances of X_{is} match the empirical variances in group i for the hatch data of Section 5. This yields a heteroskedastic variance setting. The random variables $\eta_{11}, \dots, \eta_{1n_1}, \eta_{21}, \dots, \eta_{kn_k}$ are drawn independently from four different distributions: $\mathcal{N}(0,1)$, $\mathcal{LN}(0,1)$, χ^2_3, and t_3. The reason why we exclude the t_2-distribution is that the variances of X_{is} would not exist in this case. Hence, it would not be possible to choose the parameters $\sigma_1, \dots, \sigma_{17}$ such that the variances of X_{is} equal the empirical variances. The constants m_i represent the medians of the corresponding distribution. We set $\mu_1 = \cdots = \mu_{17} = 0$ under the null hypothesis. For power simulations, we set μ_i to the empirical median of group i for the hatch data of Section 5 for all $i \in \{1, \dots, 16\}$ and μ_{17} to the empirical median of group 17 for the hatch data minus 7 (which is the constant ϵ_ℓ in the data analysis). All other parameters are set as in Section 4.1.

Control of the FWER. The empirical FWERs under the null hypothesis across all scenarios are illustrated in Figure 5. Here, we see that the results are not as surprising as for smaller sample sizes and less groups. Particularly for the bootstrap variance estimator, the empirical FWERs of the MCTPs are quite close to the desired level $\alpha = 0.05$, while the interval and kernel estimators still show an observable deviance from $\alpha = 0.05$ The Bonferroni-adjusted tests tend to be too conservative. This might be explained by the large number of tests, that is 16. The asymptotic MCTP with interval-based estimator performs slightly too liberal in the considered simulation settings.

Power results. The empirical global power, which is the rejection rate of the global null hypothesis under the alternative, has been exactly 1 for all scenarios. This means that the global null hypothesis could be rejected in all simulation runs for all settings under the alternative.

4.3 | Discussion of the Results

The simulation results are quite surprising in several ways. There are two well-known and often discussed problems with the Bonferroni adjustment in general: a loss of power (e.g., Holm 1979; Olejnik et al. 1997) and a rather conservative behavior (e.g., Westfall and Young 1989; Gordon et al. 2007; Chen et al. 2017). From the method's definition, it is clear that the conservative behavior occurs if a large number of hypotheses is simultaneously tested or the hypotheses are highly correlated. The situation that is described in other articles is vice versa for MCTPs. Hasler and Hothorn (2008) and many others (e.g., Bretz et al. 2001; Hasler and Hothorn 2008; Konietschke et al. 2013; Hasler 2014; Umlauft et al. 2019; Noguchi et al. 2020; Rubarth et al. 2022) showed in simulation studies that MCTPs hold their level of significance quite satisfactorily. Furthermore, Konietschke et al. (2013) showed that the power of the global test decision of some mean-based MCTPs is comparable to the power of an ANOVA-F-test. In fact, this is exactly what we could observe in the simulation study of Section 4.2 with larger sample sizes and many hypotheses. From these observations, one would assume that the MCTPs are the preferred method compared to Bonferroni-adjusted procedures. However, the simulation results of Section 4.1 with smaller sample sizes are in favor of the Bonferroni-adjusted permutation approach. We point out that the good behavior of the Bonferroni adjustment can only be observed for small sample sizes in combination with the permutation approach, the standard asymptotic version was often observed to be too conservative. Moreover, it is important to note that the Bonferroni correction cannot really be improved by an MCTP for large negative correlations between the test statistics. However, this situation mainly occurs for the noninferiority tests with Tukey- and Grand-mean-type matrix in our simulations (cf. Section A.6 for details). For highly positive correlated tests, it is well-known that the Bonferroni correction performs too conservative, which is simply a consequence of the Bonferroni inequality. In our simulation settings, we can observe the highest positive correlation in Dunnett-type tests with noninferiority hypothesis (median correlation 0.524 with bootstrap covariance estimator, cf. Section A.6), but cannot observe that the Bonferroni correction Permutation approach behaves very conservative. It should also be emphasized that most MCTPs and corresponding simulations or analyses use means and not quantiles as an estimand. Furthermore, VanderWeele and Mathur (2019) stated that there are still many testing problems where the behavior is not or only a little bit conservative and the tests are still rejecting even if they are

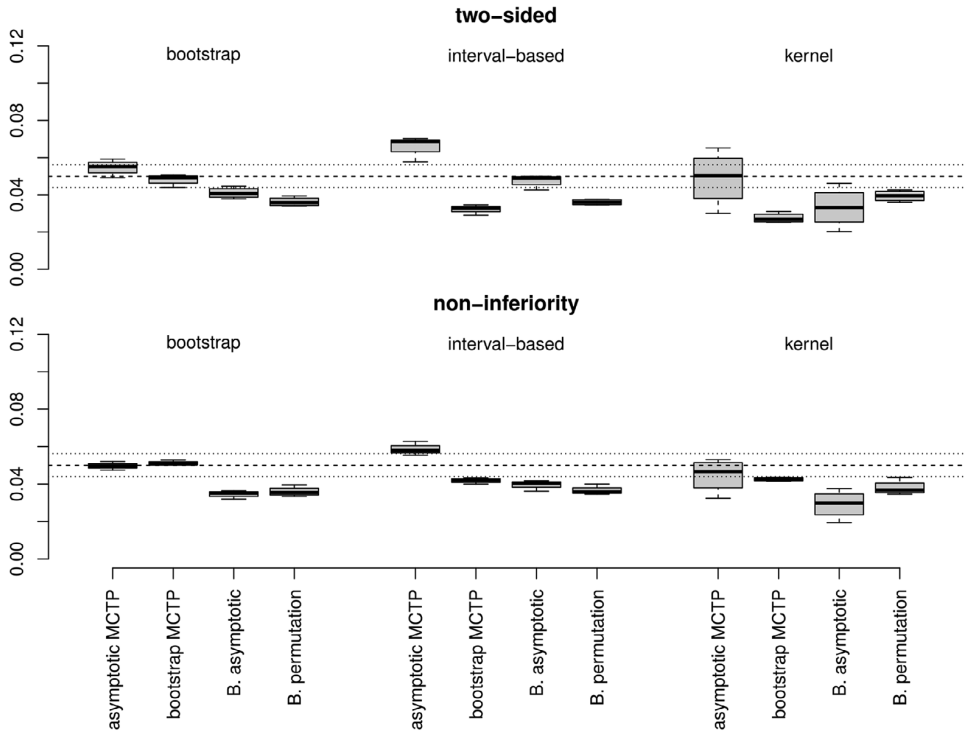

FIGURE 5 | Empirical FWERs for Dunnett-type contrasts with different hypotheses (top: two-sided and bottom: noninferiority) and variance estimators (from left to right: bootstrap, interval-based, or kernel).

less powerful than other tests that control the FWER. Moreover, the Bonferroni-adjusted QANOVA permutation approach and the MCTPs are not directly comparable as they use different techniques to derive critical values. In particular, the estimation of the covariance matrix is more crucial for the MCTPs than for the QANOVA as the latter may have a balancing effect through the studentized permutation approach. As the estimation of the underlying covariance structures is much more complex in the case of quantiles compared to classical mean-based approaches, this could be one reason for our results. In context of the simulation study of Segbehoe et al. (2022), our simulation results lead to the conclusion, that MCTPs regarding quantiles do not perform well for smaller sample sizes and more than three groups, especially less well than the Bonferroni-adjusted permutation test.

Recommendation. To conclude, we recommend to use the Bonferroni-adjusted permutation test for small sample sizes and few hypotheses due to a quite accurate FWER control and comparable power to other methods. The simulations indicate that the choice of the variance estimator has no big impact on the permutation tests decision. All methods are expected to perform similarly well for larger sample sizes regarding the FWER control. However, if the number of tests increases, the Bonferroni adjustment may lead to conservative test results. This observation is not surprising and refers to the well-known disadvantages of the Bonferroni correction. In order to ensure more powerful test decisions, we recommend to use the bootstrap MCTPs with bootstrap variance estimator in the case of many hypotheses and large sample sizes.

5 | Data Example: Early and Late Buzzards

Birds living in temperate climates have to cope with changing seasons during the year; they have to adapt to different weather conditions, temperatures, and length of daytimes (Begon and Townsend 2021). Parental care is probably one of the most important activities of birds regulated by the seasons due to the strong connection to reproduction and fitness (Caro 2005). For this energy-demanding task, most birds must rely on sufficient resources to feed their young and thus are dependent on a small time frame during the year, when enough of these resources are available (Verhulst and Nilsson 2007). To do so, most birds rely on hints from temperature or length of daylight (Verhulst and Nilsson 2007) to time hatching in the best possible way. Since human-induced climate change alters weather conditions as well as temperature developments through the year way faster than during earlier decades and centuries (Sippel et al. 2020), birds relying on these influences to time their reproduction were shown to change their reaction accordingly (Halupka and Halupka 2017). At first glance, this might seem positive as climate change leads in general to warmer temperatures and hence the reproductive period during the year should potentially increase (McDermott and DeGroote 2016). However, not all organisms react in the same way and at the same pace to these changes, leading to potential mismatches in the food web (Drever and Clark 2007).

Common buzzards (*Buteo buteo*, Figure 6) are medium-sized birds of prey and feed mostly on small mammals and birds (Walls and Kenward 2020). As being predators, they are dependent on the performance of many other organisms, not only their prey, but also their prey's food resources (Mittelbach and McGill

FIGURE 6 | Adult common buzzard during flight (left) and buzzard nestlings in the nest (right). © O. Krüger, N. Chakarov.

2019). It is known that buzzards have higher breeding success under certain weather conditions (Kostrzewa and Kostrzewa 1990; Krüger 2002, 2004). Their main prey, field voles (*Microtus arvalis*), often shows fluctuating population densities between years (Frank 1957) caused by different factors like predation pressure or snow level during winter (Boyce and Boyce 1988), also influencing breeding success in buzzards (Lehikoinen et al. 2009). In their study, Lehikoinen et al. (2009) showed as well that common buzzards in Finland started breeding earlier and shifted their range more toward the north because of the warmer climate.

Data were collected from 2006 to 2022 by the Department of Animal Behaviour in a study area in northwest of Germany (see Chakarov et al. 2013 for a description of the study area and sampling procedure). We only consider the age of the first-hatched nestling of each brood for this analysis to avoid dependencies between siblings. With the relationship between the age of the chicks and their wing length observed by Bijlmsa (1999), we are able to calculate the hatch dates of the chicks. We use R 4.4.2 (R Core Team 2024) and the implementation in R by Ottensmann (2022) for the calculation. Here, we use as a scaling the *day of the year*, where 32 means 1st February and 120 means 30th April in nonleap years. From this, the hatch dates are calculated by the day of observation minus the age in days. Therefore, the considered hatch dates are the result of a polynomial model and accordingly metric. This data are shown in Figure 7 as kernel density estimators with gaussian kernels and a bandwidth determined by the `nrd0` method as explained in Section 4.1. The complete material of this analysis as well as the data can be found in the Supporting Information.

In Figure 7, there is a high variability between years regarding the hatch dates of common buzzard nestlings. Biologists who study these animals often have the impression of particular *early* and *late* years, especially if the behavior of the buzzards differs from the years before. As this is also observable in the kernel density estimators in Figure 7, it is a motivation to search for possible reasons. In the years from 2019 onward, this pattern seemed to be changing as the years 2019, 2020, and 2021 tended to be earlier in contrast to the year 2022. This is our motivation to take 2022 as a reference year for *late* years. There are potentially several reasons for this phenomenon which are difficult to measure, but the division into two groups (early and late years) is a simplification that makes it possible to get a more accurate view. For that, we do a multiple testing procedure that identifies similar years. In context of a directed scenario, the multiple testing procedure that identifies similar years can be understood as a multiple noninferiority testing problem as described in (2). Regarding the structure of the data, heavy-tailed distributions appear quite frequently (Figure 7). The simulation studies of Ditzhaus et al. (2021) show that median-based tests have a higher power than mean-based tests in the context of heavy-tailed data. So if one wants to identify similar years to investigate possible reasons, it is in this case most suitable to use a median-based approach. The sample sizes of the $k = 17$ groups are shown in Figure 7. From this, it can be seen that the groups are highly unbalanced. This is no problem because from the simulation study we observed that all methods can deal with high variation in sample sizes between samples. As this behavior can also be observed in the simulations of Ditzhaus et al. (2021) and Baumeister et al. (2024), this seems to be a useful property of testing regarding quantiles. Consider Figure 1 in the Supporting Inforamtion for an analysis of that property in our simulation study. Since the simulation setup of Section 4 does not perfectly fit to the data example, we have conducted a further simulation motivated by the data example. The detailed description of the scenarios and the results can be found in Section 5 in the Supporting Information.

We use the median m_ℓ of the year $\ell \in \{06, 07, 08, \dots, 19, 20, 21\}$ ($m = 1$, $p_1 = 0.5$, $k = 17$ groups) of the hatch dates and $\epsilon_\ell = 7$ for all ℓ regarding to the intuitive observations of the ecologists. Hatches 1 week (7 days) later do not lead to the conclusion that a year is late. This situation leads to a Dunnett's test or many-to-one procedure and has the concrete form:

$$\mathcal{H}^I_{0,\ell} : m_{22} - m_\ell \geq 7 \quad \text{vs.} \quad \mathcal{H}^I_{1,\ell} : m_{22} - m_\ell < 7,$$
$$\ell \in \{06, 07, 08, \dots, 19, 20, 21\}. \tag{11}$$

To realize these hypotheses of interest, we use the Dunnett-type hypothesis matrix

$$\mathbf{H} = [\mathrm{Diag}(\mathbf{1}_{16}), -\mathbf{1}_{16}].$$

Note that this framework assumes independent groups, which means in terms of content that the years are assumed to be independent. This is a plausible assumption because of the high fluctuation of the buzzards in the data sample. As the data collection is based on the defined area and not on the individual buzzard, the sample size differs through the years (see Figure 7), every year birds migrate to the area, some leave it and other

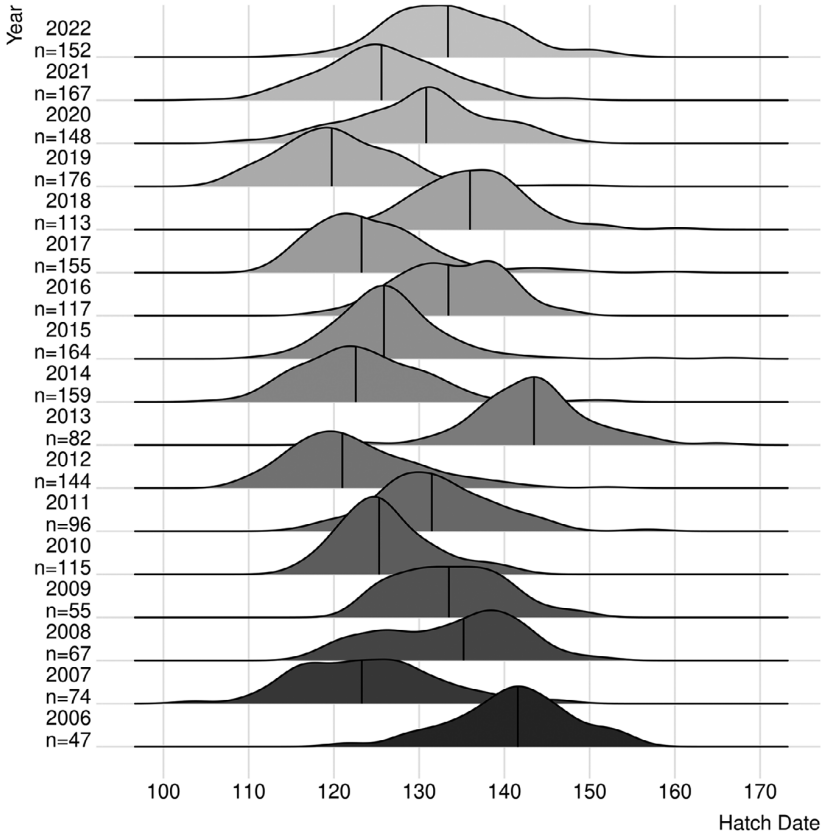

FIGURE 7 | Kernel density estimators of hatching dates (in days) for the years 2006–2022 with the sample sizes for every year. The black vertical line marks the median.

change their nest within the area. Therefore, the data are not collected as a paired sample.

In the context of the not entirely clear simulation results, we consider the four presented testing methods (cf. Section 3) with the bootstrap covariance estimator as the simulation results for larger sample sizes indicates the best performance for that covariance estimator in the bootstrap MCTP while this does not seems to be relevant in other situations and for other methods. We used 1999 iterations for both resampling methods. The results for the selected test decisions are given as p-values and as confidence intervals in Table 1 (cf. Section 3.) Here, all testing methods indicate that the null hypotheses $\mathcal{H}^I_{0,06}$, $\mathcal{H}^I_{0,08}$, $\mathcal{H}^I_{0,09}$, $\mathcal{H}^I_{0,11}$, $\mathcal{H}^I_{0,13}$, $\mathcal{H}^I_{0,16}$, $\mathcal{H}^I_{0,18}$, and $\mathcal{H}^I_{0,20}$ are rejected to a significance level of $\alpha = 0.05$. In line with the noninferiority multiple testing problem, the global hypothesis $\mathcal{H}^I_0 = \bigcap_{\ell=1}^{r} \mathcal{H}_{0,\ell}$ is also rejected. This means that the hatch days in the years 2006, 2008, 2009, 2011, 2013, 2016, 2018, and 2020 are identified as at least as late as the year 2022 because the difference of the median hatch days between these years is not significantly bigger as 1 week. These *late* years can be used for further investigation of reasons for different hatching dates. Here, the most interesting result is that the year 2020 is still categorized as an late year, although the density plots in Figure 7 suggest a seemingly high similarity of the years 2019, 2020, and 2021. However, the tests indicate that 2020 is at least as late as 2022, while 2019 and 2021 could not be identified as at least late as 2022, which is an important information to look for possible reasons. This is a good motivation for using the median as estimand because it is not sensitive to the heavy-tailed data. In line with the simulation results for bigger sample sizes, the testing procedures perform similarly. In the sense of the simulation study, special attention should be paid to the MCTP with bootstrap.

6 | Conclusion and Outlook

We have compared different approaches to solve one- and two-sided multiple testing problems regarding one or more quantiles simultaneously. To this end, we have presented and extended two Bonferroni-adjusted methods, an asymptotic and a permutation approach, and an asymptotic and a bootstrap MCTP in a comparable multiple testing framework for two-sided and noninferiority hypotheses. As a motivation for this kind of testing problems, we gave a noninferiority example from ecology, which deals with hatch dates in context of climate change. To investigate the behavior of the methods, we have conducted an intensive simulation study. Here, our main motivation was to compare Bonferroni adjustment and MCTPs in context of testing regarding quantiles. In line with VanderWeele and Mathur (2019), we have found out that the Bonferroni adjustment can be conservative, but when combined with a permutation approach in the situation of small sample sizes it performs better than its reputation. The often-read claim that the Bonferroni method is in general too conservative (Gordon et al. 2007), cannot be confirmed when inferring quantiles. We also wanted to ask the question whether the MCTPs are less conservative and have more power

TABLE 1 | Test results of the four testing procedures (from left to the right: *asymptotic MCTP, bootstrap MCTP, Bonferroni asymptotic*, and *Bonferroni permutation*, cf. Section 3) with bootstrap covariance estimator regarding the testing problem given in Equation 11. The test results are given as p-value and as the right value of the one-sided confidence interval $(-\infty, \cdot]$ for $m_{22} - m_\ell$ (cf. Equation 7). Interpretation: Rejecting local null hypotheses $\mathcal{H}_{0\ell}^I$ means that we can rule out that the median hatch date of year ℓ is at least 7 days later than for year 2022. The p-values for the bootstrap MCTP and the Bonferroni permutation consider the number of resampling iterations and are calculated by eq. (17.7) in Lehmann and Romano (2022). This testing problem uses Dunnett-type contrasts, the median hatch date of the year 2022 is compared with the median hatch dates from the years 2006 to 2021. The first column contains the differences of the empirical medians between the year 2022 and the respective other years.

	Asymp. MCTP		**Boot. MCTP**		**B. asymp.**		**B. perm.**	
Test	***p*-Value**	**SCI**	***p*-Value**	**SCI**	***p*-Value**	**SCI**	***p*-Value**	**SCI**
$\hat{m}_{22} - \hat{m}_{06} = -8.19$	<0.0001	−5.01	<0.0005	−4.93	<0.0001	−4.83	<0.0006	−4.94
$\hat{m}_{22} - \hat{m}_{07} = 10.11$	1.0000	14.15	1.0000	14.25	1.0000	14.37	1.0000	14.04
$\hat{m}_{22} - \hat{m}_{08} = -1.83$	<0.0001	2.71	<0.0006	2.83	<0.0001	2.97	<0.0006	2.87
$\hat{m}_{22} - \hat{m}_{09} = -0.11$	<0.0001	3.77	<0.0006	3.87	<0.0001	3.99	<0.0006	4.35
$\hat{m}_{22} - \hat{m}_{10} = 8.08$	0.9998	10.11	1.0000	10.17	1.0000	10.23	1.0000	10.05
$\hat{m}_{22} - \hat{m}_{11} = 1.90$	<0.0001	4.82	<0.0006	4.90	<0.0001	4.98	<0.0006	5.08
$\hat{m}_{22} - \hat{m}_{12} = 12.39$	1.0000	14.94	1.0000	15.01	1.0000	15.09	1.0000	15.04
$\hat{m}_{22} - \hat{m}_{13} = -10.09$	<0.0001	−7.51	<0.0006	−7.44	<0.0001	7.36	<0.0006	−7.31
$\hat{m}_{22} - \hat{m}_{14} = 10.82$	1.0000	13.44	1.0000	13.51	1.0000	13.59	1.0000	13.91
$\hat{m}_{22} - \hat{m}_{15} = 7.51$	0.9930	9.97	0.9990	10.03	1.0000	10.10	1.0000	10.41
$\hat{m}_{22} - \hat{m}_{16} = -0.05$	<0.0001	2.95	<0.0006	3.02	<0.0001	3.12	<0.0006	3.23
$\hat{m}_{22} - \hat{m}_{17} = 10.14$	1.0000	13.01	1.0000	13.08	1.0000	13.17	1.0000	13.10
$\hat{m}_{22} - \hat{m}_{18} = -2.58$	<0.0001	0.16	<0.0006	0.23	<0.0001	0.31	<0.0006	0.21
$\hat{m}_{22} - \hat{m}_{19} = 13.64$	1.0000	16.01	1.0000	16.07	1.0000	16.15	1.0000	16.12
$\hat{m}_{22} - \hat{m}_{20} = 2.56$	<0.0001	4.91	<0.0006	4.97	<0.0001	5.05	<0.0006	5.09
$\hat{m}_{22} - \hat{m}_{21} = 7.80$	0.9967	10.43	0.9990	10.50	1.0000	10.59	1.0000	10.66

than Bonferroni-adjusted approaches. Our clear answer in this quantile-based setting with small samples is: no. This is because of the behavior of the Bonferroni-adjusted permutation approach, which is very stable. Independently from the considered distributions, the covariance structure or the sample sizes, its empirical FWER control was quite accurate and there was almost no power loss compared to the MCTPs. In contrast to the asymptotic and resampling-based MCTP approaches, the permutation-based method does not seem to need bigger sample sizes to work well. For both small and large samples, the resampling-based methods show a clear improvement in the test performance.

We also want to point out that hypotheses formulated in terms of quantiles can be useful in lots of situations. This is particularly important in the context of data that refers to animal and human behavior, as this situations are known to be rather skewed and can be rarely modeled as homoskedastic and normally distributed (e.g., Gardiner et al. 2014). As multiple testing problems occur very often in this field of science (Farcomeni 2008), our analyses can be helpful in the selection of the appropriate method.

For future research, it remains to create an implementation in R for the presented methods as well as for other quantile-based methods for factorial designs, for example, Ditzhaus et al. (2021) and Baumeister et al. (2024). In addition, it can be investigated how multiple testing regarding multivariate quantiles can be realized by extending the QMANOVA of Baumeister et al. (2024). Furthermore, it would be interesting to have more systematic comparisons between MCTPs and other multiple testing procedures like the Bonferroni adjustment for other estimands of interest. Especially for mean-based methods, it would be interesting to investigate if a similar simulation-based comparison comes to the same conclusion as our simulation does. Then a general statement could be made about whether this relationship between the behavior of the Bonferroni correction and sample size occurs systematically. As Besag et al. (1995, Sec. 6.3) introduced quantile-based simultaneous credible regions, there are also Bayesian approaches that could be compared with the methods presented in this paper in further comparisons. From this, we hope to gain a better overview of the behavior of MCTPs in relation to FWER control, power, and further concepts.

Acknowledgments

M.M. and M.D. gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft - 314838170, GRK 2297 MathCoRe. The work of M.B. and M.P. has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance Centers within the UA Ruhr. The work of K.-P.G. has been partly supported by the Friedrich–Ebert–Stiftung (FES) and the work of K.-P.G and N.C. has been partly supported by the SFB TRR 212 (NC^3) (Project Number 396780709) of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). We thank Meinolf Ottensmann for his help with the calculation of the hatch dates. Furthermore, we thank Paavo Sattler for many helpful advices on multiple testing.

Open access funding enabled and organized by Projekt DEAL.

Author Contributions (CRediT)

Marléne Baumeister: data curation, formal analysis, methodology, project administration, software, validation, visualization, writing–original draft. **Merle Munko:** formal analysis, methodology, software, validation, visualization, writing–original draft. **Kai-Philipp Gladow:** data curation, investigation, resources, writing–original draft. **Marc Ditzhaus:** conceptualization, funding acquisition, methodology, supervision, writing–review and editing. **Nayden Chakarov:** investigation, resources, writing–review and editing. **Markus Pauly:** conceptualization, funding acquisition, methodology, supervision, writing–review and editing.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study, for example, the data example, simulation scripts, and results are openly available in TUDOdata at http://doi.org/10.17877/TUDODATA-2025-M6TDKFDE.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

References

Baumeister, M., M. Ditzhaus, and M. Pauly. 2024. "Quantile-Based MANOVA: A New Tool for Inferring Multivariate Data in Factorial Designs." *Journal of Multivariate Analysis* 199: 105246.

Begon, M., and C. R. Townsend. 2021. *Ecology: From Individuals to Ecosystems*. Wiley.

Besag, J., P. Green, D. Higdon, and K. Mengersen. 1995. "Bayesian Computation and Stochastic Systems." *Statistical Science* 10, no. 1: 3–41.

Bijlmsa, R. 1999. "Sex Determination of Nestling Common Buzzards *Buteo buteo*." *Limosa* 72: 1–10.

Bonett, D. G., and R. M. Price. 2002. "Statistical Inference for a Linear Function of Medians: Confidence Intervals, Hypothesis Testing, and Sample Size Requirements." *Psychological Methods* 7, no. 3: 370–383.

Boyce, C. C. K., and J. L. Boyce. 1988. "Population Biology of *Microtus arvalis*. I. Lifetime Reproductive Success of Solitary and Grouped Breeding Females." *Journal of Animal Ecology* 57, no. 3: 711–722.

Bretz, F., A. Genz, and L. A. Hothorn. 2001. "On the Numerical Availability of Multiple Comparison Procedures." *Biometrical Journal* 43, no. 5: 645–656.

Bretz, F., T. Hothorn, and P. H. Westfall. 2011. *Multiple Comparisons Using R*. CRC Press.

Caro, T. 2005. *Antipredator Defenses in Birds and Mammals*. University of Chicago Press.

Chakarov, N., R. M. Jonker, M. Boerner, J. I. Hoffman, and O. Krüger. 2013. "Variation at Phenological Candidate Genes Correlates With Timing of Dispersal and Plumage Morph in a Sedentary Bird of Prey." *Molecular Ecology* 22, no. 21: 5430–5440.

Chen, S.-Y., Z. Feng, and X. Yi. 2017. "A General Introduction to Adjustment for Multiple Comparisons." *Journal of Thoracic Disease* 9, no. 6: 1725–1729.

Chung, E., and J. P. Romano. 2013. "Exact and Asymptotically Robust Permutation Tests." *Annals of Statistics* 41, no. 2: 484–507.

Chung, E., and J. P. Romano. 2016. "Multivariate and Multiple Permutation Tests." *Journal of Econometrics* 193, no. 1: 76–91.

Ditzhaus, M., R. Fried, and M. Pauly. 2021. "QANOVA: Quantile-Based Permutation Methods for General Factorial Designs." *TEST* 30, no. 4: 960–979.

Ditzhaus, M., M. Yu, and J. Xu. 2023. "Studentized Permutation Method for Comparing Two Restricted Mean Survival Times With Small Sample From Randomized Trials." *Statistics in Medicine* 42, no. 13: 2226–2240.

Djira, G., M. Hasler, D. Gerhard, L. Segbehoe, and F. Schaarschmidt. 2020. *mratios: Ratios of Coefficients in the General Linear Model*. https://doi.org/10.32614/CRAN.package.quantreg.

Djira, G. D., and L. A. Hothorn. 2009. "Detecting Relative Changes in Multiple Comparisons With an Overall Mean." *Journal of Quality Technology* 41, no. 1: 60–65.

Dobler, D., S. Friedrich, and M. Pauly. 2020. "Nonparametric MANOVA in Meaningful Effects." *Annals of the Institute of Statistical Mathematics* 72, no. 4: 997–1022.

Dobler, D., and M. Pauly. 2018. "Bootstrap- and Permutation-Based Inference for the Mann–Whitney Effect for Right-Censored and Tied Data." *TEST* 27, no. 3: 639–658.

Drever, M. C., and R. G. Clark. 2007. "Spring Temperature, Clutch Initiation Date and Duck Nest Success: A Test of the Mismatch Hypothesis." *Journal of Animal Ecology* 76, no. 1: 139–148.

Dunn, O. J. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56, no. 293: 52–64.

Dunnett, C. W. 1955. "A Multiple Comparison Procedure for Comparing Several Treatments With a Control." *Journal of the American Statistical Association* 50, no. 272: 1096–1121.

Farcomeni, A. 2008. "A Review of Modern Multiple Hypothesis Testing, With Particular Attention to the False Discovery Proportion." *Statistical Methods in Medical Research* 17, no. 4: 347–388.

Frank, F. 1957. "The Causality of Microtine Cycles in Germany (Second Preliminary Research Report)." *Journal of Wildlife Management* 21, no. 2: 113–121.

Gabriel, K. R. 1969. "Simultaneous Test Procedures—Some Theory of Multiple Comparisons." *Annals of Mathematical Statistics* 40, no. 1: 224–250.

Gardiner, J. C., Z. Luo, X. Tang, and R. V. Ramamoorthi. 2014. "Fitting Heavy-Tailed Distributions to Health Care Data by Parametric and Bayesian Methods." *Journal of Statistical Theory and Practice* 8, no. 4: 619–652.

Goeman, J. J., and A. Solari. 2022. "Comparing Three Groups." *American Statistician* 76, no. 2: 168–176.

Gordon, A., G. Glazko, X. Qiu, and A. Yakovlev. 2007. "Control of the Mean Number of False Discoveries, Bonferroni and Stability of Multiple Testing." *Annals of Applied Statistics* 1, no. 1: 179–190.

Halupka, L., and K. Halupka. 2017. "The Effect of Climate Change on the Duration of Avian Breeding Seasons: A Meta-Analysis." *Proceedings of the Royal Society B: Biological Sciences* 284, no. 1867: 20171710.

Hasler, M. 2014. "Multiple Contrast Tests for Multiple Endpoints in the Presence of Heteroscedasticity." *International Journal of Biostatistics* 10, no. 1: 17–28.

Hasler, M., and L. A. Hothorn. 2008. "Multiple Contrast Tests in the Presence of Heteroscedasticity." *Biometrical Journal* 50, no. 5: 793–800.

Hauck, W., and S. Anderson. 1984. "A New Statistical Procedure for Testing Equivalence in 2-Group Comparative Bioavailability Trials." *Journal of Pharmacokinetics and Biopharmaceutics* 12, no. 1: 83–91.

Hemerik, J., and J. Goeman. 2018. "Exact Testing With Random Permutations." *TEST* 27, no. 4: 811–825.

Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6, no. 2: 65–70.

Hothorn, T., F. Bretz, and P. Westfall. 2008. "Simultaneous Inference in General Parametric Models." *Biometrical Journal* 50, no. 3: 346–363.

Koenker, R. 2024. *Quantreg: Quantile Regression*.

Koenker, R., and G. Bassett. 1978. "Regression Quantiles." *Econometrica* 46, no. 1: 33.

Konietschke, F., S. Bösiger, E. Brunner, and L. A. Hothorn. 2013. "Are Multiple Contrast Tests Superior to the ANOVA?" *International Journal of Biostatistics* 9, no. 1: 63–73.

Kostrzewa, A., and R. Kostrzewa. 1990. "The Relationship of Spring and Summer Weather with Density and Breeding Performance of the Buzzard *Buteo buteo*, Goshawk *Accipiter gentilis* and Kestrel *Falco tinnunculus*." *IBIS* 132, no. 4: 550–559.

Krüger, O. 2002. "Dissecting Common Buzzard Lifespan and Lifetime Reproductive Success: The Relative Importance of Food, Competition, Weather, Habitat and Individual Attributes." *Oecologia* 133, no. 4: 474–482.

Krüger, O. 2004. "The Importance of Competition, Food, Habitat, Weather and Phenotype for the Reproduction of Buzzard *Buteobuteo*." *Bird Study* 51, no. 2: 125–132.

Lehikoinen, A., P. Byholm, E. Ranta, et al. 2009. "Reproduction of the Common Buzzard at Its Northern Range Margin Under Climatic Change." *Oikos* 118, no. 6: 829–836.

Lehmann, E., and J. P. Romano. 2022. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer International Publishing.

Marcus, R., P. Eric, and K. R. Gabriel. 1976. "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance." *Biometrika* 63, no. 3: 655–660.

McDermott, M. E., and L. W. DeGroote. 2016. "Long-Term Climate Impacts on Breeding Bird Phenology in Pennsylvania, USA." *Global Change Biology* 22, no. 10: 3304–3319.

McKean, J. W., and R. M. Schrader. 1984. "A Comparison of Methods for Studentizing the Sample Median." *Communications in Statistics - Simulation and Computation* 13, no. 6: 751–773.

Mittelbach, G. G., and B. J. McGill. 2019. *Community Ecology*. Oxford University Press.

Mukerjee, H., T. Robertson, and F. T. Wright. 1987. "Comparison of Several Treatments With a Control Using Multiple Contrasts." *Journal of the American Statistical Association* 82, no. 399: 902–910.

Munko, M., M. Ditzhaus, D. Dobler, and J. Genuneit. 2024. "RMST-Based Multiple Contrast Tests in General Factorial Designs." *Statistics in Medicine* 43, no. 10: 1849–1866.

Noguchi, K., R. S. Abel, F. Marmolejo-Ramos, and F. Konietschke. 2020. "Nonparametric Multiple Comparisons." *Behavior Research Methods* 52, no. 2: 489–502.

Olejnik, S., J. Li, S. Supattathum, and C. J. Huberty. 1997. "Multiple Testing and Statistical Power With Modified Bonferroni Procedures." *Journal of Educational and Behavioral Statistics* 22, no. 4: 389–406.

Ottensmann, M. 2022. *R-Package: ButeoAgeR*. https://github.com/mottensmann/ButeoAgeR 0.1.0, Bielefeld, Germany.

Pauly, M., E. Brunner, and F. Konietschke. 2015. "Asymptotic Permutation Tests in General Factorial Designs." *Journal of the Royal Statistical Society: Series B* 77, no. 2: 461–473.

Pigeot, I. 2000. "Basic Concepts of Multiple Tests—A Survey." *Statistical Papers* 41, no. 1: 3–36.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Roy, S. N. 1953. "On a Heuristic Method of Test Construction and Its Use in Multivariate Analysis." *Annals of Mathematical Statistics* 24, no. 2: 220–238.

Rubarth, K., P. Sattler, H. G. Zimmermann, and F. Konietschke. 2022. "Estimation and Testing of Wilcoxon–Mann–Whitney Effects in Factorial Clustered Data Designs." *Symmetry* 14, no. 2: 244.

Ruxton, G. D., and G. Beauchamp. 2008. "Time for Some A Priori Thinking About Post Hoc Testing." *Behavioral Ecology* 19, no. 3: 690–693.

Sattler, P., A. C. Bathke, and M. Pauly. 2022. "Testing Hypotheses About Covariance Matrices in General MANOVA Designs." *Journal of Statistical Planning and Inference* 219: 134–146.

Schuirmann, D. 1987. "A Comparison of the 2 One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics* 15, no. 6: 657–680.

Schumi, J., and J. T. Wittes. 2011. "Through the Looking Glass: Understanding Non-Inferiority." *Trials* 12, no. 1: 106.

Scott, I. A. 2009. "Non-Inferiority Trials: Determining Whether Alternative Treatments are Good Enough." *Medical Journal of Australia* 190, no. 6: 326–330.

Segbehoe, L. S., F. Schaarschmidt, and G. D. Djira. 2022. "Simultaneous Confidence Intervals for Contrasts of Quantiles." *Biometrical Journal* 64, no. 1: 7–19.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley.

Shaffer, J. P. 1986. "Modified Sequentially Rejective Multiple Test Procedures." *Journal of the American Statistical Association* 81, no. 395: 826–831.

Silverman, B. W. 1998. *Density Estimation for Statistics and Data Analysis*. Number 26 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Sippel, S., N. Meinshausen, E. M. Fischer, E. Székely, and R. Knutti. 2020. "Climate Change Now Detectable From Any Single Day of Weather at Global Scale." *Nature Climate Change* 10, no. 1: 35–41.

Tukey, J. W. 1994. "The Problem of Multiple Comparisons. Unpublished Manuscript Reprinted." In *The Collected Works of John W. Tukey*, Vol. 8, edited by H. I. Braun. Chapman & Hall.

Umlauft, M., M. Placzek, F. Konietschke, and M. Pauly. 2019. "Wild Bootstrapping Rank-Based Procedures: Multiple Testing in Nonparametric Factorial Repeated Measures Designs." *Journal of Multivariate Analysis* 171: 176–192.

van der Vaart, A., and J. Wellner. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag.

VanderWeele, T. J., and M. B. Mathur. 2019. "Some Desirable Properties of the Bonferroni Correction: Is the Bonferroni Correction Really So Bad?" *American Journal of Epidemiology* 188, no. 3: 617–618.

Verhulst, S., and J.-Å. Nilsson. 2007. "The Timing of Birds' Breeding Seasons: A Review of Experiments That Manipulated Timing of Breeding." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, no. 1490: 399–410.

Walls, S., and R. Kenward. 2020. *The Common Buzzard*. Bloomsbury Publishing.

Westfall, P. H., and S. S. Young. 1989. "P Value Adjustments for Multiple Tests in Multivariate Binomial Models." *Journal of the American Statistical Association* 84, no. 407: 780–786.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File 1: bimj70065-sup-0001-SuppMat.pdf.

Appendix A

A.1 | Covariance Estimators

Let $\mathbf{\Sigma}^{(i)} = (\mathbf{\Sigma}^{(i)}_{ab})_{a,b\in\{1,\dots,m\}}$ denote the covariance matrix of group $i \in \{1,\dots,k\}$. In our analysis, we consider the following three different covariance estimators for $\mathbf{\Sigma}^{(i)}_{ab}, a, b \in \{1,\dots,m\}$, as discussed in Ditzhaus et al. (2021):

i. **Kernel estimator**. The main idea of the kernel estimator is to replace the unknown densities in (5) by kernel density estimators. Therefore, let $K_i : \mathbb{R} \to [0,\infty)$ with $\int_{\mathbb{R}} K_i(x)\,\mathrm{d}x = 1$ denote a Lebesgue density, $h_{ni} \to 0$ as $n \to \infty$ a bandwidth, and

$$\widehat{f}_{K_i,i}(x) = (n_i h_{ni})^{-1} \sum_{i=1}^{n_i} K_i\left(\frac{x - X_{ij}}{h_{ni}}\right)$$

the kernel density estimator for f_i for all $i \in \{1,\dots,k\}$. Then, the kernel estimator for $\mathbf{\Sigma}^{(i)}_{ab}$ is given by

$$\widehat{\mathbf{\Sigma}}^{(i),K}_{ab} = \frac{n}{n_i} \frac{\min\{p_a,p_b\} - p_a p_b}{\widehat{f}_{K_i,i}(\widehat{q}_{ia})\widehat{f}_{K_i,i}(\widehat{q}_{ib})}.$$

ii. **Bootstrap estimator**. For the bootstrap estimator, we use the fact that the mean squared error of the bootstrapped sample quantile, which can be calculated as

$$\hat{\sigma}^*_i(p_r) := \left(n_i \sum_{j=1}^{n_i} (X^{(i)}_{j:n_i} - \hat{q}_{ir})^2 P_{ijr}\right)^{1/2} \quad \text{for all } r \in \{1,\dots,m\}$$

converges in probability to the asymptotic standard deviation of the corresponding sample quantile $\sqrt{\kappa_i \mathbf{\Sigma}^{(i)}_{rr}} = \sqrt{p_r - p_r^2}/f_i(q_{ir})$ (Ditzhaus et al. 2021), where $X^{(i)}_{j:n_i}$ denote the jth smallest element of the ordered ith sample and

$$P_{ijr} := B_{n_i,(j-1)/n_i}((-\infty, \lceil n_i p_r\rceil - 1]) - B_{n_i,j/n_i}((-\infty, \lceil n_i p_r\rceil - 1])$$

for $B_{n_i,p}$ denoting the binomial distribution with size parameter n_i and success probability p. Furthermore, eq. (8) in Ditzhaus et al. (2021) shows that $\mathbf{\Sigma}^{(i)}_{ab}$ only depends on $\mathbf{\Sigma}^{(i)}_{aa}$, $\mathbf{\Sigma}^{(i)}_{bb}$, p_a and p_b through

$$\mathbf{\Sigma}^{(i)}_{ab} = \sqrt{\mathbf{\Sigma}^{(i)}_{aa}\mathbf{\Sigma}^{(i)}_{bb}} \frac{\min\{p_a,p_b\} - p_a p_b}{\sqrt{(p_a - p_a^2)(p_b - p_b^2)}}. \tag{A1}$$

Thus, the bootstrap estimator for $\mathbf{\Sigma}^{(i)}_{ab}$ is given by

$$\widehat{\mathbf{\Sigma}}^{(i),B}_{ab} = \frac{n}{n_i}\hat{\sigma}^*_i(p_a)\hat{\sigma}^*_i(p_b)\frac{\min\{p_a,p_b\} - p_a p_b}{\sqrt{(p_a - p_a^2)(p_b - p_b^2)}}.$$

iii. **Interval-based estimator**. For the interval-based estimator, we use the extended estimator for the standard deviation of the pth sample quantile (Ditzhaus et al. 2021), which is motivated by an estimator of McKean and Schrader (1984) based on a standardized confidence interval. The extended estimator of Ditzhaus et al. (2021) is given by

$$\hat{\sigma}^{PB}_i(p) := n_i^{1/2} \frac{X^{(i)}_{u_i(p):n_i} - X^{(i)}_{l_i(p):n_i}}{2z_{1-\alpha^*_{n_i}(p)/2} + 2n_i^{-1/2}} \quad \text{for all } p \in (0,1),$$

where $l_i(p) := \max\{1, \lfloor n_i p - z_{1-\alpha/2}\sqrt{n_i p(1-p)}\rfloor\}$, $u_i(p) := \min\{n_i, \lfloor n_i p + z_{1-\alpha/2}\sqrt{n_i p(1-p)}\rfloor\}$, $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution and

$$\alpha^*_{n_i}(p) := 1 - \sum_{j=l_i(p)+1}^{u_i(p)-1} \binom{n_i}{j} p^j (1-p)^{n_i - j}.$$

In Ditzhaus et al. (2021), it is shown that $\hat{\sigma}^{PB}_i(p_a)$ is consistent for the asymptotic standard deviation of the corresponding sample quantile $\sqrt{\kappa_i \mathbf{\Sigma}^{(i)}_{aa}}$. By using (A1), we obtain the interval-based estimator for $\mathbf{\Sigma}^{(i)}_{ab}$

$$\widehat{\mathbf{\Sigma}}^{(i),PB}_{ab} = \frac{n}{n_i}\hat{\sigma}^{PB}_i(p_a)\hat{\sigma}^{PB}_i(p_b)\frac{\min\{p_a,p_b\} - p_a p_b}{\sqrt{(p_a - p_a^2)(p_b - p_b^2)}}.$$

A.2 | Details on the Bonferroni-Adjusted Permutation QANOVA

Here, we want to explain the details of the QANOVA permutation approach. Draw the permuted samples $X^\pi_{i1},\dots,X^\pi_{in_i}, i \in \{1,\dots,k\}$, without replacement from the pooled sample $X_{11},\dots,X_{1n_1},X_{21},\dots,X_{kn_k}$. As in Ditzhaus et al. (2021), let $F := \sum_{i=1}^k \kappa_i F_i$ denote the pooled cumulative distribution function and assume the following.

Assumption A.1. We assume that F is differentiable with uniformly continuous derivative f and that $f(F^{-1}(p_j)) > 0$ for all $j \in \{1,\dots,m\}$. Moreover, we assume $|n_i/n - \kappa_i| = O(n^{-1/2})$ as $n \to \infty$ for all $i \in \{1,\dots,k\}$.

Note that the latter assumption equals Assumption 4 in Ditzhaus et al. (2021). Similarly as Assumption 2.1, this assumption cannot really be checked in practice because usually F is not known. However, ties in the pooled data indicate that F cannot be continuous and, thus, not differentiable. The assumption $|n_i/n - \kappa_i| = O(n^{-1/2})$ guarantees that the group fractions converge sufficiently fast to their limits. Then, the permutation counterpart of the test statistics are defined as

$$T^\pi_n(\mathbf{h}_\ell) := \sqrt{n}\frac{\mathbf{h}'_\ell \widehat{\mathbf{q}}^\pi}{\sqrt{\mathbf{h}'_\ell \widehat{\mathbf{\Sigma}}^\pi \mathbf{h}_\ell}}, \quad \ell \in \{1,\dots,r\}.$$

By Lemmas S.1, S.2, and S.3 in the Supplement of Ditzhaus et al. (2021), we have

$$T^\pi_n(\mathbf{h}_\ell) \xrightarrow{d^*} \mathcal{N}(0,1) \tag{A2}$$

as $n \to \infty$ for each $\ell \in \{1,\dots,r\}$ under $|n_i/n - \kappa_i| = O(n^{-1/2})$ for all $i \in \{1,\dots,k\}$ whenever the kernel or interval-based covariance estimator is used, where here and throughout $\xrightarrow{d^*}$ denote conditional convergence in distribution in probability given the data $X_{11},X_{12},\dots,X_{21},\dots,X_{k1},\dots$. For the bootstrap estimator, we need the stronger assumption that $|n_i/n - \kappa_i| = o(n^{-1})$ holds, which means that the group fractions converge sufficiently fast to their limits. To show the consistency of the bootstrap estimator in this case, we first note that $X^\pi_{11},\dots,X^\pi_{kn_k} \sim \sum_{i=1}^k n_i/nF_i$ (unconditionally) independent and identically distributed. Hence, we can construct random variables $Y_{11},\dots,Y_{kn_k} \sim F$ independent and identically distributed with $\mathrm{P}(Y_{ij} \neq X^\pi_{ij}) \le |n_i/n - \kappa_i|$ for all $i \in \{1,\dots,k\}, j \in \{1,\dots,n_i\}$. Thus, we get

$$\mathrm{P}\Big(\exists j \in \{1,\dots,n_i\} \;:\; Y_{ij} \neq X^\pi_{ij}\Big) \le \sum_{j=1}^{n_i} |n_i/n - \kappa_i| \to 0$$

as $n \to \infty$ for all $i \in \{1,\dots,k\}$. Since the bootstrap estimator based on $Y_{11},\dots,Y_{kn_k}$ is consistent as discussed in Ditzhaus et al. (2021), it easily follows that the permutation counterpart of the bootstrap estimator is consistent as well. Mathematically, (A2) means

$$\sup_{x\in\mathbb{R}} \left|\mathrm{P}\big(T^\pi_n(\mathbf{h}_\ell) \le x \mid X_{11},\dots,X_{1n_1},X_{21},\dots,X_{kn_k}\big) - \Phi(x)\right| \xrightarrow{P} 0$$

as $n \to \infty$, where Φ denotes the standard normal distribution function. Thus, each test statistic $T^\pi_n(\mathbf{h}_\ell)$ can mimic the distribution of $T_n(\mathbf{h}_\ell, \epsilon_\ell)$ asymptotically. However, the joint distribution of the whole vector of test statistics $(T^\pi_n(\mathbf{h}_1),\dots,T^\pi_n(\mathbf{h}_r))$ turns out to converge to a centered normal distribution with different covariance matrix than in (8) in general. Hence, this approach is not able to mimic the joint distribution asymptotically and a correcting procedure for multiple testing, for example, a Bonferroni correction, needs to be applied. Therefore, let $q^\pi_{\ell,\beta}$

and $\widetilde{q}^{\pi}_{\ell,\beta}$ denote the β-quantiles of the conditional distribution of $T^{\pi}_n(\mathbf{h}_\ell)$ and $|T^{\pi}_n(\mathbf{h}_\ell)|$, respectively, given the data for all $\ell \in \{1,\dots,r\}$. By (A2), the quantiles are converging in probability to quantiles of the standard normal distribution or its absolute value, respectively. The Bonferroni-adjusted permutation tests can be obtained by setting $q_\ell = q^{\pi}_{\ell,1-\alpha/r}$ and $\widetilde{q}_\ell = \widetilde{q}^{\pi}_{\ell,1-\alpha/r}$ in Section 3.

A.3 | Details on the Asymptotic Multiple Contrast Testing Procedures (MCTPs)

In this section, we prove that the critical values in Section 3.2 converge in probability to the $(1-\alpha)$-quantiles of $\max_{\ell\in\{1,\dots,r\}} Z_\ell$ and $\max_{\ell\in\{1,\dots,r\}} |Z_\ell|$, respectively, for $(Z_1,\dots,Z_r)' \sim \mathcal{N}(0, \mathbf{DH\Sigma H'D})$. Therefore, we first state an auxiliary lemma.

Lemma A.1. *Let $(Z_1,\dots,Z_r) \sim F$, where $F : \mathbb{R}^r \to [0,1]$ denotes a continuous distribution function, and $(Y_{n1},\dots,Y_{nr})$ be a sequence of random vectors with $(Y_{n1},\dots,Y_{nr}) \xrightarrow{d^*} (Z_1,\dots,Z_r)$ as $n\to\infty$ conditionally on a random variable $\mathbf{X}$. Moreover, denote by G_n the conditional distribution function of $\max_{\ell\in\{1,\dots,r\}} Y_{n\ell}$ or $\max_{\ell\in\{1,\dots,r\}} |Y_{n\ell}|$ given $\mathbf{X}$ and by G the distribution function of $\max_{\ell\in\{1,\dots,r\}} Z_\ell$ or $\max_{\ell\in\{1,\dots,r\}} |Z_\ell|$, respectively. If G is strictly increasing on $[a,b] \subset \mathbb{R}$ with $G(a) < 1-\alpha < G(b)$ for $\alpha \in (0,1)$, we have $G_n^{-1}(1-\alpha) \xrightarrow{P} G^{-1}(1-\alpha)$.*

Proof. By the conditional convergence in distribution, we get $\max_{\ell\in\{1,\dots,r\}} Y_{n\ell} \xrightarrow{d^*} \max_{\ell\in\{1,\dots,r\}} Z_\ell$ and $\max_{\ell\in\{1,\dots,r\}} |Y_{n\ell}| \xrightarrow{d^*} \max_{\ell\in\{1,\dots,r\}} |Z_\ell|$ as $n\to\infty$ conditionally on $\mathbf{X}$ by the continuous mapping theorem. Since F is continuous, G is continuous. Hence, it follows $\sup_{t\in\mathbb{R}} |G_n(t) - G(t)| \xrightarrow{P} 0$. By Lemma S3 in the Supplement of Munko et al. (2024) together with the subsequence criterion, we obtain $G_n^{-1}(1-\alpha) \xrightarrow{P} G^{-1}(1-\alpha)$. □

In Section 3.2, the conditional convergence in distribution follows from the consistency of the covariance estimators. Hence, $q_{1-\alpha}$ and $\widetilde{q}_{1-\alpha}$ are converging in probability to the corresponding quantiles of $\max_{\ell\in\{1,\dots,r\}} Z_\ell$ and $\max_{\ell\in\{1,\dots,r\}} |Z_\ell|$, respectively.

A.4 | Details on the Bootstrap MCTPs

For the groupwise bootstrap MCTP, Theorem 3.6.1 in van der Vaart and Wellner (1996) implies $\sqrt{n_i}(\widehat{F}_i^* - \widehat{F}_i) \xrightarrow{d^*} B\circ F_i$ on $D(\mathbb{R})$ as $n\to\infty$ for all $i\in\{1,\dots,k\}$, where $D(\mathbb{R})$ denotes the Skorohod space on $\mathbb{R}$ equipped with the sup-norm and B denotes a Brownian bridge on $[0,1]$. By the delta method (van der Vaart and Wellner 1996, Theorem 3.9.11), it follows that we have conditional convergence in distribution $\sqrt{n}\left(\widehat{q}^*_{ij} - \widehat{q}_{ij}\right)_{j\in\{1,\dots,m\}} \xrightarrow{d^*} \mathbf{Z}_i$ as $n\to\infty$ for all $i\in\{1,\dots,k\}$ similarly as in the proof of Proposition 1 in the Supplement of Ditzhaus et al. (2021). Moreover, the consistency of the groupwise bootstrap counterpart of the kernel and interval-based covariance estimator follows as in Lemmas S.2 and S.3 in the Supplement of Ditzhaus et al. (2021) by just replacing $\widehat{F}^{\pi}_i$ by $\widehat{F}^*_i$ and f by f_i. Hence, combining everything with Slutsky's lemma and the continuous mapping theorem yields

$$(T^*_n(\mathbf{h}_1),\dots,T^*_n(\mathbf{h}_r))' \xrightarrow{d^*} \mathcal{N}(\mathbf{0}, \mathbf{DH\Sigma H'D})$$

as $n\to\infty$ whenever the kernel or interval-based covariance estimator is used. Hence, even the joint limit distribution in (8) can be approximated by the groupwise bootstrap. By Lemma A.1, $q^*_{1-\alpha}$ and $\widetilde{q}^*_{1-\alpha}$ are converging in probability to the corresponding quantiles of $\max_{\ell\in\{1,\dots,r\}} Z_\ell$ and $\max_{\ell\in\{1,\dots,r\}} |Z_\ell|$, respectively, whenever the kernel or interval-based covariance estimator is used.

A.5 | Algorithms for the Bootstrap MCTP (Algorithms 1 and 2)

ALGORITHM 1 | Bootstrap MCTP algorithm for the two-sided testing problem.

Input: Original samples $X_{i1},\dots,X_{in_i}$ for $i\in\{1,\dots,k\}$, probabilities $p_1,\dots,p_m$, contrasts $\mathbf{h}_1,\dots,\mathbf{h}_r$, constants $\varepsilon_1,\dots,\varepsilon_r$, global significance level α, and number of bootstrap samples B.
Calculate $\widehat{\mathbf{q}}$ and $\widehat{\mathbf{\Sigma}}$.
for $\ell = 1,\dots,r$ **do**
  Calculate the original test statistic $T_n(\mathbf{h}_\ell,\varepsilon_\ell) := \sqrt{n}\frac{\mathbf{h}'_\ell\widehat{\mathbf{q}}-\varepsilon_\ell}{\sqrt{\mathbf{h}'_\ell\widehat{\mathbf{\Sigma}}\mathbf{h}_\ell}}$.
end for
Bootstrap Procedure:
for $b = 1,\dots,B$ **do**
  Draw bootstrap samples $X^*_{i1},\dots,X^*_{in_i} \sim \hat{F}_i, i\in\{1,\dots,k\}$, independently conditionally on the data.
  Calculate $\widehat{\mathbf{q}}^*$ and $\widehat{\mathbf{\Sigma}}^*$ based on $X^*_{i1},\dots,X^*_{in_i}, i\in\{1,\dots,k\}$.
  for $\ell = 1,\dots,r$ **do**
    Calculate the bootstrap test statistic $T^*_n(\mathbf{h}_\ell) := \sqrt{n}\frac{\mathbf{h}'_\ell(\widehat{\mathbf{q}}^*-\widehat{\mathbf{q}})}{\sqrt{\mathbf{h}'_\ell\widehat{\mathbf{\Sigma}}^*\mathbf{h}_\ell}}$.
  end for
  Compute $\widetilde{M}_b := \max_{\ell\in\{1,\dots,r\}}|T^*_n(\mathbf{h}_\ell)|$.
end for
Estimate the quantile $\widetilde{q}^*_{1-\alpha}$ as empirical $(1-\alpha)$-quantiles of $\widetilde{M}_1,\dots,\widetilde{M}_B$.
Test Decisions:
for $\ell = 1,\dots,r$ **do**
  Reject $\mathcal{H}_{0,\ell}$ if and only if $|T_n(\mathbf{h}_\ell,\varepsilon_\ell)| > \widetilde{q}^*_{1-\alpha}$.
end for
Reject the global hypothesis $\mathcal{H}_0 = \bigcap_{\ell=1}^r \mathcal{H}_{0,\ell}$ if and only if $\max_{\ell\in\{1,\dots,r\}} |T_n(\mathbf{h}_\ell,\varepsilon_\ell)| > \widetilde{q}^*_{1-\alpha}$.
Output: Multiple test decisions of the bootstrap MCTP for the two-sided testing problem.

ALGORITHM 2 | Bootstrap MCTP algorithm for the noninferiority testing problem.

Input: Original samples $X_{i1}, \dots, X_{in_i}$ for $i \in \{1, \dots, k\}$, probabilities $p_1, \dots, p_m$, contrasts $\mathbf{h}_1, \dots, \mathbf{h}_r$, constants $\epsilon_1, \dots, \epsilon_r$, global significance level α, and number of bootstrap samples B.
Calculate $\widehat{\mathbf{q}}$ and $\widehat{\mathbf{\Sigma}}$.
for $\ell = 1, \dots, r$ **do**
  Calculate the original test statistic $T_n(\mathbf{h}_\ell, \epsilon_\ell) := \sqrt{n}\frac{\mathbf{h}_\ell' \widehat{\mathbf{q}} - \epsilon_\ell}{\sqrt{\mathbf{h}_\ell' \widehat{\mathbf{\Sigma}} \mathbf{h}_\ell}}$.
end for
Bootstrap Procedure:
for $b = 1, \dots, B$ **do**
  Draw bootstrap samples $X_{i1}^*, \dots, X_{in_i}^* \sim \widehat{F}_i, i \in \{1, \dots, k\}$, independently conditionally on the data.
  Calculate $\widehat{\mathbf{q}}^*$ and $\widehat{\mathbf{\Sigma}}^*$ based on $X_{i1}^*, \dots, X_{in_i}^*, i \in \{1, \dots, k\}$.
  for $\ell = 1, \dots, r$ **do**
    Calculate the bootstrap test statistic $T_n^*(\mathbf{h}_\ell) := \sqrt{n}\frac{\mathbf{h}_\ell'(\widehat{\mathbf{q}}^* - \widehat{\mathbf{q}})}{\sqrt{\mathbf{h}_\ell' \widehat{\mathbf{\Sigma}}^* \mathbf{h}_\ell}}$.
  end for
  Compute $M_b := \max_{\ell \in \{1,\dots,r\}} T_n^*(\mathbf{h}_\ell)$.
end for
Estimate the quantiles $q_{1-\alpha}^*$ as the empirical $(1-\alpha)$-quantile of $M_1, \dots, M_B$.
Test Decisions:
for $\ell = 1, \dots, r$ **do**
  Reject $\mathcal{H}_{0,\ell}^I$ if and only if $T_n(\mathbf{h}_\ell, \epsilon_\ell) > q_{1-\alpha}^*$.
end for
Reject the global hypothesis $\mathcal{H}_0^I = \bigcap_{\ell=1}^r \mathcal{H}_{0,\ell}^I$ if and only if $\max_{\ell \in \{1,\dots,r\}} T_n(\mathbf{h}_\ell, \epsilon_\ell) > q_{1-\alpha}^*$.
Output: Multiple test decisions of the bootstrap MCTP for the non-inferiority testing problem.

A.6 | Correlation Between the Test Statistics in Our Simulations

For a better understanding and investigation of the behavior of the different test procedures in our simulation study of Section 4, we report summaries of the correlations between the test statistics in this section. Therefore, we calculated the empirical correlations for the $N_{sim} = 5000$ test statistics resulting from the 5000 data sets used in the simulation for each setting. The same is done for the absolute values of the test statistics, which are used for the two-sided multiple testing problem (cf. Section 3). In order to get a broad overview over the correlations, the minimal (Min.), maximal (Max.), median, and minimal absolute (min abs.) correlations are reported in Table A1.

The Bonferroni correction is known to perform too conservative for a large positive correlation. The largest negative correlations are realized for the noninferiority tests for Tukey- and Grand-mean-type contrast matrix with correlations less than -0.52. For the Grand-mean-type contrast matrix, we do not observe a visible difference of the asymptotic MCTP compared to the Bonferroni-adjusted asymptotic test for the noninferiority testing problem in Figure 3. This is a consequence of the rather negative correlations between the test statistics for this scenario which yield a good performance of the Bonferroni correction.

TABLE A1 | Correlation summary between the different test statistics in our simulation study across all settings under the null hypothesis for the different scenarios.

Contrast matrix	Testing problem	Variance estimator	Min.	Max.	Median	Min. abs.
Dunnett	Two-sided	Bootstrap	0.067	0.666	0.276	0.067
Dunnett	Two-sided	Interval-based	0.063	0.653	0.263	0.063
Dunnett	Two-sided	Kernel	0.068	0.695	0.272	0.068
Dunnett	Noninferiority	Bootstrap	0.259	0.820	0.524	0.259
Dunnett	Noninferiority	Interval-based	0.258	0.814	0.520	0.258
Dunnett	Noninferiority	Kernel	0.256	0.823	0.513	0.256
Tukey	Two-sided	Bootstrap	−0.029	0.666	0.203	0.000
Tukey	Two-sided	Interval-based	−0.033	0.653	0.203	0.000
Tukey	Two-sided	Kernel	−0.030	0.695	0.205	0.001
Tukey	Noninferiority	Bootstrap	−0.593	0.820	0.247	0.000
Tukey	Noninferiority	Interval-based	−0.589	0.814	0.249	0.000
Tukey	Noninferiority	Kernel	−0.589	0.823	0.241	0.000
Grand-mean	Two-sided	Bootstrap	0.008	0.249	0.116	0.008
Grand-mean	Two-sided	Interval-based	0.006	0.256	0.122	0.006
Grand-mean	Two-sided	Kernel	0.007	0.308	0.133	0.007
Grand-mean	Noninferiority	Bootstrap	−0.525	0.130	−0.321	0.097
Grand-mean	Noninferiority	Interval-based	−0.532	0.131	−0.319	0.092
Grand-mean	Noninferiority	Kernel	−0.524	0.137	−0.322	0.096

Article 3

Baumeister, M., Thiel, K. E., Matits, L.,
Zimmermann, G., Pauly, M., & Sattler, P. (2025).

Multivariate and Multiple Contrast Testing
in General Covariate-adjusted Factorial Designs

arXiv:2506.15292.

Multivariate and Multiple Contrast Testing in General Covariate-adjusted Factorial Designs

Short title: Multiple Contrast Testing in MANCOVA

Marléne Baumeister[*,1,2], Konstantin Emil Thiel[1,3], Lynn Matits[4,5], Georg Zimmermann[3,6,7], Markus Pauly[1,2], Paavo Sattler[1]

June 19, 2025

Abstract

Evaluating intervention effects on multiple outcomes is a central research goal in a wide range of quantitative sciences. It is thereby common to compare interventions among each other and with a control across several, potentially highly correlated, outcome variables. In this context, researchers are interested in identifying effects at both, the global level (across all outcome variables) and the local level (for specific variables). At the same time, potential confounding must be accounted for. This leads to the need for powerful multiple contrast testing procedures (MCTPs) capable of handling multivariate outcomes and covariates. Given this background, we propose an extension of MCTPs within a semiparametric MANCOVA framework that allows applicability beyond multivariate normality, homoscedasticity, or non-singular covariance structures. We illustrate our approach by analysing multivariate psychological intervention data, evaluating joint physiological and psychological constructs such as heart rate variability.

[*]Corresponding author: e-mail: baumeister@statistik.tu-dortmund.de
[1]Department of Statistics, TU Dortmund University, Germany
[2]Research Center Trustworthy Data Science and Security, UA Ruhr, Germany
[3]Research Program Biomedical Data Science, Paracelsus Medical University Salzburg, Austria.
[4]Clinical & Biological Psychology, Institute of Psychology and Education, Ulm University, Ulm, Germany
[5]Sports and Rehabilitation Medicine, Department of Medicine, Ulm University Hospital, Ulm, Germany
[6]Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University, Salzburg, Austria
[7]Department of Artificial Intelligence and Human Interfaces, Paris Lodron University, Salzburg, Austria

Keywords: Multiple Testing, Covariate Adjustment, Multivariate Factorial Design, Interventional Study, Bootstrap.

1. Introduction

In biological, medical, and psychological research, there is a strong need for multiple testing methods, as such questions often arise in factorial design that are common for these fields. Typically, post hoc tests are considered after rejecting a global hypothesis when comparing more than two groups. This occurs, e.g., in clinical studies, or in ecological studies. An example of the latter is the comparison of the relative reproductive success (fitness) of birds grouped by sex and colour morph (Boerner & Krüger, 2009). On the other hand, in psychological intervention studies, there are often only two groups (intervention and control), but multiple, often highly correlated outcomes are measured. Indeed, correlated outcomes are inherent in the structure of intervention studies: in particular, if several related measures are taken from the same individual at a single time point, it is natural to assume a (strong) dependency between these outcome variables. In such settings, joint modelling of these outcomes is more suitable, and statistical testing usually becomes more reliable when it accounts for the underlying dependencies. Warne (2014) explicitly recommends using a multivariate model in such situations to avoid type I error inflation.

To exemplify the practical issues, we consider a synthetic dataset (Thiel et al., 2025) based on original data from the intervention-based *HypnoTreat* study conducted at the University of Ulm (Karrasch, Matits, et al., 2023; Karrasch, Mavioğlu, et al., 2023; Karrasch et al., 2022). The study examined the effects of a single relaxation hypnosis session on psychological and biological variables in chronically stressed individuals, such as heart rate variability. Here, the intervention-induced changes are observed in multiple variables measuring the same physiological construct. Consequently, we are interested in an overall global effect ("Does hypnosis influence HRV?") and specific local effects ("Does hypnosis influence a single specific parameter?"). This requires testing of multiple hypotheses. Moreover, modelling and inferring these questions jointly in one model, requires a multivariate approach. Additionally, the data set contains confounding covariates such as perceived chronic stress and suggestibility (ability to be hypnotised), which have to be accounted for. In fact, including covariates in a statistical analysis usually leads to an increase in power as predictive covariates can explain variability in the outcomes, which improves detection of factorial effects (Thiel et al., 2024). Therefore, it is recommended to adjust for covariates if they are of predictive character (Kahan et al., 2014). In line with this, also regulatory authorities (European Medicines Agency (EMA), 2015; U.S. Department of Health and Human Services Food and Drug Administration, 2023) recommend covariate-adjustment in randomized clinical trials.

The co-occurences of multivariate outcomes, predictive covariates, and multiple testing problems, motivate the adaption of multiple contrast testing procedures (MCTPs) to a semiparametric MANCOVA framework. MCTPs are a powerful solution for multiple testing as they redefine the rejection of the global hypothesis: the global hypothesis is rejected simultaneously if one of the local hypotheses is rejected. By construction, it is transparent which of the local tests are responsible for the global rejection and both local and global test decisions are coherent and consonant (Bretz et al., 2011). Traditional approaches, such as ANOVA followed by adjusted post hoc t-tests, do not have these advantages. Konietschke et al. (2013) pointed out that MCTPs provides more information about the process of rejecting the hypotheses. This is because the testing principle of MCTPs complies with the union-intersection principle introduced by Roy (1953). Moreover, as MCTPs are simultaneous testing procedures, they allow for the construction of compatible confidence intervals.

Furthermore, in many situations MCTPs turn out to be more powerful than methods of classical p-value adjustment like the Bonferroni procedure (Bretz et al., 2011). Because of this, MCTPs are known to be effective for many models and estimands. There are different methods for various univariate (Baumeister et al., 2025; Bretz et al., 2001; Hasler & Hothorn, 2008; Konietschke et al., 2013), multivariate (Hasler, 2014; Sattler et al., 2024) and even high dimensional and functional scenarios with metric outcomes (Konietschke et al., 2021; Munko et al., 2023). Moreover, there are even rank-based MCTPs for univariate outcomes (Konietschke et al., 2012; Noguchi et al., 2020), repeated measures (Umlauft et al., 2019), and other complex designs (Rubarth et al., 2022). Hasler and Hothorn (2011) and Hasler (2014) where the first to introduce MCTPs for multiple endpoints, but they did not allow for covariate adjustment and assume a parametric model. More recently, Becher et al. (2025) introduced covariate-adjusted MCTPs, but only allow for one univariate outcome. To reach our goal, we consider a more general semiparametric MANCOVA model, as studied in Zimmermann et al. (2020). The method allows for global testing in factorial designs and considers a covariate-adjusted mean as estimand. As it allows for covariance heteroscedasticity and some types of singularity, the model is very flexible. However, multiple testing was not considered so far. We close this gap, by proposing a MCTP for this general framework. In particular, our contribution is an MCTP for multivariate covariate-adjusted means

i. that is asymptotically valid in general semiparametric models without assuming normality, while

ii. allowing for potential covariance heteroscedasticity and singularity,

iii. that covers various local and global testing problems such as multivariate Dunnett- or Tukey-type hypotheses, especially in intervention designs.

To improve small sample performance, we consider two resampling approaches: parametric and wild bootstrapping. Both have demonstrated accurate performances, not only in Zimmermann et al. (2020), but also in other multivariate factorial designs (Friedrich & Pauly, 2018; Friedrich et al., 2017; Konietschke et al., 2015), and also in multiple testing (Munko et al., 2023, 2024; Umlauft et al., 2019).

The paper is structured as follows. In Section 2 we present the semiparametric MANCOVA model and state central limit theorems. All statistical methods are given in Section 3, including singularity-robust covariance estimator, the general multiple testing problem, and specific cases for concrete testing problems. In addition, we state and explain asymptotic guarantees for the bootstrap, the determination of proper critical values, and the resulting MCTP including local p-values. In Section 4, we evaluate the method's small-sample properties via extensive simulation. We thereby examine family-wise type I error rate (FWER) control and power of the MCTP in comparison with existing methods. Then, we present an illustrative data analysis based on the *HypnoTreat* study in Section 5. The paper closes with a discussion in Section 6.

2. Statistical Model and Set-Up

We consider a general MANCOVA set-up with d-dimensional random variables $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijd})'$ representing the outcome of individual $j \in \{1, \ldots, n_i\}$ in group $i \in \{1, \ldots, k\}$. Attached to each outcome vector there is a c-dimensional individual-specific covariate vector $\mathbf{z}_{ij} = (z_{ij1}, \ldots, z_{ijc})'$ and we pool all outcome and covariate vectors in the n-dimensional vector $\mathbf{Y} = (\mathbf{Y}'_{11}, \ldots, \mathbf{Y}'_{kn_k})'$ and the $n \times c$ matrix $\mathbf{Z} = (\mathbf{z}_{11}, \ldots, \mathbf{z}_{kn_k})'$, where $n := \sum_{i=1}^{k} n_i$. In the following, let $\mathbf{I}_d$ denote the d-dimensional identity matrix, $\mathbf{1}_d$ a d-dimensional vector and $\mathbf{J}_d$ a d-dimensional quadratic matrix containing only 1s. Moreover, $\oplus$ denotes the direct sum and $\otimes$ denotes the Kronecker product of matrices. Additionally, we introduce a vector of n error variables $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_{11}, \ldots, \boldsymbol{\epsilon}'_{kn_k})'$, $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijd})'$, a vector of k adjusted means $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_k)'$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})'$ and a vector of c regression coefficients $\boldsymbol{\nu} = (\boldsymbol{\nu}'_1, \ldots, \boldsymbol{\nu}'_c)'$, $\boldsymbol{\nu}_m = (\nu_{m1}, \ldots, \nu_{md})'$, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n_i\}$, $m \in \{1, \ldots, c\}$. Then our general semiparametric MANCOVA model is given by

$$\mathbf{Y} = \tilde{\mathbf{M}}\boldsymbol{\mu} + \tilde{\mathbf{Z}}\boldsymbol{\nu} + \boldsymbol{\epsilon},$$

where $\tilde{\mathbf{M}} = \bigoplus_{i=1}^{k} (\mathbf{1}_{n_i} \otimes \mathbf{I}_d)$ and $\tilde{\mathbf{Z}} = \mathbf{Z} \otimes \mathbf{I}_d$. Thereby, $\tilde{\mathbf{X}} = (\tilde{\mathbf{M}}, \tilde{\mathbf{Z}})$ is the *design matrix* of a linear model, where $\tilde{\mathbf{M}}$ characterises the factorial part and $\tilde{\mathbf{Z}}$ the regression part. As usual for a regression model, we assume that the errors in $\boldsymbol{\epsilon}_{ij}$ are independent with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and that a group-specific covariance matrix $\boldsymbol{\Sigma}_i := \mathrm{Cov}(\boldsymbol{\epsilon}_{ij})$, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n_i\}$ exists. These assumptions will later be part of Assumption (M1). From this it follows that the observations in $\mathbf{Y}$ are assumed to be independent and identically distributed per group.

As the regression coefficients in the vector $\boldsymbol{\nu}$ depends on the dimension ℓ, $\ell \in \{1, \ldots, d\}$, but not on the group i, $i \in \{1, \ldots, k\}$, they do not allow unequal regressions coefficients for different groups. Allowing for unequal regression coefficients leads to uninterpretable coefficients as the magnitude of the treatment effect is not the same at different levels of $\mathbf{Z}$ (Huitema, 2011). We also refer to Figure 11.1 in Huitema (2011). The covariance of $\boldsymbol{\epsilon}$ is given by $\boldsymbol{\Sigma} := \mathrm{Cov}(\boldsymbol{\epsilon}) = \bigoplus_{i=1}^{k} (\mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_i)$. As suggested in Zimmermann et al. (2020), the vector $\boldsymbol{\mu}$ can be estimated by the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\mu}} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_k')'$, that is

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{Y}}_{i\cdot} - \left(\bar{\mathbf{z}}_{i\cdot} \otimes \mathbf{1}_d'\right) \hat{\boldsymbol{\nu}}, \tag{1}$$

where the dot notation means averaging over all subjects in group i. The vector $\hat{\boldsymbol{\nu}} = (\hat{\boldsymbol{\nu}}_1', \ldots, \hat{\boldsymbol{\nu}}_c')'$ is the OLS estimator of $\boldsymbol{\nu}$, where $\hat{\boldsymbol{\nu}}_m = (\hat{\nu}_{m1}, \ldots, \hat{\nu}_{md})'$ for every $m \in \{1, \ldots, c\}$. To define $\hat{\boldsymbol{\nu}}$ in matrix notation we use the matrices $\mathbf{M} = \bigoplus_{i=1}^{k} \mathbf{1}_{n_i}$, $\mathbf{P}_M := \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'$ and $\mathbf{W} := (\mathbf{I}_n - \mathbf{P}_M)\mathbf{Z}$ and define $\hat{\boldsymbol{\nu}} = [(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' \otimes \mathbf{I}_d]\mathbf{Y}$, where $(\mathbf{I}_n - \mathbf{P}_M)$ adjusts the covariates $\mathbf{Z}$ in such a way that they are correctly multiplied with the factorial part and the classical multivariate OLS estimator is calculated by $\mathbf{W}$. To see the connection with the classical formulation of linear models, the OLS estimator for $\boldsymbol{\beta} = (\boldsymbol{\mu}', \boldsymbol{\nu}')'$ may also be written as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\nu}}')' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y}$.

In order to show the asymptotic behaviour of this estimator and to derive asymptotic MCTPs based thereon, the following assumptions are made (Zimmermann et al., 2020), where here and throughout all convergences are understood as $n \to \infty$.

(M1) The errors $\boldsymbol{\epsilon}_{ij}$ are independent and identically distributed in every group $i \in \{1, \ldots, k\}$ with $\mathrm{E}(\boldsymbol{\epsilon}_{ij}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{\Sigma}_i$ and $\mathrm{E}(\|\boldsymbol{\epsilon}_{ij}\|^4) < \infty$ for all $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n_i\}$.

(M2) The variance is positive, i.e. $\sigma_{i\ell}^2 := \mathrm{Var}(\epsilon_{ij\ell}) > 0$ for all $i \in \{1, \ldots, k\}$ and $\ell \in \{1, \ldots, d\}$.

(M3) The groups do not vanish, i.e. $n_i/n \to \kappa_i > 0$.

(M4) The matrix of covariates $\mathbf{Z}$ has full column rank, i.e. the columns of $\mathbf{Z}$ are linearly independent of each other, and they should be independent of the columns of $\bigoplus_{i=1}^{k} \mathbf{1}_{n_i}$.

(M5) $1/n_i \sum_{j=1}^{n_i} z_{ijm} \to \pi_{im} \in \mathbb{R}$ for all $i \in \{1, \ldots, k\}$ and $m \in \{1, \ldots, c\}$.

(M6) $1/n_i \sum_{j=1}^{n_i} \mathbf{z}_{ij}\mathbf{z}_{ij}' \to \Pi_i \in \mathbb{R}^{c \times c}$ for all $i \in \{1, \ldots, k\}$.

Note that (M1) does not postulate a specific distribution class (such as normality). In particular, the distributions can differ between groups. Singularity of the covariance matrix $\boldsymbol{\Sigma}$ is allowed through (M2). Assumption (M3) is a standard assumption in asymptotic

frameworks with several groups while (M4) avoids collinearity. As stated in Zimmermann et al. (2020), (M5) and (M6) have technical reasons.

We need the following central limit theorem, to derive the asymptotic behaviour of test statistics. It is proven in the Appendix of Zimmermann et al. (2020), see (A1) in the Proof of Theorem 1 therein. We state it here as a separate theorem.

Proposition 1. *Let* $\boldsymbol{\beta} = (\boldsymbol{\mu}', \boldsymbol{\nu}')'$ *and* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\nu}}')'$. *Then, under (M1), (M3), (M4), (M5) and (M6) it holds*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathbf{N} \sim \mathcal{N}(\mathbf{0}_{d\,(c+k)}, \boldsymbol{\Lambda}),$$

where $\boldsymbol{\Lambda} := \lim_{n\to\infty} n(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{\Sigma}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}$.

3. Statistical Methods

3.1. Heteroscedasticity- and Singularity-robust Estimation of $\boldsymbol{\Lambda}$

For statistical inference we are not only interested in an estimator for $\boldsymbol{\beta}$, but also in an estimator for $\boldsymbol{\Lambda}$, the limiting covariance matrix of $\boldsymbol{\beta}$. Consider the vector of residuals $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}}_1', \ldots, \hat{\boldsymbol{\epsilon}}_k')' = \mathbf{Y} - \tilde{\mathbf{M}}\hat{\boldsymbol{\mu}} - \tilde{\mathbf{Z}}\hat{\boldsymbol{\nu}}$ and, more precisely, for subject j in group i the residual $\hat{\boldsymbol{\epsilon}}_{ij} = \mathbf{Y}_{ij} - \hat{\boldsymbol{\mu}}_i - (\mathbf{z}_{ij}' \otimes \mathbf{I}_d)\hat{\boldsymbol{\nu}} \in \mathbb{R}^d$. We then define the squared residuals as

$$\hat{\boldsymbol{\Sigma}}_{ij} = \hat{\boldsymbol{\epsilon}}_{ij}\hat{\boldsymbol{\epsilon}}_{ij}'. \tag{2}$$

We now use the multivariate generalisation of the classical regression covariance estimator by Eicker (1963), i.e. we consider an adjusted block diagonal matrix of the matrices (2) as the mid part of a sandwich covariance estimator. As our model allows for heteroscedasticity, we adjust the sandwich estimator for that. Classical adjustments for heteroscedasticity are adaptable for multivariate outcomes and do not influence the statistical inference of this model as they converge to 1, see Welz et al. (2023) and Zimmermann et al. (2020).

Therefore, all of them are applicable in this framework. To follow the recommendation of Zimmermann et al. (2020), we propose a multivariate generalization of the HC4-adjustment (Cribari-Neto, 2004), which means that $(1 - p_{ij})^{-\delta_{ij}/2}$ is multiplied with every squared residual $\hat{\boldsymbol{\Sigma}}_{ij}$, where p_{ij} is the (i,j)-th diagonal element of the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n_i\}$ with $\delta_{ij} := \min\{4, p_{ij}/(n^{-1}\sum_{r=1}^{k}\sum_{s=1}^{n_r} p_{rs}\}$. In the general ANCOVA model studied in Zimmermann et al. (2019), the HC4-adjusted covariance

estimator was shown to be most effective compared to other heteroscedasticity-consistent approaches. Therefore, we consider the adjusted squared residuals

$$\hat{\boldsymbol{\Sigma}}_{ij}^{H} = (1 - p_{ij})^{-\delta_{ij}} \hat{\boldsymbol{\Sigma}}_{ij}$$

and the block diagonal matrix $\hat{\boldsymbol{\Sigma}} = \bigoplus_{i=1}^{k} \bigoplus_{j=1}^{n_i} \hat{\boldsymbol{\Sigma}}_{ij}^{H}$ for all $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n_i\}$. If this matrix is used as the center matrix of a sandwich covariance estimator (Eicker, 1963), the upper left $dk \times dk$ dimensional matrix block of

$$\hat{\boldsymbol{\Lambda}} = n(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\hat{\boldsymbol{\Sigma}}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \tag{3}$$

is a consistent estimator for $\boldsymbol{\Lambda}_{11}$, the upper left block of $\boldsymbol{\Lambda}$, and is defined as $\hat{\boldsymbol{\Lambda}}_{11}$. To be robust against singularity, we follow the idea of Friedrich and Pauly (2018), and only use the diagonal elements of (3) as covariance matrix estimator and zeros otherwise. We combine both ideas to get an heteroscedasty and singularity robust covariate adjusted covariance estimator by

$$\hat{\mathbf{D}} := \left(\hat{\boldsymbol{\Lambda}}_{11}\right)_0,$$

where the subscript zero means that only the diagonal elements are kept while all other matrix elements (on the off-diagonal) are set to zero. Combining Eicker (1963) with the Assumptions (M1)- (M6) implies that the estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ are consistent for $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, respectively. With the same argument it follows that the sandwich estimator $\hat{\boldsymbol{\Lambda}}$ is consistent for $\boldsymbol{\Lambda}$. Applying the Continuous Mapping Theorem, it thus follows that the covariance estimator $\hat{\mathbf{D}}$ is also consistent for $\mathbf{D} = (\boldsymbol{\Lambda}_{11})_0$, see also Theorem 1 and 2 in Zimmermann et al. (2020) for similar results.

3.2. Multiple Testing Problem

Within this framework we are able to formulate multiple testing problems consisting of r tests regarding the vector of adjusted means $\boldsymbol{\mu}$. To this end we consider r contrast vectors $\mathbf{h}_s = (\mathbf{h}_{s1}', \ldots, \mathbf{h}_{sk}')' \in \mathbb{R}^{kd}$, $\mathbf{h}_{si} = (h_{si1}, \ldots, h_{sid})'$ for all $s \in \{1, \ldots, r\}$. The vector $\mathbf{h}_s$ is a contrast vector iff $\sum_{i=1}^{k}\sum_{\ell=1}^{d} h_{si\ell} = 0$, and we combine all of them in the contrast matrix $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_r)' \in \mathbb{R}^{r \times kd}$. Then, the local hypotheses are $\mathcal{H}_{0,s} : \mathbf{h}_s'\boldsymbol{\mu} = \mathbf{0}$ for all $s \in \{1, \ldots, r\}$, defining a family of local null hypotheses

$$\Omega = \left\{\mathcal{H}_{0,s} : \mathbf{h}_s'\boldsymbol{\mu} = \mathbf{0},\ s \in \{1, \ldots, r\}\right\}. \tag{4}$$

This family corresponds to the global hypothesis $\mathcal{H}_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$. That $\mathbf{h}_s$ has to be a contrast vector is not necessary for mathematical reasons, but contrast vectors characterise the questions of interest.

Examples for Covered Multiple Testing Problems. The chosen family of hypotheses Ω covers various multiple testing problems of interest. In fact, we are able to consider the multivariate and covariate adjusted versions of well-known multiple testing problems (cf. Bretz et al., 2011; Konietschke et al., 2013) by choosing different matrices $\mathbf{H}$. To build a hypothesis matrix for a multivariate testing problem from the respective univariate one $\mathbf{H}_u$ we simply use the Kronecker product of $\mathbf{H}_u$ and $\mathbf{I}_d$, that is $\mathbf{H} = \mathbf{H}_u \otimes \mathbf{I}_d$. For example, there can be realized the following multivariate testing procedures with this technique:

1. **Multiple testing in a multivariate two-sample problem.** For $k = 2$ groups and an arbitrary dimension d we can compare the adjusted means $\mu_{i,\ell}$ for every endpoint: $\mathcal{H}_{0,\ell} : \mu_{1\ell} = \mu_{2,\ell}$, $\ell \in \{1, \ldots, d\}$. This multiple testing problem occurs if multiple correlated endpoints have to be analysed and the dimension-wise hypotheses are of interest. The hypothesis can be defined by the matrix $\mathbf{H} = (1, -1) \otimes \mathbf{I}_d$. This testing problem is considered in the data example in Section 5.

2. **Multivariate many-to-one comparison.** For arbitrary k and d the Kronecker product of the Dunnett-type matrix (Dunnett, 1955) and $\mathbf{I}_d$ leads to the hypotheses $\mathcal{H}_{0,i\ell} : \mu_{1\ell} = \mu_{i\ell}$, $i \in \{2, \ldots, k\}$, $\ell \in \{1, \ldots, d\}$, which compares component-wise the adjusted means of group 1 with the adjusted mean of all other groups.

3. **Multivariate all-pair comparison.** The equivalent use of the Tukey-type matrix (Tukey, 1994) gives the family of hypotheses including all pair-wise comparisons: $\mathcal{H}_{0,i_1 i_2 \ell} : \mu_{i_1\ell} = \mu_{i_2\ell}, i_1, i_2 \in \{1, \ldots, k\}$, $i_1 \neq i_2$, $\ell \in \{1, \ldots, d\}$.

4. **Multivariate grand-mean comparison.** The choose of the Grand-mean-type matrix introduced by Djira and Hothorn (2009) leads to a component-wise comparison of the adjusted group means with the overall mean of the group-wise adjusted means $\bar{\mu}_\ell := k^{-1} \sum_{i=1}^{k} \mu_{i\ell}$: $\mathcal{H}_{0,i} : \mu_{i\ell} = \bar{\mu}_\ell, i \in \{1, \ldots, k\}$, $\ell \in \{1, \ldots, d\}$.

To infer the local null hypothesis $\mathcal{H}_{0,s}$, we consider the local test statistics

$$A_n(\mathbf{h}_s) = \sqrt{n} \frac{\mathbf{h}_s' \hat{\boldsymbol{\mu}}}{\sqrt{\mathbf{h}_s' \hat{\mathbf{D}} \mathbf{h}_s}}. \tag{5}$$

For simultaneously testing (4) while adjusting for multiplicity, we must analyse the joint distribution of these statistics, i.e. the distribution of the vector $(A_n(\mathbf{h}_1), \ldots, A_n(\mathbf{h}_r))' = (\mathbf{H}\hat{\mathbf{D}}\mathbf{H})_0^{-1/2} \sqrt{n} \mathbf{H} \hat{\boldsymbol{\mu}} = \mathbf{A}_n(\mathbf{H})$. Its asymptotic distribution is given in the following theorem:

Theorem 2. *Under the Assumption of the semiparametric MANCOVA model (M1)-(M6), it holds:*

1. *Under* $\mathcal{H}_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ *the vector of test statistics* $\mathbf{A}_n(\mathbf{H})$ *converges in distribution to a multivariate normal distribution, i.e.*

$$\mathbf{A}_n(\mathbf{H}) = (A_n(\mathbf{h}_1), \ldots, A_n(\mathbf{h}_r))' \xrightarrow{d} \mathbf{B},$$

where $\mathbf{B} = (B_1, \ldots, B_r)$ *is* r*-dimensional and has the expectation* $\mathrm{E}(\mathbf{B}) = \mathbf{0}$ *and the covariance matrix*

$$\mathbf{R} := \mathrm{Cov}\,(\mathbf{B}) = \left(\mathbf{H}\mathbf{D}\mathbf{H}'\right)_0^{-\frac{1}{2}} \mathbf{H}\boldsymbol{\Lambda}_{11}\mathbf{H}' \left(\mathbf{H}\mathbf{D}\mathbf{H}'\right)_0^{-\frac{1}{2}}. \tag{6}$$

2. *Under* $\mathcal{H}_1 : \mathbf{H}\boldsymbol{\mu} \neq \mathbf{0}$, $\mathbf{A}_n(\mathbf{H})$ *converges in probability to* ∞.

The proof of Theorem 2 and all other proofs can be found in the Appendix A. It is a consequence of Theorem 2 that B_s, $s \in \{1, \ldots, r\}$, are normally distributed:

$$B_s \sim \mathcal{N}\left(0, \frac{\mathbf{h}_s'\boldsymbol{\Lambda}_{11}\mathbf{h}_s}{\mathbf{h}_s'\mathbf{D}\mathbf{h}_s}\right). \tag{7}$$

If we would follow the traditional approach of constructing MCTPs, we would use the asymptotic distribution of $\mathbf{B}$ to derive multivariate equicoordinate quantiles, and to define local and global test decisions. This idea follows from the union-intersection principle (Roy, 1953) and was made numerically available by Bretz et al. (2001). Due to standardization of the individual test statistics, a multivariate equicoordinate quantile, with equal values q_γ in all dimensions can be defined. Thereby, q_γ is the γ-quantile of the maximum tests statistic $\max_{s\in\{1,\ldots,r\}} |A_n(\mathbf{h}_s)|$ at the same time. However, this technique is not feasible in our framework since the test statistics $A_n(\mathbf{h}_s)$ have limiting distributions depending on $\mathbf{h}_s$, $s \in \{1, \ldots, r\}$. Consequently, it is not possible to compute equicoordinate quantiles. To overcome this problem we consider the idea of Munko et al. (2024) to adjust the level of significance for local tests such that the global level is controlled and the different distributions are taken into account. It is based on the concept of simultaneous confidence bands of Bühlmann (1998). For the adaption of this approach, we additionally consider asymptotically correct bootstrap methods presented in the following section.

We note, that it is possible to consider asymptotically valid MCTPs with equicoordinate quantiles for the covariate-adjusted mean by using $\hat{\boldsymbol{\Lambda}}_{11}$ instead of $\hat{\boldsymbol{D}}$. In the Supplementary Material we explain this in detail, and also discuss that this can be seen as an extension of the MCTPs from Hasler (2014) to multiple endpoints. However, these asymptotic MCTPs have the disadvantage, that they additionally require positive definite covariances $\boldsymbol{\Lambda}_i$, $i \in \{1, \ldots, k\}$, which is a stronger assumption than (M2). As described in Friedrich and Pauly (2018) and Zimmermann et al. (2020) singularity can occur in multivariate data if the outcome vector has strong linear dependencies. That is why we opted to focus on this more general framework.

3.3. Bootstrapping

To obtain suitable resampling methods we adapt the two bootstrap methods of Zimmermann et al. (2020), wild and a parametric bootstrap. The idea of the wild bootstrap is to produce variation by multiplying random variables to the residuals. This is why we first have to generate n independent identically distributed random variables T_{ij} independently from the data with $\mathrm{E}(T_{11}) = 0$, $\mathrm{Var}(T_{11}) = 1$ and $\sup_{i,j} \mathrm{E}(T_{ij}^4) < \infty$ for $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n_i\}$. Methodologically, we can consider some T_{ij} which fulfils this conditions, but practically we have to decide for some specific ones. As they turned out to be successful in Zimmermann et al. (2020) and the choice of weights do not seem to have much influence on the methods performance, we choose Rademacher random variables for T_{ij}, which means that $\mathrm{P}(T_{11} = -1) = \mathrm{P}(T_{11} = 1) = 1/2$. Then, the elements of the wild bootstrap sample are defined as

$$\mathbf{Y}_{ij}^* := \frac{T_{ij}}{\sqrt{1 - p_{ij,ij}}} \hat{\boldsymbol{\epsilon}}_{ij}$$

for every $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n_i\}$. This process is carried out for every subject and does not depends on the component $\ell \in \{1, \ldots, d\}$, which receives the dependence structure within the subjects. With the bootstrap sample we can generate the wild bootstrap OLS estimator $\hat{\boldsymbol{\beta}}^* := (\hat{\boldsymbol{\mu}}^{*\prime}, \hat{\boldsymbol{\nu}}^{*\prime})' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y}^*$ and the wild bootstrap covariance estimator

$$\hat{\mathbf{D}}^* := \left(\hat{\boldsymbol{\Lambda}}_{11}^*\right)_0 = \left(n(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\hat{\boldsymbol{\Sigma}}^*\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\right)_0,$$

where $\hat{\boldsymbol{\Sigma}}^* = \bigoplus_{i=1}^k \bigoplus_{j=1}^{n_i} \hat{\boldsymbol{\Sigma}}_{ij}^{*H}$ and $\hat{\boldsymbol{\Sigma}}_{ij}^{*H}$ are the heteroscedasticity-robust squared residuals calculated with the wild bootstrap data for every $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n_i\}$. From the second part of Theorem 2 in Zimmermann et al. (2019) the consistency of $\hat{\mathbf{D}}^*$, i.e., $\hat{\mathbf{D}}^* \xrightarrow{P} \mathbf{D}$ can be concluded, because they argue that the wild bootstrap residuals $\hat{\boldsymbol{\epsilon}}_{ij}^*$ are consistent. To define a wild bootstrap version of the test statistic $A_n^*(\mathbf{h}_s)$ we simply use the wild bootstrap estimators $\boldsymbol{\mu}^*$ and $\hat{\mathbf{D}}^*$ instead of the original estimators in formula (5), $s \in \{1, \ldots, r\}$:

$$A_n^*(\mathbf{h}_s) = \sqrt{n} \frac{\mathbf{h}_s' \hat{\boldsymbol{\mu}}^*}{\sqrt{\mathbf{h}_s' \hat{\mathbf{D}}^* \mathbf{h}_s}}. \tag{8}$$

Note that the s test statistics are calculated with the same bootstrap sample. The following Theorem states that the asymptotic distribution of the wild bootstrap test statistic $A_n^*(\mathbf{h}_s)$ is the same as the distribution of $A_n(\mathbf{h}_s)$ under $\mathcal{H}_{0,s}$.

Theorem 3. *Let $s \in \{1, \ldots, r\}$. The wild bootstrap test statistic $A_n^\star(\mathbf{h}_s)$ given in (8) converges conditionally given the data in distribution to the real-valued random vector $\mathbf{B}$,*

which characterises also the asymptotic distribution of $A_n(\mathbf{h}_s)$ *under* $\mathcal{H}_{0,s}$ *(see Theorem 2), i.e.,*

$$\sup \left| \mathrm{P}\left(A_n^*\left(\mathbf{h}_s\right) \leq x | \mathbf{Y}\right) - \mathrm{P}\left(\mathbf{B} \leq x\right) \right| \xrightarrow{p} 0. \tag{9}$$

The idea of the parametric bootstrap approach is to estimate the group-wise covariance and use these estimators to generate independent d-dimensional observation vectors from the normal distribution:

$$\mathbf{Y}_{ij}^{\star} \sim \mathcal{N}\left(\mathbf{0}, \hat{\mathbf{\Sigma}}_i\right),$$

for every group $i \in \{1, \ldots, k\}$ and every subject $j \in \{1, \ldots, n_i\}$, where

$$\hat{\mathbf{\Sigma}}_i = \frac{1}{n_i - c - 1} \sum_{j=1}^{n_i} \hat{\boldsymbol{\epsilon}}_{ij} \hat{\boldsymbol{\epsilon}}_{ij}'$$

for every group $i \in \{1, \ldots, k\}$. Analogously to the wild bootstrap approach we can calculate the parametric bootstrap versions $\hat{\boldsymbol{\beta}}^{\star}$ and $\hat{\mathbf{D}}^{\star}$ of the estimators and the test statistic

$$A_n^{\star}(\mathbf{h}_s) = \sqrt{n} \frac{\mathbf{h}_s' \hat{\boldsymbol{\mu}}^{\star}}{\sqrt{\mathbf{h}_s' \hat{\mathbf{D}}^{\star} \mathbf{h}_s}}. \tag{10}$$

for $s \in \{1, \ldots, r\}$. Again, the s test statistics are calculated from the same bootstrap sample. In the Proof of Theorem 4 of Zimmermann et al. (2020), it is shown that $\hat{\mathbf{D}}^{\star}$ is a consistent estimator for $\mathbf{D}$. And similar to the wild bootstrap approach we are able to state that the asymptotic distribution of the parametric bootstrap test statistic $A_n^{\star}(\mathbf{h}_s)$ is the same as the distribution of $A_n(\mathbf{h}_s)$ under $\mathcal{H}_{0,s}$.

Theorem 4. *Let* $s \in \{1, \ldots, r\}$. *The parametric bootstrap test statistic* $A_n^{\star}(\mathbf{h}_s)$ *given in (10) converges conditionally given the data in distribution to the real-valued random vector* $\mathbf{B}$*, which characterises also the asymptotic distribution of* $A_n(\mathbf{h}_s)$ *under* $\mathcal{H}_{0,s}$ *(see Theorem 2), i.e.,*

$$\sup \left| \mathrm{P}\left(A_n^{\star}\left(\mathbf{h}_s\right) \leq x | \mathbf{Y}\right) - \mathrm{P}\left(\mathbf{B} \leq x\right) \right| \xrightarrow{p} 0. \tag{11}$$

3.4. Determination of Critical Values

To get suitable critical values for the test decision in the semiparametric MANOVA model we draw B bootstrap samples with one of the methods above. Consider for every sample $b \in \{1, \ldots, B\}$ the vector of test statistics $(A_n^{\circ,b}(\mathbf{h}_1), \ldots, A_n^{\circ,b}(\mathbf{h}_r))$, $\circ \in \{*, \star\}$. Define $q_{s,1-\gamma}^{\circ}$, the $(1-\gamma)$-quantile of $|A_n^{\circ,1}(\mathbf{h}_s)|, \ldots, |A_n^{\circ,B}(\mathbf{h}_s)|$, $s \in \{1, \ldots, r\}$, $\circ \in \{*, \star\}$. The

idea of Munko et al. (2024) is to adjust the significance level γ for each local test such that the level α is controlled globally. For that, define the estimated family-wise type I error rate with the critical value $q^{\circ}_{r,\gamma}$:

$$\mathrm{FWER}^{\circ}_{n}(\gamma) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{\exists\, s \in \{1,\dots,r\} : |A^{\circ,b}_{n}(\mathbf{h}_s)| > q^{\circ}_{s,1-\gamma}\right\}$$

for $\circ \in \{*, \star\}$ and $\gamma \in [0,1]$. Then, Munko et al. (2024) define the adjusted level $\gamma_n(\alpha)$ as:

$$\gamma_n(\alpha) := \max\left\{\gamma \in \left\{0, \frac{1}{B}, \dots, \frac{B-1}{B}\right\} | \mathrm{FWER}^{\circ}_{n}(\gamma) \leq \alpha\right\}, \tag{12}$$

which means that $\gamma_n(\alpha)$ is the largest value such that $\mathrm{FWER}^{\circ}_{n}(\gamma)$ is bounded by the global level of significance α. Here, the maximum is evaluated over the set $\{0, 1/B, \dots, (B-1)/B\}$, because the quantiles can only take B different values. With the adjusted level $\gamma_n(\alpha)$ we are able to formulate some decision rules regarding the local and global hypotheses for the two different bootstrap approaches. For every $s \in \{1,\dots,r\}$ and $\circ \in \{*, \star\}$ the hypothesis $\mathcal{H}_{0,s}$ of Ω is rejected if and only if $|A_n(\mathbf{h}_s)| > q^{\circ}_{s,1-\gamma_n(\alpha)}$ or equivalently $|A_n(\mathbf{h}_s)|/q^{\circ}_{s,1-\gamma_n(\alpha)} > 1$, i.e. we can define tests

$$\varphi^{\circ}_{n,s} = \mathbb{1}\{|A_n(\mathbf{h}_s)| > q^{\circ}_{s,1-\gamma_n(\alpha)}\}.$$

It is possible to construct simultaneous confidence intervals from the multiple testing procedure for $\mathbf{h}_s'\boldsymbol{\mu}$ with the global confidence level $1-\alpha$:

$$\left[\mathbf{h}_s'\hat{\boldsymbol{\mu}} \pm q^{\circ}_{s,1-\gamma_n(\alpha)} \frac{\sqrt{\mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s}}{\sqrt{n}}\right].$$

In line with the classic MCTPs the global hypothesis $\mathcal{H}_0$ is rejected, if and only if at least one $\mathcal{H}_{0,s}$ is rejected. This leads to the global test

$$\varphi^{\circ}_{n} = \max_{s\in\{1,\dots,r\}} \mathbb{1}\left\{\frac{|A_n(\mathbf{h}_s)|}{q^{\circ}_{s,1-\gamma_n(\alpha)}} > 1\right\}, \circ \in \{*, \star\}.$$

This formulation incorporates the different distributions of $A_n(\mathbf{h}_s)$, $s \in \{1,\dots,r\}$ by considering individual quantiles $q^{\circ}_{s,1-\gamma_n(\alpha)}$ as critical values, but uses the same level of significance $\gamma_n(\alpha)$ for all tests. To ensure, that the level of significance of the global test and the family-wise type I error rate of Ω is controlled asymptotically we state the following Theorem:

Theorem 5. *Let $T \subset \{1,\dots,r\}$ denote the subset of true hypotheses $\mathcal{H}_{0,s}$ of Ω. Then, φ°_{n}, $\circ \in \{*, \star\}$ is an asymptotic level-α test, i.e. with $B = B(n) \to \infty$ as $n \to \infty$, it holds*

that

$$\lim_{n\to\infty} \mathrm{P}\left(\exists\, s \in T:\ |A_n(\mathbf{h}_s)| > q^{\circ}_{s,1-\gamma_n(\alpha)}\right) \leq \alpha,$$

where equality holds, if $T = \{1, \ldots, r\}$.

Theorem 3 and Theorem 4 state that the conditional distributions of $A^{\circ}_n(\mathbf{h}_s)$, $\circ \in \{*, \star\}$ are asymptotically the same as the distribution of $A_n(\mathbf{h}_s)$ under $\mathcal{H}_{0,s}$. Under $\mathcal{H}_{1,s}$ we have the situation that the test statistic $A_n(\mathbf{h}_s)$ is divergent and and $A^{\circ}_n(\mathbf{h}_s)$, $\circ \in \{*, \star\}$ has still the same normal distribution. Therefore, the tests $\varphi^{\circ}_{n,s}$ and φ°_n are consistent.

We considered also a version of this tests where we do not assume that the test statistics $A^{\circ}_n(\mathbf{h}_s)$, $\circ \in \{*, \star\}$, are symmetrically distributed and therefore do not use the absolute value of the test statistics, but two asymmetric critical values from an optimized local level that takes the asymmetry into account. It turns out that this method is not superior to the symmetric version. That is why we state the methodology and some simulation results in the Supplementary Material.

3.5. P-Values

In this section, we are going to introduce local and global p-values for the bootstrap MCTPs. For the asymptotic MCTPs sketched in Section 3.2, we refer to the Supplementary Material. Moreover, we are going to show that the comparison of these p-values against an adjusted significance level provides an equivalent way of expressing both the local test $\varphi^{\circ}_{n,s}$ and the global test φ°_n. Comparing p-values with an adjusted level of significance comes with a noteworthy advantage: while the test decision of $\varphi^{\circ}_{n,s}$ through the critical values is based on a different value $q^{\circ}_{s,1-\gamma_n(\alpha)}$ for each test $s \in \{1, \ldots r\}$, p-value-based test decisions require only one adjusted significance level for all tests. In case of the bootstrap MCTPs, this adjusted significance level is defined as $\gamma_n(\alpha)$. In contrast to classical p-values, it has to be calculated dependently from α. Notably, the consideration of an adjusted level of significance is conceptually analogous to Bonferroni-adjustment (Dunn, 1961), where an adjusted significance level is obtained by dividing the global significance level α by the number of tests. This local level can be used to get a test decision, either it is used to calculate a critical value or a p-value. This analogy eases comparability between MCTPs and Bonferroni, and therefore, will subsequentially help us illustrating MCTPs (cf. Section 5).

We use the definition of Munko et al. (2024) to introduce local p-values for bootstrap

MCTPs as follows:

$$p_{n,s} := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{|A_n^{\circ,b}(\mathbf{h}_s)| \geq |A_n(\mathbf{h}_s)|\right\} \tag{13}$$

To give an intuition, $p_{n,s}$ is the fraction of bootstrap runs, for which the bootstrap test statistic $A_n^{\circ,b}(\mathbf{h}_s)$ is at least as large as $A_n(\mathbf{h}_s)$ in absolute values. Recall from Theorem 3 and Theorem 4 that $A_n^{\circ,b}(\mathbf{h}_s)$ asymptotically mimics the distribution of $A_n(\mathbf{h}_s)$ under $\mathcal{H}_{0,s}$. Therefore, it is easy to see that the definition of $p_{n,s}$ is in line with the well-established interpretation of p-values as the probability that the test statistic assumes a value at least as extreme as the one observed when the null hypothesis holds (e.g. Woodward, 2013). Eventually, a global p-value is obtained by $p_n := \min\{p_{n,1}, \ldots, p_{n,r}\}$. The following proposition shows equivalence to the tests defined above:

Proposition 6.

(i) For each $s \in \{1, \ldots r\}$, it holds $p_{n,s} \leq \gamma_n(\alpha)$ if and only if $\varphi_{n,s}^{\circ} = 1$,

(ii) It holds that $p_n \leq \gamma_n(\alpha)$ if and only if $\varphi_n^{\circ} = 1$.

4. Simulations

In order to analyse the small sample performance of the developed testing procedures we did an extensive simulation study on the Linux HPC cluster of TU Dortmund University (LiDo3) via the computing environment R, version 4.2.1 R Core Team (2022). For each simulation scenario we considered 5000 simulation runs, 2000 bootstrap iterations and the level of significance $\alpha = 0.05$. We chose a simulation set-up similarly to Zimmermann et al. (2020), accordingly, the basis of our simulations is the following data-generation process:

$$\mathbf{Y}_i = \boldsymbol{\mu}_i + \mathbf{Z}_i\boldsymbol{\nu} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i = \boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{X}_i,\ i \in \{1, \ldots, k\}.$$

Here, $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$ is a matrix of standardized random data from different distributions: standard normal distribution (N), t-distribution with 3 degrees of freedom (t_3), χ^2- distribution with 3 degrees of freedom (χ_3^2), logarithmic standard normal distribution (LN) and double exponential distribution (DExp). We chose the balanced and unbalanced sample sizes $\mathbf{n}^{(1)} = (10, \ldots, 10)' \in \mathbb{R}^k$, $\mathbf{n}^{(2)} = (20, 10, \ldots, 10)' \in \mathbb{R}^k$, $\mathbf{n}^{(3)} = (10, \ldots, 10, 20)' \in \mathbb{R}^k$ and its multiples $K \cdot \mathbf{n}^{(t)}, t \in \{1, 2, 3\}$. The matrix $\boldsymbol{\Sigma}_i^{\frac{1}{2}} \in \mathbb{R}^{d \times d}$ is the square-root of a certain covariance matrix. We consider different scenarios:

1. Homoscedastic covariance: $\boldsymbol{\Sigma}_i = \mathbf{I}_d + 0.5 \cdot (\mathbf{J}_d - \mathbf{I}_d)$, $i \in \{1, \ldots, k\}$.

2. Heteroscedastic covariance: $\boldsymbol{\Sigma}_i = \mathbf{I}_d + 0.5 \cdot (\mathbf{J}_d - \mathbf{I}_d)$ for $i \in \{1, \ldots, k-1\}$ and $\boldsymbol{\Sigma}_k = 2 \cdot \mathbf{I}_d + 0.5 \cdot (\mathbf{J}_d - \mathbf{I}_d)$.

3. Singular covariance: $\boldsymbol{\Sigma}_i = \begin{cases} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.25 \end{pmatrix}, & d = 2 \\ \begin{pmatrix} 6 & 3 & 3 \\ 3 & 2 & 3 \\ 3 & 3 & 6 \end{pmatrix}, & d = 3 \\ \begin{pmatrix} 6 & 3 & 3 & 3 \\ 3 & 6 & 3 & 3 \\ 3 & 3 & 2.5 & 3 \\ 3 & 3 & 3 & 6 \end{pmatrix}, & d = 4 \end{cases}, i \in \{1, \ldots, k\}.$

If the second covariance setting is combined with the sample sizes $K \cdot \mathbf{n}^{(2)}$ and $K \cdot \mathbf{n}^{(3)}$ the situations of negative respective positive pairing occur, where the group with the smaller sample size has the bigger respective the smaller variance. The number of covariates is set to $c = 2$ for all simulations and the matrix of covariates $\mathbf{Z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}) \in \mathbb{R}^{n \times c}$ is drawn from uniform distributions separately: $\mathbf{z}_{i1} \sim \mathcal{U}((-10, 10))$ and for $\mathbf{z}_{i2}$ the first half is drawn from $\mathcal{U}((0, 5))$ and the second half from $\mathcal{U}((-2, -1))$. Additionally, we accept the generated covariates only under certain conditions of dispersion, see the R-script `covariates.R` in the Supplementary Material for further details. The regression coefficients $\boldsymbol{\nu} \in \mathbb{R}^{c \times d}$ are set to $\boldsymbol{\nu} = (-0.5, \mathbf{1}'_{d-2}, -1 | 1.5, 2 \cdot \mathbf{1}'_{d-2}, 3)$. This choice is in line with the simulation set-up in Zimmermann et al. (2020). For $d = 2$ they did a small simulation to ensure the association between the covariates and the selected outcomes with that choice of $\boldsymbol{\nu}$. Additionally, we did a similar simulation for dimensions $d \in \{3, 4\}$ by fitting univariate models and checking if the covariates are significant at the level of 5%. We chose a sample size of 40 per group and did 1000 simulation runs. Especially, this choice of $\boldsymbol{\nu}$ ensures the linear relationship between the components 1 and d of the outcomes in the singular covariance setting.

In all simulation runs, we compare the method of multiple contrast test procedures with wild (*MCTP-wild*) and parametric (*MCTP-param*) bootstrap with the asymptotic multiple contrast test procedure explained in Section 3.2 (*MCTP-norm*, *MCTP-t*) and with the MANCATS by Zimmermann et al. (2020). The latter is also considered with wild (*MANCATS-wild*) and parametric (*MANCATS-param*) bootstrap and we made it comparable in our multiple testing problem by simply using the Bonferroni-adjustment (Dunn, 1961). Note that the asymptotic MCTPs are not defined in the singular covariance setting (3). Therefore, we compare six methods in this simulation study. We simulated also the asymmetric MCTPs explained in Section 3.4 in all scenarios. These methods are not considered throughout this paper but there are some analyses in the Supplementary Material. All simulation results can be found in the Supplementary Material.

Testing Problem	Groups	Dimensions
Two-sample	2	$2, 3, 4$
Dunnett	3	$2, 3, 4$
Dunnett	4	$2, 3$
Tukey	3	$2, 3$
Tukey	4	2

Table 1: Combinations of considered testing problems, groups k and dimensions d in the simulation study regarding FWER-control. For the testing problems, see the explanation in Section 3.2.

4.1. Simulation Results Under the Null Hypothesis

For simulations under the null hypothesis we ensure that our data fulfils the global null hypothesis $\mathcal{H}_0$ and set $\boldsymbol{\mu}_i = \mathbf{0}$ for all groups $i \in \{1, \ldots, k\}$. In the analysis of the control of the family-wise type I error rate (FWER) we consider small and moderate sample sizes, consequently, the choice of the sample size multiplier K is widely and the sample are multiplied with $K \in \{1, 2, 3, 4\}$ as explained above. To get a comprehensive overview about the FWER-control of the testing procedures, we considered various numbers of groups and dimensions as well as different testing problems. In Table 1, we present the combinations we included in our simulation study. As an example we present in the paper plots regarding the simulations settings considering Dunnett's testing problem and $d \in \{2, 3, 4\}$ dimensions and $k = 3$ groups. In Figure 1 this simulations results are presented as boxplots (from R-package `ggplot2`, see Wickham, 2016) split by the sample size multiplier K. Here, it is observable that the empirical FWER of the resampling MCTPs is in the 95% binomial interval $[4.4, 5.6]$ for most simulation settings even for the smaller samples. The MCTP with parametric bootstrap tends to be a little bit more conservative than the alternative with wild bootstrap in most settings. The asymptotic MCTPs tend to be liberal for smaller sample sizes, this effect is stronger for the asymptotic MCTPs with multivariate normal distribution. The Bonferroni-adjusted MANCATS tend to a conservative behaviour as expected. For all methods, there are a few settings with a highly liberal behaviour. In Figure 2 the subgroup of settings with negative pairing are plotted as bee swarms (Eklund & Trimble, 2021) split by the distributions and we can see that liberal behaviour occurs in settings with negative pairing and the χ^2_3- or the logarithmic normal distribution. A moderately liberal behaviour can also be observed in the plot for the normal distribution. Here, the resampling MCTPs are less liberal than the asymptotic one with t-distribution. This liberal behaviour is not surprising as it is also observed in other multivariate semiparametric models for factorial designs that negative pairing can be a problem, see for example Konietschke et al. (2015). In the logarithmic normal and in the t_3-distributed settings of the presented data in Figure 1, the resampling MCTPs tend to a conservative behaviour (empirical FWER smaller than 4.4). As the MCTP with parametric bootstrap tends to be a bit more conservative than

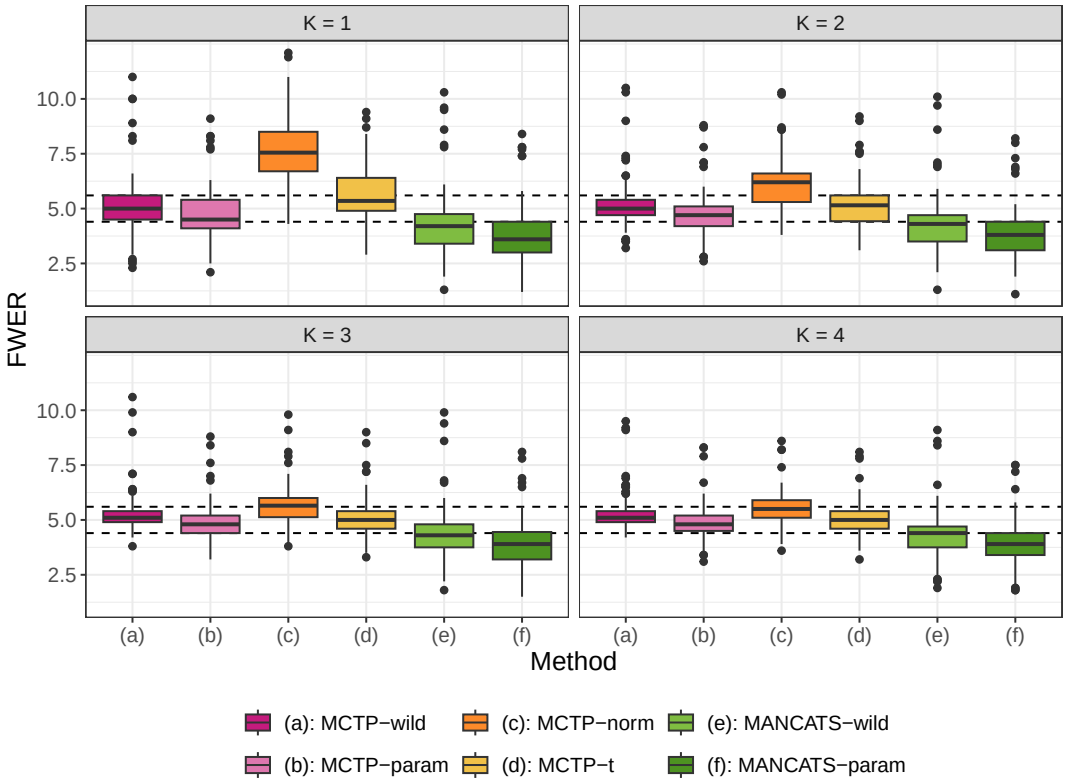

Figure 1: Empirical family-wise type error rate in % for Dunnett's testing problem with $k = 3$ and $d \in \{2, 3, 4\}$ split by the sample size multiplier $K \in \{1, 2, 3, 4\}$, i.e., $\mathbf{n} = K \cdot \mathbf{n}^{(t)}$, $t \in \{1, 2, 3\}$.

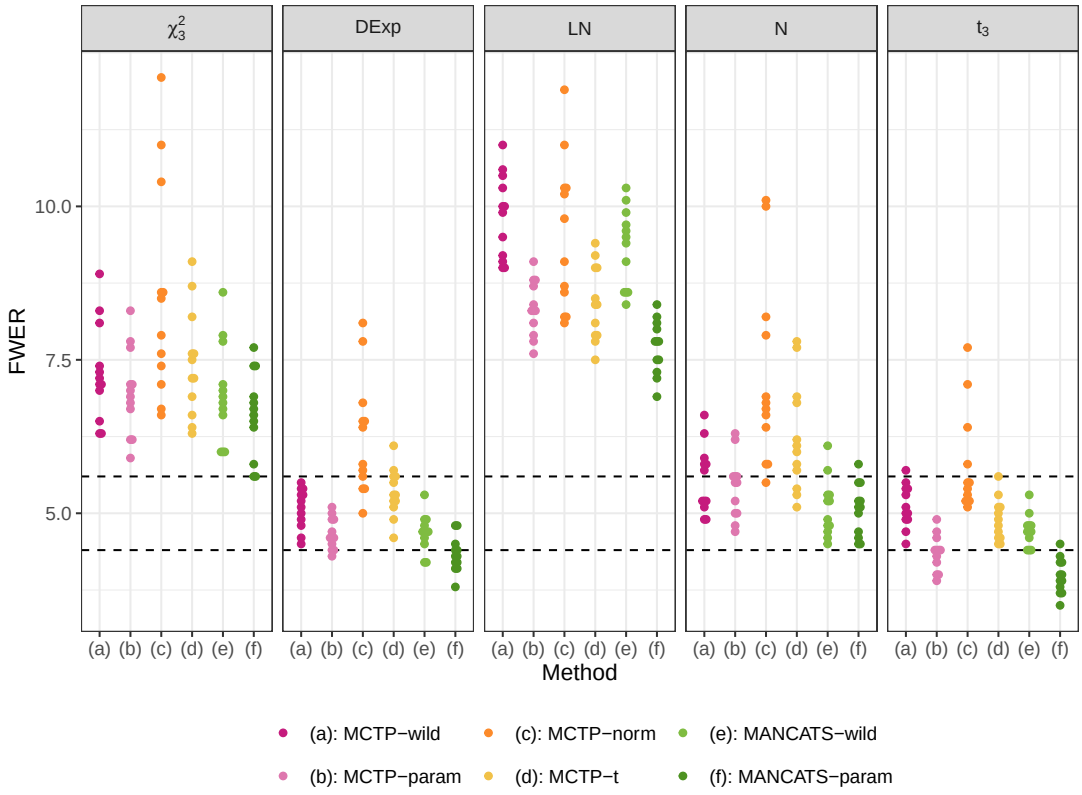

Figure 2: Empirical family-wise type I error rate in % for Dunnett's testing problem with $k = 3$ and $d \in \{2, 3, 4\}$ for the settings with negative pairing (the bigger covariance occurs in the smaller sample size for a heteroscedastic unbalanced setting) split by the distributions: standard normal distribution (N), t-distribution with 3 degrees of freedom (t_3), χ^2-distribution with 3 degrees of freedom (χ^2_3), logarithmic standard normal distribution (LN) and double exponential distribution (DExp).

the MCTP with wild bootstrap the MCTP with parametric bootstrap produced more settings with an FWER lower than 4.4. In the Supplemental Material, we present all simulation results split by the distribution and the covariance settings as bee swarm plots (Eklund & Trimble, 2021). From this plots, it can be observed that the FWER-control of all methods is not really stable for the logarithmic normal distribution with the heteroscedastic covariance. This phenomenon is more pronounced if the dimension d is higher. It has to be pointed out that the asymptotic MCTP with multivariate t-distribution performs similarly well as the resampling MCTPs especially for the bigger sample sizes. But a problem of this method is that it is not defined for singular covariance scenarios and can not be applied in these settings. It turns out that the simulations results are not much different for other testing problems or numbers of groups and dimensions. Nevertheless, further plots regarding simulation results can be found in the Supplementary Material.

All in all, the simulation study regarding FWER leads to a recommendation of the two resampling MCTPs. These methods allow potential singularity and work also for small samples and in heterogeneous data, which is an advantage compared to the asymptotic MCTPs. Here, the asymptotic MCTP with t-distribution can be recommended for non-singular data with moderate sample sizes. As the version with parametric bootstrap tends to be a bit conservative in comparison to the approach with wild bootstrap, this method can be recommended if someone is interested in more conservative test decisions or in the situation of moderate negative pairing. Here the phenomena that lead to liberal and conservative behaviour counteract each other.

4.2. Simulation Results Under the Alternative Hypothesis

For further insights we also performed simulations under the alternative. As the power is known to be higher for bigger sample sizes we consider in our power simulations the smaller sample sizes with multiplier $K \in \{1, 2\}$. And we consider only settings with $d = 3$ dimensions, $k = 3$ groups and Dunnett's testing problem. To ensure that the data does not fulfil the null hypothesis we add $\delta \in \{0.5, 1, 1.5, 2, 3\}$ in three ways to $\boldsymbol{\mu}_i$, $i \in \{1, 2, 3\}$:

1. Shift alternative: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$, $\boldsymbol{\mu}_3 = \delta \cdot \mathbf{1}_{n_i d}$,
2. One-point alternative: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$, $\boldsymbol{\mu}_3 = (\delta, 0, \ldots, 0)$,
3. Trend alternative: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$, $\boldsymbol{\mu}_3 = (\delta, \delta/2, \ldots, \delta/d)$.

As the asymptotic MCTP with multivariate normal distribution has a liberal behaviour under the null hypothesis we exclude this method from our power analysis. The results for this method can still be found in the Supplementary Material.

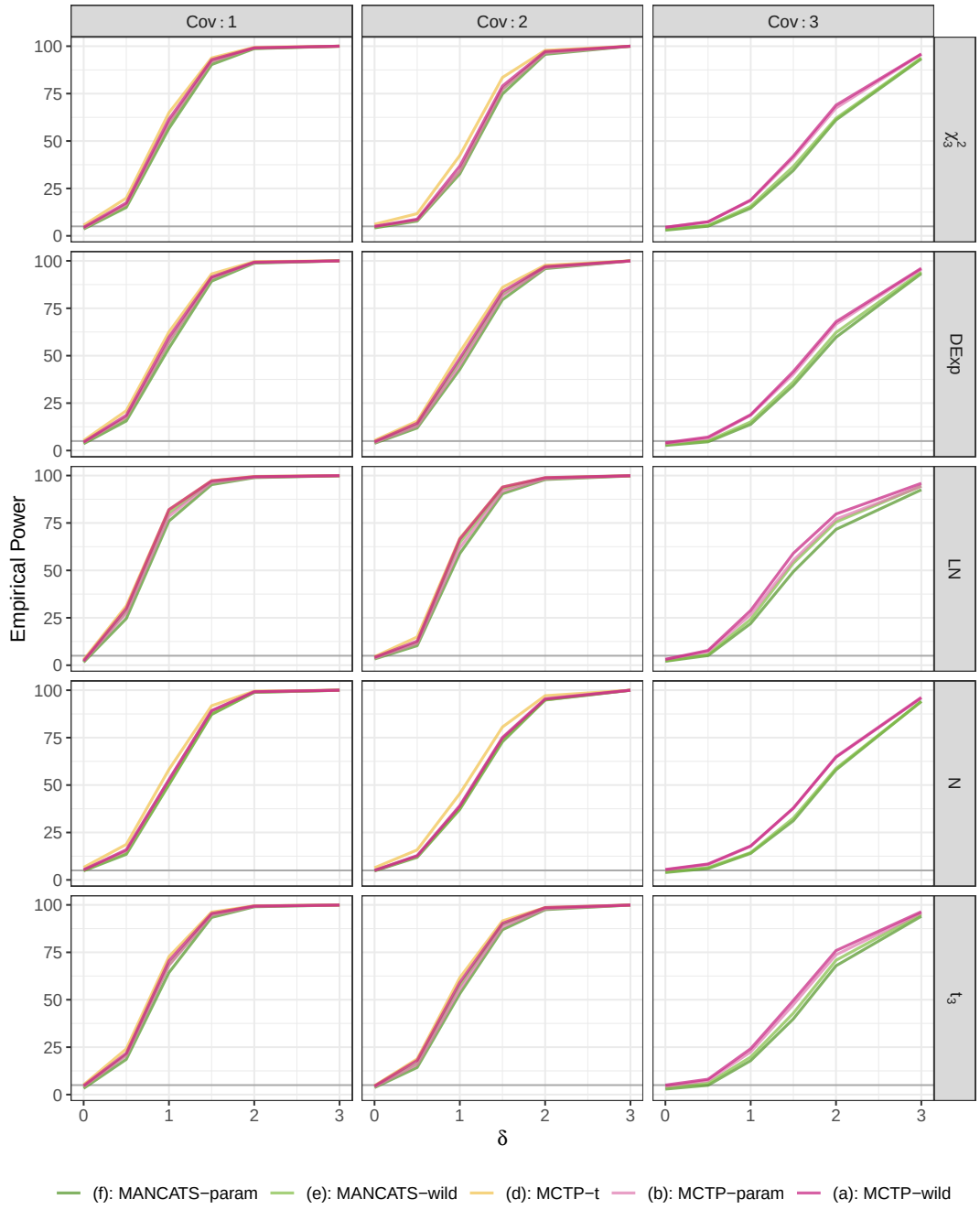

Figure 3: Empirical power in % through a shift alternative for Dunnett's testing problem with $k = 3$ and $d = 3$, for the sample $1 \cdot n_3^{(1)} = (10, 10, 10)$ and the covariance settings: homoscedastic (1), heteroscedastic (2) and singular (3), and distributions: standard normal distribution (N), t-distribution with 3 degrees of freedom (t_3), χ^2-distribution with 3 degrees of freedom (χ^2_3), logarithmic standard normal distribution (LN) and double exponential distribution (DExp).

In Figure 3, the power of the shift alternative of the 12 settings with the sample $1 \cdot n_3^{(1)} = (10, 10, 10)$ is plotted. Here, it can be observed that the five presented methods have a similar power. For the singular covariance setting (3) the power is smaller in comparison to the other covariance settings for all methods and the resampling MCTPs have a better performance than the other methods. For some heteroscedastic settings (χ_3^2-and normal distribution) the MCTP with multivariate t-distribution has a little better power than the other methods. This is not surprising because of the liberal behaviour of the tests in these settings. As expected the power simulations with the multiplier $K = 2$ show a faster increase of the power than for multiplier $K = 1$, which is in line with the proven asymptotic behaviour of the methods. In the Supplementary Material, there are power plots for all settings with sample factor $K = 1$, especially for the different ways to create data under the alternatives. As the one-point alternative is more difficult to detect for a statistical test, the power increases much slower for this alternative, especially for the singular covariance setting (3). The increase of the power under the trend alternative is between the shift and the one-point alternative, which is not surprising. In the settings with negative pairing (sample size $1 \cdot n_3^{(2)} = (20, 10, 10)$, heteroscedastic covariance (2)), it is observable for all considered alternatives that the power increases very slowly, especially for $\delta = 0.5$ the power is very small compared to other settings. This phenomenon occurs in all considered methods, similarly to the bad performance for negative pairing regarding the FWER-control. The power results for $K = 2$ can be found in the Supplementary Material. To conclude, when considering the power, there is no reason not to continue recommending the resampling MCTPs. In general, these methods have a high power, even in singular settings. Especially, the power is higher than the power of the Bonferroni-corrected MANCATs.

5. Data Analysis

In this section, we illustrate the practical application of our MCTPs.

Hypnosis interventional study. Our analysis is motivated from a real interventional study, the *HypnoTreat* study, conducted at Ulm University (Karrasch, Matits, et al., 2023; Karrasch, Mavioğlu, et al., 2023; Karrasch et al., 2022). As the data from the HypnoTreat is not published, we generated a synthetic dataset that preserves key characteristics of the HypnoTreat study, including design, sample sizes, range, skewness, and dependency structures. The resulting dataset is published in Thiel et al. (2025). HypnoTreat examined the effects of a single relaxation hypnosis session on several heart rate variability (HRV) parameters in chronically stressed individuals. HRV refers to the variability in the interval between successive heartbeats (inter-beat intervals), reflecting the heart's capacity to

respond to internal and external stimuli and maintain homeostasis under varying environmental demands (Rajendra Acharya et al., 2006). HRV is influenced by both acute and chronic stress (Kim et al., 2018), and therefore, serves as a non-invasive marker of autonomic nervous system function and the dynamic interaction between its sympathetic and parasympathetic branches.

In HypnoTreat, a total of 45 participants was randomly assigned to either an intervention group (listening to a 20-minute relaxation hypnosis via headphones) or a control group (watching a 20-minute documentary about the universe). HRV was assessed using five standard time- and frequency-domain parameters, assessed at baseline and at post-intervention (Kim et al., 2018; Sammito et al., 2012; Shaffer & Ginsberg, 2017):

SDNN: standard Deviation of Normal-to-Normal Intervals (measured in ms); reflects overall HRV and indicates both sympathetic and parasympathetic activity.

RMSSD: root Mean Square of Successive Differences (measured in ms); primarily reflects short-term HRV and serves as a reliable marker of parasympathetic activity.

HF: high Frequency (measured in ms^2); mainly associated with parasympathetic activity.

LF: low Frequency (measured in ms^2); predominantly reflects sympathetic activity.

VLF: very Low Frequency, (measured in ms^2); less well understood, but linked to other stress-regulating systems such as the endocrine and immune systems.

Besides these HRV parameters, two important confounding variables have been observed at baseline: *suggestibility* and *perceived chronic stress*. Chronic stress was assessed using the German translation of the Perceived Stress Scale-14 (Cohen et al., 1983), which is a sum of 4-point Likert scales and ranges from 0 to 56 (**PSS**). Suggestibility, defined as an individual's responsiveness to hypnotic procedures, was measured using the German version of the Harvard Group Scale of Hypnotic Susceptibility (**HGSHA**) (Bongartz, W., 1982; Shor & Orne, 1963), where a total score was calculated based on the sum of responses to 12 items.

Data modelling. To analyse this longitudinal dataset, we first calculate the change from baseline of all HRV parameters, that is, we subtract the baseline value from the post-intervention values. We now consider the newly gained change from baseline variables as the dependent variables of our analysis. Note that using change from baseline is a simple way of correcting for baseline imbalances between subjects. Figure 4 displays the change from baseline for all HRV parameters in boxplots. Visual investigation suggests possible treatment effects in some variables. In the synthetic dataset, we introduced an artificial shift effect in the hypnosis group in two variables SDNN and VLF (Thiel et al., 2025). Consequently, we are interested in whether the statistical inference tools proposed in this

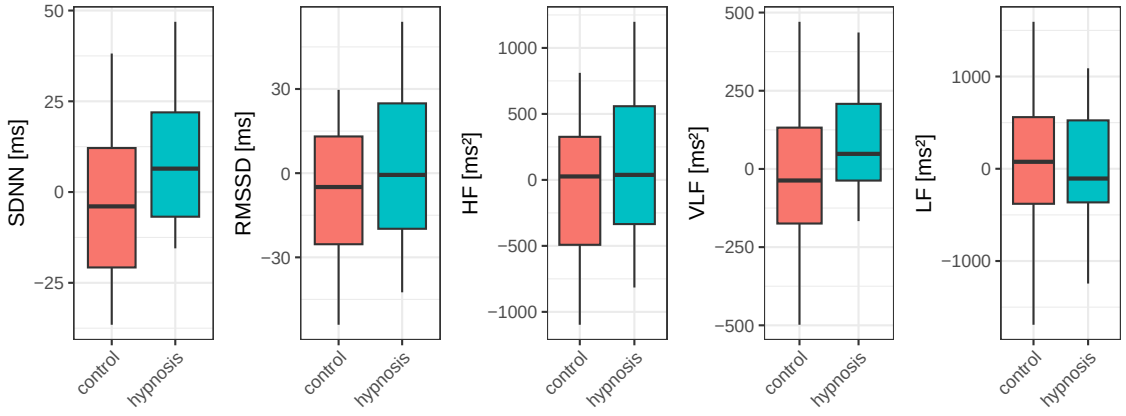

Figure 4: Change from baseline of five heart rate variability (HRV) parameters.

manuscript are able to detect these effects at a significance level of $\alpha = 0.05$, especially when controlling for multiplicity, which is clearly required for separate investigation of the five variables. In fact, we are interested in an overall effect and in specific local effects, which leads to a multiple multivariate two-sample testing problem as explained in Section 3.2. It can be realized by the hypothesis matrix $\mathbf{H}_d := (1, -1) \otimes \mathbf{I}_5$. Then, the global hypothesis is characterized as $\mathcal{H}_0^d : \mathbf{H}_d\boldsymbol{\mu} = \mathbf{0} \Leftrightarrow \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Moreover, the local hypotheses are defined by the rows of $\mathbf{H}_d$ and are consequently given by $\mathcal{H}_{0,\ell}^d : \mu_{1\ell} = \mu_{2,\ell}$, $\ell \in \{1, \ldots, 5\}$.

Before using an inferential tool, we inspect the correlation structure in the dataset in Figure 5. We observe a high correlation between the HRV parameters. For instance, SDNN shows correlations of 0.76 – 0.87 with the other HRV parameters, and the correlation between RMDSS and HF is even 0.95. The high level of linear dependency reflected by these values suggests a multivariate modelling approach for the HRV parameters. The subsequently presented approach offers such an advantage: when applied to our data, it models the joint distribution of test statistics for covariate-adjusted means of the HRV parameters (cf. Theorem 2). Importantly, the resampling MCTPs are also capable of leveraging information from the baseline covariates HGSHA and PSS. Figure 5 shows small to medium correlation coefficients of HGSHA and PSS with several HRV parameters, which suggests that precision of group comparisons may be improved by controlling for these covariates. For the considered methods, this is done in Equation 1, where the mean estimators of HRV parameters are corrected for a linear dependency on the covariates.

Estimators. We consider covariate-adjusted means as estimands. With the notation from Section 3, we have $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1', \hat{\boldsymbol{\mu}}_2')'$, where $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}_{i\ell})'_{\ell=1,\ldots 5}$ for treatment groups $i \in \{1, 2\}$. Here, $\ell \in \{1 \ldots, 5\}$ denotes the individual change from baseline HRV parameters. Using the hypothesis matrix $\mathbf{H}_d$, we obtain the vector of covariate-adjusted mean differences $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ between the two treatment groups. Table 2 displays the absolute values of these

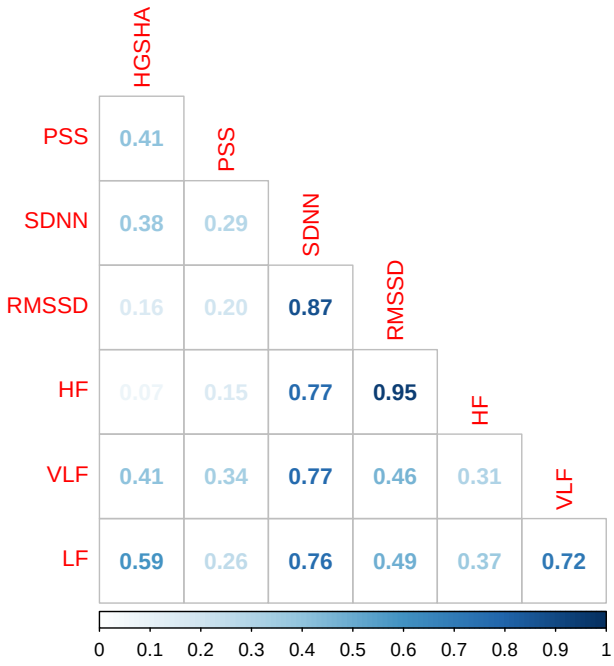

Figure 5: (Pearson) correlation matrix of two baseline covariates (HGSHA, PSS) and change from baseline of five HRV parameters (SDNN, RMSSD, HF, LF, VLF).

	SDNN	**RMSSD**	**HF**	**VLF**	**LF**
unadjusted	12.95	9.80	157.23	115.96	100.38
covariate-adjusted	17.03	11.43	167.89	156.93	103.14

Table 2: Absolute value of unadjusted and covariate-adjusted mean differences of change from baseline HRV parameters between the two treatment groups.

covariate-adjusted mean differences as well as unadjusted mean differences for comparison. We observe that covariate adjustment leads to larger absolute mean differences. Note that for the two parameters SDNN and VLF, for which we know that there exists a treatment effect, the absolute mean difference is even increased by approximately 1/3.

Adjusted significance levels & p-values. We now apply the MCTPs proposed in Section 3 and Section 4: (i) a wild and (ii) a parametric bootstrap MCTP, as well as (iii) a normally-distributed, and (iv) a t-distributed asymptotic MCTP as explained in Section 3.2. Moreover, we compare these MCTPs with (v) a wild and (vi) a parametric bootstrap MANCATS-based test. For each of these tests, we compute local p-values for each contrast $\hat{\mu}_{1\ell} - \hat{\mu}_{2\ell}$, $\ell \in \{1, \ldots, 5\}$. Alongside with these local p-values, we compute adjusted significance levels γ to compare all p-values. For the bootstrap MCTPs, local p-values are defined in Equation (13) and γ is defined in Equation (12), see Section 3.5 for the methodology. For the asymptotic MCTPs, local p-values are defined in the Supplementary Material and γ is obtained by plugging in the numerical value of the equicoordinate quantile into the

	γ	SDNN	RMSSD	HF	VLF	LF
(i) MCTP-wild	0.0145	**0.0050**	0.1270	0.3475	**0.0110**	0.5905
(ii) MCTP-param	0.0150	0.0155	0.2180	0.4390	**0.0075**	0.6275
(iii) MCTP-norm	0.0157	**0.0025**	0.1228	0.3603	**0.0072**	0.5923
(iv) MCTP-t	0.0162	**0.0043**	0.1302	0.3654	**0.0102**	0.5950
(v) MANCATS-wild	0.0100	**0.0040**	0.1180	0.3540	**0.0080**	0.6100
(vi) MANCATS-param	0.0100	**0.0075**	0.1445	0.3640	0.0115	0.5720

Table 3: Local p-values for covariate-adjusted mean differences of five change from baseline HRV parameters and adjusted significance levels γ for various testing methods. Significant test decisions are marked by p-values in bold. The number of bootstrap runs was set to 2000 and the significance level for the global hypothesis (= at least one local hypothesis is rejected) was set to $\alpha = 0.05$.

(univariate) standard normal- or t-distribution function. For the MANCATS-based tests, γ is obtained by standard Bonferroni adjustment.

All approaches aim to control the global FWER α while providing coherent local and global test decisions. As we have investigated in Section 4, the individual methods differ in their capability of actually controlling α. Nevertheless, the representation in terms of local p-values and adjusted significance levels γ guarantees comparability between the methods. Table 3 displays this representation for each method (i) – (vi). Here, we observe that the adjusted significance level γ obtained from Bonferroni is with 0.0100 smaller than all other adjusted levels. The largest levels γ can be observed with the asymptotic MCTPs (iii) and (iv), where the levels of the bootstrap MCTPs (i) and (ii) are only slightly below. This represents the conservative behaviour of the Bonferroni-adjustment and the advantage of the MCTPs: they produce moderate adjusted levels while still controlling the FWER by leveraging dependencies between local test statistics. Note that a more conservative behaviour of the Bonferroni-adjusted methods is expected in case of more than 5 local hypotheses. The order of the levels γ displayed in Table 3 are also in line with our simulation results from Section 4, which attested (iii) and (iv) a somewhat liberal behaviour and (v) and (vi) a rather conservative behaviour. Notably, all methods (i) – (vi) reject at least one local hypothesis and 4 out of 6 methods correctly detect both effects in SDNN and in VLF.

6. Discussion

We extended the framework of multiple contrast testing procedures (MCTPs) to a general semiparametric MANCOVA model, enabling multivariate multiple comparisons based on covariate-adjusted means. The proposed approach allows for heteroscedastic data, singular covariance matrices, and deviations from multivariate normality. This makes it particularly

suited for complex interventional studies with multiple outcomes. In this context, the main advantage of the model is the possibility to analyse several highly correlated outcomes together in one model. Moreover, our model covers further multivariate testing problems like many-to-one and all-pair comparisons.

From a technical point of view, we considered a generalized calculation of critical values to comply with the flexible semiparametric model. This is different to traditional MCTPs that make use of equicorrdinate quantiles, and also necessitates resampling-based inference approaches. For the latter, we followed Friedrich and Pauly (2018) and Zimmermann et al. (2020), and studied parametric and wild bootstrap approaches. Both are theoretically justified within our model framework and yield favourable small-sample performance, as confirmed by extensive simulation studies. In particular, our findings from the simulation study are in line with the theoretical results and demonstrate reliable type-I-error control and competitive power, even on small sample sizes and on singular data. The illustrative data analysis, based on a synthetic dataset mimicking a psychological intervention study, highlights the practical relevance of the presented methods. It exemplifies how multivariate group comparisons can benefit from simultaneous covariate adjustment and multiplicity control.

For the future, several methodological extensions are promising: the recent preprint by Sattler et al. (2024) provides multivariate MCTPs based on quadratic form type statistics instead of linear ones, which allows multiple testing on a vector of group-wise means. This procedure could also be adapted for covariate-adjusted vectors of means. Thinking ahead, covariate-adjusted multiple testing of other estimands beyond the mean, e.g. based on quantiles or nonparametric relative effect, are of high interest: in particular, an extension of the quantile based (M)ANOVA approaches of Baumeister et al. (2024) and Ditzhaus et al. (2021) to allow for covariate-adjustments is tempting. Similarly, a combination of covariate-adjustments (Bathke & Brunner, 2003; Thiel et al., 2024) and MCTPs (Konietschke et al., 2012; Noguchi et al., 2020) for nonparametric relative effects would be interesting. Both, quantile- and rank-based approaches, could be an opportunity to overcome the bad performance on skewed data, which was observed in the simulation study for log-normal data. An R implementation of our proposed MCTPs and those by Zimmermann et al. (2020) is currently in preparation. Until then, the software implementation we used for the data analysis is provided in the Supplementary Material.

CRediT Authorship Contribution Statement

Marléne Baumeister: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation; **Konstantin Emil Thiel:** Data Curation, Formal Analysis, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation; **Lynn Matits:** Data Curation, Investigation, Resources, Writing – Original Draft Preparation; **Markus Pauly:** Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing; **Georg Zimmermann:** Conceptualization, Writing – Review & Editing; **Paavo Sattler:** Formal Analysis, Methodology, Supervision, Writing – Review & Editing.

Acknowledgement

MB and MP has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as project 271512359. LM was supported by a Ph.D. scholarship from the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

We thank Merle Munko for the fruitful exchange about the methodology of multiple contrast tests and our student assistant Mitja Schemmer for the help with the programming of the simulations.

Declarations of interest: none

Data Availability Statement

The data that support the findings of this work e.g. all simulation results, simulation scripts and the scripts for the data analysis are openly available in TUDOdata at https://doi.org/10.17877/TUDODATA-2025-MANOBADL.

References

Bathke, A., & Brunner, E. (2003). A nonparametric alternative to analysis of covariance. In *Recent advances and trends in nonparametric statistics* (pp. 109–120). Elsevier.

Baumeister, M., Ditzhaus, M., & Pauly, M. (2024). Quantile-based MANOVA: A new tool for inferring multivariate data in factorial designs. *Journal of Multivariate Analysis*, *199*, 105246.

Baumeister, M., Munko, M., Gladow, K.-P., Ditzhaus, M., Chakarov, N., & Pauly, M. (2025, April 28). *Early and Late Buzzards: Comparing Different Approaches for Quantile-based Multiple Testing in Heavy-Tailed Wildlife Research Data.* arXiv: 2409.14926 [stat].

Becher, M., Hothorn, L. A., & Konietschke, F. (2025). Analysis of Covariance in General Factorial Designs Through Multiple Contrast Tests Under Variance Heteroscedasticity. *Statistics in Medicine*, *44*(7), e70018.

Boerner, M., & Krüger, O. (2009). Aggression and fitness differences between plumage morphs in the common buzzard (Buteo buteo). *Behavioral Ecology*, *20*(1), 180–185.

Bongartz, W. (1982). Harvard Group Scale of Hypnotic Susceptibility Form A: Deutsche Fassung. *Projekt Experimentelle Hypnose, Universität Konstanz.*

Bretz, F., Genz, A., & A. Hothorn, L. (2001). On the Numerical Availability of Multiple Comparison Procedures. *Biometrical Journal*, *43*(5), 645–656.

Bretz, F., Hothorn, T., & Westfall, P. H. (2011). *Multiple comparisons using R.* CRC Press. OCLC: ocn643322735.

Bühlmann, P. (1998). Sieve bootstrap for smoothing in nonstationary time series. *The Annals of Statistics*, *26*(1).

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, *24*(4), 385.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*(2), 215–233.

Ditzhaus, M., Fried, R., & Pauly, M. (2021). QANOVA: Quantile-based permutation methods for general factorial designs. *TEST*, *30*(4), 960–979.

Djira, G. D., & Hothorn, L. A. (2009). Detecting Relative Changes in Multiple Comparisons with an Overall Mean. *Journal of Quality Technology*, *41*(1), 60–65.

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64.

Dunnett, C. W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, *50*(272), 1096–1121.

Eicker, F. (1963). Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions. *The Annals of Mathematical Statistics*, *34*(2), 447–456.

Eklund, A., & Trimble, J. (2021). *Beeswarm: The bee swarm plot, an alternative to stripchart.* manual.

European Medicines Agency (EMA). (2015). Guideline on adjustment for baseline covariates in clinical trials.

Friedrich, S., Konietschke, F., & Pauly, M. (2017). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, *113*, 38–52.

Friedrich, S., & Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, *165*, 166–179.

Hasler, M. (2014). Multiple Contrast Tests for Multiple Endpoints in the Presence of Heteroscedasticity. *The International Journal of Biostatistics*, *10*(1), 17–28.

Hasler, M., & Hothorn, L. A. (2008). Multiple Contrast Tests in the Presence of Heteroscedasticity. *Biometrical Journal*, *50*(5), 793–800.

Hasler, M., & Hothorn, L. A. (2011). A Dunnett-Type Procedure for Multiple Endpoints. *The International Journal of Biostatistics*, *7*(1).

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies* (2nd ed). Wiley. OCLC: ocn700466125.

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials*, *15*(1), 139.

Karrasch, S., Bongartz, W., Behnke, A., Matits, L., & Kolassa, I.-T. (2022). The Effects of a Single Relaxation Hypnosis Session on Mental Stress in Chronically Stressed Individuals: An Explorative Experiment. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *51*(3–4), 247–262.

Karrasch, S., Matits, L., Bongartz, W., Mavioğlu, R. N., Gumpp, A. M., Mack, M., Tumani, V., Behnke, A., Steinacker, J. M., & Kolassa, I.-T. (2023). An exploratory study of hypnosis-induced blood count changes in chronically stressed individuals. *Biological Psychology*, *178*, 108527.

Karrasch, S., Mavioğlu, R. N., Matits, L., Gumpp, A. M., Mack, M., Behnke, A., Tumani, V., Karabatsiakis, A., Bongartz, W., & Kolassa, I.-T. (2023). Randomized controlled trial investigating potential effects of relaxation on mitochondrial function in immune cells: A pilot experiment. *Biological Psychology*, *183*, 108656.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry Investigation*, *15*(3), 235–245.

Konietschke, F., Bathke, A. C., Harrar, S. W., & Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301.

Konietschke, F., Bösiger, S., Brunner, E., & Hothorn, L. A. (2013). Are Multiple Contrast Tests Superior to the ANOVA? *The International Journal of Biostatistics*, *9*(1), 63–73.

Konietschke, F., Hothorn, L. A., & Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, *6*, 738–759.

Konietschke, F., Schwab, K., & Pauly, M. (2021). Small sample sizes: A big data problem in high-dimensional data analysis. *Statistical Methods in Medical Research*, *30*(3), 687–701.

Munko, M., Ditzhaus, M., Dobler, D., & Genuneit, J. (2024). RMST-based multiple contrast tests in general factorial designs. *Statistics in Medicine*, *43*(10), 1849–1866.

Munko, M., Ditzhaus, M., Pauly, M., Smaga, Ł., & Zhang, J.-T. (2023, June 27). *General multiple tests for functional data*. arXiv: 2306.15259 [stat]. Retrieved January 25, 2024, from http://arxiv.org/abs/2306.15259

Noguchi, K., Abel, R. S., Marmolejo-Ramos, F., & Konietschke, F. (2020). Nonparametric multiple comparisons. *Behavior Research Methods*, *52*(2), 489–502.

R Core Team. (2022). *R: A language and environment for statistical computing*. manual. Vienna, Austria, R Foundation for Statistical Computing.

Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: A review. *Medical and Biological Engineering and Computing*, *44*(12), 1031–1051.

Roy, S. N. (1953). On a Heuristic Method of Test Construction and its use in Multivariate Analysis. *The Annals of Mathematical Statistics*, *24*(2), 220–238.

Rubarth, K., Sattler, P., Zimmermann, H. G., & Konietschke, F. (2022). Estimation and Testing of Wilcoxon–Mann–Whitney Effects in Factorial Clustered Data Designs. *Symmetry*, *14*(2), 244.

Sammito, S., Thielmann, B., Klussmann, A., Deußen, A., Braumann, K.-M., & Böckelmann, I. (2012). Nutzung der Herzschlagfrequenz und der Herzfrequenzvariabilit ¨at in der Arbeitsmedizin und der Arbeitswissenschaft. *Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) e. V., 2014*.

Sattler, P., Pauly, M., & Munko, M. (2024, November 15). *Quadratic Form based Multiple Contrast Tests for Comparison of Group Means*. arXiv: 2411.10121 [stat].

Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, *5*, 258.

Shor, R. E., & Orne, E. C. (1963). Norms on the Harvard Group Scale of Hypnotic Susceptibility, Form A. *International Journal of Clinical and Experimental Hypnosis*, *11*(1), 39–47.

Thiel, K. E., Matits, L., & Baumeister, M. (2025). *HypnoTreatSynth dataset: Effects of relaxation hypnosis on heart rate variability in chronically stressed individuals*. TUDOdata.

Thiel, K. E., Sattler, P., Bathke, A. C., & Zimmermann, G. (2024, December 23). *Resampling NANCOVA: Nonparametric Analysis of Covariance in Small Samples*. arXiv: 2412.17513 [stat].

Tukey, J. W. (1994). The problem of multiple comparisons. Unpublished manuscript reprinted. In H. I. Braun (Ed.), *The Collected Works of John W. Tukey, Braun, H. I. (ed.)* (Vol. Volume 8). Chapman & Hall.

Umlauft, M., Placzek, M., Konietschke, F., & Pauly, M. (2019). Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *Journal of Multivariate Analysis*, *171*, 176–192.

U.S. Department of Health and Human Services Food and Drug Administration. (2023). Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products.

Warne, R. (2014). A Primer on Multivariate Analysis of Variance (MANOVA) for Behavioral Scientists. *Practical Assessment, Research & Evaluation*, *19*, 17.

Welz, T., Viechtbauer, W., & Pauly, M. (2023). Cluster-robust estimators for multivariate mixed-effects meta-regression. *Computational Statistics & Data Analysis*, *179*, 107631.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Woodward, M. (2013, December). *Epidemiology: Study Design and Data Analysis, Third Edition* (3rd). Chapman and Hall/CRC.

Zimmermann, G., Pauly, M., & Bathke, A. C. (2019). Small-sample performance and underlying assumptions of a bootstrap-based inference method for a general analysis of covariance model with possibly heteroskedastic and nonnormal errors. *Statistical Methods in Medical Research*, *28*(12), 3808–3821.

Zimmermann, G., Pauly, M., & Bathke, A. C. (2020). Multivariate analysis of covariance with potentially singular covariance matrices and non-normal responses. *Journal of Multivariate Analysis*, *177*, 104594.

A. Proofs

A.1. Proof of Theorem 2

To derive the asymptotics of $\mathbf{A}_n(\mathbf{H})$ and $A_n(\mathbf{h}_s)$, $s \in \{1, \ldots, r\}$, we can deduce the following statements from Proposition 1. To extract the vectors $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ from $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ we use the vector $\tilde{\mathbf{h}}_s' = (\mathbf{h}_s', \mathbf{0}_c')' \in \mathbb{R}^{dk+c}$ and the zero-inflated covariance estimator

$$\hat{\tilde{\mathbf{D}}} = \hat{\mathbf{D}} \oplus \mathbf{0}_{c,c}$$

for all $s \in \{1, \ldots, r\}$. The matrix $\tilde{\mathbf{D}}$ is obtained from $\mathbf{D}$ in the same way. The zero-inflated contrast matrix $\tilde{\mathbf{H}} = (\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_r)' \in \mathbb{R}^{r \times (dk+c)}$ is consequently built with the zero-inflated contrast vectors. Then, it holds $\mathcal{H}_{0,s} : \tilde{\mathbf{h}}_s\boldsymbol{\beta} = \mathbf{h}_s\boldsymbol{\mu} = \mathbf{0}$, $\mathcal{H}_0 : \tilde{\mathbf{H}}\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ and $\tilde{\mathbf{h}}_s'\tilde{\mathbf{D}}\tilde{\mathbf{h}}_s = \mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s$. Under $\mathcal{H}_0$, by Slutzky's and the Continuous Mapping Theorem we can describe the distribution of the vector of test statistics:

$$\begin{aligned} A_n(\mathbf{H}) &= (\mathbf{H}\hat{\mathbf{D}}\mathbf{H}')_0^{-1/2}\sqrt{n}\mathbf{H}\hat{\boldsymbol{\mu}} = (\tilde{\mathbf{H}}\hat{\tilde{\mathbf{D}}}\tilde{\mathbf{H}}')_0^{-1/2}\sqrt{n}\tilde{\mathbf{H}}\hat{\boldsymbol{\beta}} \\ &= (\tilde{\mathbf{H}}\hat{\tilde{\mathbf{D}}}\tilde{\mathbf{H}}')_0^{-1/2}\tilde{\mathbf{H}}\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} (\tilde{\mathbf{H}}\tilde{\mathbf{D}}\tilde{\mathbf{H}}')_0^{-1/2}\tilde{\mathbf{H}}\mathbf{Z} =: \mathbf{B}, \end{aligned}$$

From the distribution of $\mathbf{Z}$ it follows that $\mathbf{B} = (B_1, \ldots, B_r)$ has a multivariate normal distribution with expectation $\mathrm{E}(\mathbf{A}_n(\mathbf{H})) = \mathbf{0}$ and covariance matrix

$$\mathbf{R} := \mathrm{Cov}\left(\mathbf{A}_n(\mathbf{H})\right) = \left(\mathbf{H}\mathbf{D}\mathbf{H}'\right)_0^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Lambda}_{11}\mathbf{H}'\left(\mathbf{H}\mathbf{D}\mathbf{H}'\right)_0^{-\frac{1}{2}}.$$

This is the first assertion. Furthermore, $A_n(\mathbf{h}_s)$ is for every $s \in \{1, \ldots, r\}$ already under $\mathcal{H}_{0,s}$ asymptotically distributed like B_s. In general, we can argue that $|A_n(\mathbf{h}_s)|/\sqrt{n}$ converges in probability:

$$\frac{1}{\sqrt{n}}\left|A_n(\mathbf{h}_s)\right| = \frac{|\mathbf{h}_s'\hat{\boldsymbol{\mu}}|}{\sqrt{\mathbf{h}_s'\hat{\mathbf{D}}\mathbf{h}_s}} \xrightarrow{p} \frac{|\mathbf{h}_s'\boldsymbol{\mu}|}{\left|\sqrt{\mathbf{h}_s'\mathbf{D}\mathbf{h}_s}\right|}.$$

Under $\mathcal{H}_{1,s}$ this limiting value is greater than zero. From this we can conclude that under $\mathcal{H}_{1,s}$ the test statistics $A_n(\mathbf{h}_s)$ converges in probability to ∞ for all $s \in \{1, \ldots, r\}$, which is the second assertion.

A.2. Proof of Theorem 3

In Zimmermann et al. (2020) statement (C1) in the proof of Theorem 3 is a Central Limit Theorem for the wild bootstrap:

$$\sqrt{n}\hat{\boldsymbol{\beta}}^* \xrightarrow{d} \mathbf{N} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}\right) \quad \text{given the data.}$$

From this and from the consistency of $\hat{\mathbf{D}}^*$ (Zimmermann et al., 2020, Proof of Theorem 3) it follows similarly to the original test statistic:

$$A_n^*(\mathbf{h}_s) = \sqrt{n}\frac{\mathbf{h}_s'\hat{\boldsymbol{\mu}}^*}{\sqrt{\mathbf{h}_s'\hat{\mathbf{D}}^*\mathbf{h}_s}} = \tilde{\mathbf{h}}_s'\frac{\sqrt{n}\hat{\boldsymbol{\beta}}^*}{\sqrt{\tilde{\mathbf{h}}_s'\hat{\tilde{\mathbf{D}}}^*\tilde{\mathbf{h}}_s}} \xrightarrow{d} \tilde{\mathbf{h}}_s'\frac{\mathbf{N}}{\sqrt{\tilde{\mathbf{h}}_s'\tilde{\mathbf{D}}\tilde{\mathbf{h}}_s}} = \mathbf{B}.$$

This is close to the assertion.

A.3. Proof of Theorem 4

In the Proof of Theorem 4 in Zimmermann et al. (2020) there is the following statement given the data:

$$\sqrt{n}\hat{\boldsymbol{\beta}}^\star \sim \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Lambda}}_n\right),$$

where $\hat{\boldsymbol{\Lambda}}_n = \mathrm{Cov}(\sqrt{n}\hat{\boldsymbol{\beta}}^\star|\mathbf{Y})$ is the conditional covariance estimator of $\hat{\boldsymbol{\beta}}^\star$, which converges in probability to $\boldsymbol{\Lambda}$. From this it follows the assertion.

A.4. Proof of Theorem 5

It is stated in the proof of Theorem 6 in Munko et al. (2024), that Lemma S5 and S6 in the Supplement of Munko et al. (2024) imply $q^\circ_{s,1-\gamma_n(\alpha)} \xrightarrow{P} q_{s,\mathrm{FWER}_n^{-1}(\alpha)}$, $s \in \{1,\dots,r\}$. The two Lemmas are applicable because of the identical framework in Munko et al. (2024). Let $T \subset \{1,\dots,r\}$ be the set of true hypotheses. By Slutzky's Theorem and Equation (7) it follows

$$\left(A_n(\mathbf{h}_s), q^\circ_{s,1-\gamma_n(\alpha)}\right)_{s\in T} \xrightarrow{d} \left(B, q_{s,\mathrm{FWER}_n^{-1}(\alpha)}\right)_{s\in T}.$$

Analogously to the proof of Theorem 6 in Munko et al. (2024) we show that

$$\begin{aligned}
\Pr\left(\max_{s\in T} A_n(\mathbf{h}_s) > q^\circ_{s,1-\gamma_n(\alpha)}\right) &= 1 - \Pr\left(\forall s \in T : A_n(\mathbf{h}_s) \le q^\circ_{s,1-\gamma_n(\alpha)}\right) \\
&\longrightarrow 1 - \Pr\left(\forall s \in T : B \le q_{s,\mathrm{FWER}_n^{-1}(\alpha)}\right) \\
&\le 1 - \Pr\left(\forall s \in \{1,\dots,r\} : B \le q_{s,\mathrm{FWER}_n^{-1}(\alpha)}\right) \qquad (14) \\
&= \mathrm{FWER}_n\left(\mathrm{FWER}_n^{-1}(\alpha)\right) = \alpha.
\end{aligned}$$

Note that there is an equality in (14) if $T = \{1,\dots,r\}$.

A.5. Proof of Proposition 6

For (i), let $s \in \{1, \ldots, r\}$ be fixed. Firstly, assume $p_{n,s} \leq \gamma_n(\alpha)$ and for a proof by contradiction $\varphi^{\circ}_{n,s} = 0$, which is equivalent to $|A_n(\mathbf{h}_s)| \leq q^{\circ}_{s,1-\gamma_n(\alpha)}$. Then,

$$p_{n,s} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{|A_n^{\circ,b}(\mathbf{h}_s)| \geq |A_n(\mathbf{h}_s)|\right\} \geq \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{|A_n^{\circ,b}(\mathbf{h}_s)| \geq q^{\circ}_{s,1-\gamma_n(\alpha)}\right\} = \gamma_n(\alpha) + \frac{1}{B},$$

where the last equality holds by the definition of the quantile $q^{\circ}_{s,1-\gamma_n(\alpha)}$. From this, it follows the contradiction $\gamma_n(\alpha) \geq \gamma_n(\alpha) + 1/B$. Secondly, we assume $\varphi^{\circ}_{n,s} = 1$. Then, $|A_n(\mathbf{h}_s)| > q^{\circ}_{s,1-\gamma_n(\alpha)}$. From this, with the definition of the quantile $q^{\circ}_{s,1-\gamma_n(\alpha)}$, it can be concluded

$$p_{n,s} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{|A_n^{\circ,b}(\mathbf{h}_s)| \geq |A_n(\mathbf{h}_s)|\right\} \leq \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{|A_n^{\circ,b}(\mathbf{h}_s)| \geq q^{\circ}_{s,1-\gamma_n(\alpha)}\right\} \leq \gamma_n(\alpha).$$

To prove (ii), we argue that $p_n = \min\{p_{n,1}, \ldots, p_{n,r}\} \leq \gamma_n(\alpha)$ if and only if there exist $s \in \{1, \ldots, r\}$ such that $p_{n,s} \leq \gamma_n(\alpha)$. Due to (i), this is equivalent to the situation that there exist $s \in \{1, \ldots, r\}$ with $\varphi^{\circ}_{n,s} = 1$, which is the same as $\varphi^{\circ}_n = 1$. Compare also Proposition 1 in Munko et al. (2024).

Bibliography of the Dissertation

Anderson, M. J. (2001). A New Method for Non-Parametric Multivariate Analysis of Variance. *Austral Ecology*, *26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Babu, G., & Rao, C. (1988). Joint Asymptotic Distribution of Marginal Quantiles and Quantile Functions in Samples from a Multivariate Population. *Journal of Multivariate Analysis*, *27*(1), 15–23. https://doi.org/10.1016/0047-259X(88)90112-1

Bartlett, M. S. (1939). A Note on Tests of Significance in Multivariate Analysis. *Mathematical Proceedings of the Cambridge Philosophical Society*, *35*(2), 180–185. https://doi.org/10.1017/S0305004100020880

Bathke, A., & Brunner, E. (2003). A Nonparametric Alternative to Analysis of Covariance. In *Recent advances and trends in nonparametric statistics* (pp. 109–120). Elsevier.

Baumeister, M., Ditzhaus, M., & Pauly, M. (2024). Quantile-Based MANOVA: A New Tool for Inferring Multivariate Data in Factorial Designs. *Journal of Multivariate Analysis*, *199*, 105246. https://doi.org/10.1016/j.jmva.2023.105246

Baumeister, M., Munko, M., Gladow, K.-P., Ditzhaus, M., Chakarov, N., & Pauly, M. (2025a). Early and Late Buzzards: Comparing Different Approaches for Quantile-Based Multiple Testing in Heavy-Tailed Wildlife Research Data. *Biometrical Journal*, *67*(4), e70065. https://doi.org/10.1002/bimj.70065

Baumeister, M., Munko, M., Gladow, K.-P., Ditzhaus, M., Chakarov, N., & Pauly, M. (2025b). Supplementary Material of "Early and Late Buzzards: Comparing Different Approaches for Quantile-Based Multiple Testing in Heavy-Tailed Wildlife Research Data". https://doi.org/10.17877/TUDODATA-2025-M6TDKFDE

Baumeister, M., Thiel, K. E., Matits, L., Pauly, M., Zimmermann, G., & Sattler, P. (2025). Supplementary Material of "Multivariate and Multiple Contrast Testing in General Covariate-Adjusted Factorial Designs". https://doi.org/10.17877/TUDODATA-2025-MANOBADL

Baumeister, M., Thiel, K. E., Matits, L., Zimmermann, G., Pauly, M., & Sattler, P. (2025, June). Multivariate and Multiple Contrast Testing in General Covariate-Adjusted Factorial Designs. https://doi.org/10.48550/arXiv.2506.15292v1

Becher, M., Hothorn, L. A., & Konietschke, F. (2025). Analysis of Covariance in General Factorial Designs Through Multiple Contrast Tests Under Variance Heteroscedasticity. *Statistics in Medicine*, *44*(7), e70018. https://doi.org/10.1002/sim.70018

Bonett, D. G., & Price, R. M. (2002). Statistical Inference for a Linear Function of Medians: Confidence Intervals, Hypothesis Testing, and Sample Size Requirements. *Psychological Methods*, *7*(3), 370–383. https://doi.org/10.1037/1082-989X.7.3.370

Bretz, F., Genz, A., & A. Hothorn, L. (2001). On the Numerical Availability of Multiple Comparison Procedures. *Biometrical Journal*, *43*(5), 645–656. https://doi.org/10.1002/1521-4036(200109)43:5<645::AID-BIMJ645>3.0.CO;2-F

Bretz, F., Hothorn, T., & Westfall, P. H. (2011). *Multiple Comparisons Using R*. CRC Press.

Brunner, E., Dette, H., & Munk, A. (1997). Box-Type Approximations in Nonparametric Factorial Designs. *Journal of the American Statistical Association*, *92*(440), 1494–1502. https://doi.org/10.1080/01621459.1997.10473671

Brunner, E., & Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal*, *42*(1), 17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U

Brunner, E., Munzel, U., & Puri, M. L. (2002). The Multivariate Nonparametric Behrens-Fisher Problem. *Journal of Statistical Planning and Inference*, *108*(1), 37–53. https://doi.org/10.1016/S0378-3758(02)00269-0

Bühlmann, P. (1998). Sieve Bootstrap for Smoothing in Nonstationary Time Series. *The Annals of Statistics*, *26*(1). https://doi.org/10.1214/aos/1030563978

Chung, E., & Romano, J. P. (2013). Exact and Asymptotically Robust Permutation Tests. *The Annals of Statistics*, *41*(2), 484–507. https://doi.org/10.1214/13-AOS1090

Deuchler, G. (1914). Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. *Z pädagog Psychol*, *15*, 114–31.

Ditzhaus, M., Fried, R., & Pauly, M. (2021). QANOVA: Quantile-Based Permutation Methods for General Factorial Designs. *TEST*, *30*(4), 960–979. https://doi.org/10.1007/s11749-021-00758-y

Djira, G. D., & Hothorn, L. A. (2009). Detecting Relative Changes in Multiple Comparisons with an Overall Mean. *Journal of Quality Technology*, *41*(1), 60–65. https://doi.org/10.1080/00224065.2009.11917760

Dobler, D., Friedrich, S., & Pauly, M. (2020). Nonparametric MANOVA in Meaningful Effects. *Annals of the Institute of Statistical Mathematics*, *72*(4), 997–1022. https://doi.org/10.1007/s10463-019-00717-3

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Dunnett, C. W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, *50*(272), 1096–1121. https://doi.org/10.1080/01621459.1955.10501294

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression*. Springer.

Fisher, R. A. (1919). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. *Annals of Eugenics*, *6*(4), 391–398. https://doi.org/10.1111/j.1469-1809.1935.tb02120.x

Friedrich, S., & Pauly, M. (2018). MATS: Inference for Potentially Singular and Heteroscedastic MANOVA. *Journal of Multivariate Analysis*, *165*, 166–179. https://doi.org/10.1016/j.jmva.2017.12.008

Gabriel, K. R. (1969). Simultaneous Test Procedures–Some Theory of Multiple Comparisons. *The Annals of Mathematical Statistics*, *40*(1), 224–250. https://doi.org/10.1214/aoms/1177697819

Hasler, M. (2014). Multiple Contrast Tests for Multiple Endpoints in the Presence of Heteroscedasticity. *The International Journal of Biostatistics*, *10*(1), 17–28. https://doi.org/10.1515/ijb-2012-0015

Hasler, M., & Hothorn, L. A. (2008). Multiple Contrast Tests in the Presence of Heteroscedasticity. *Biometrical Journal*, *50*(5), 793–800. https://doi.org/10.1002/bimj.200710466

Hasler, M., & Hothorn, L. A. (2011). A Dunnett-Type Procedure for Multiple Endpoints. *The International Journal of Biostatistics*, *7*(1). https://doi.org/10.2202/1557-4679.1258

Hauck, W., & Anderson, S. (1984). A New Statistical Procedure for Testing Equivalence in 2-Group Comparative Bioavailability Trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*(1), 83–91. https://doi.org/10.1007/BF01063612

Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., White, I. R., & the Simulation Panel of the STRATOS initiative. (2024). Phases of Methodological Research in Biostatistics—Building the Evidence Base for New Methods. *Biometrical Journal*, *66*(1), 2200222. https://doi.org/10.1002/bimj.202200222

Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, *2*(3), 360–378.

Hotelling, H. (1951). A Generalized T-Test and Measure of Multivariate Dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.*, 19.

Huitema, B. E. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies* (2nd ed). Wiley.

Janssen, A., & Pauls, T. (2003). How Do Bootstrap and Permutation Tests Work? *The Annals of Statistics*, *31*(3), 768–806. https://doi.org/10.1214/aos/1056562462

Karrasch, S., Bongartz, W., Behnke, A., Matits, L., & Kolassa, I.-T. (2022). The Effects of a Single Relaxation Hypnosis Session on Mental Stress in Chronically Stressed Individuals: An Explorative Experiment. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *51*(3-4), 247–262. https://doi.org/10.1026/1616-3443/a000679

Karrasch, S., Matits, L., Bongartz, W., Mavioğlu, R. N., Gumpp, A. M., Mack, M., Tumani, V., Behnke, A., Steinacker, J. M., & Kolassa, I.-T. (2023). An Exploratory Study of Hypnosis-Induced Blood Count Changes in Chronically Stressed Individuals. *Biological Psychology*, *178*, 108527. https://doi.org/10.1016/j.biopsycho.2023.108527

Karrasch, S., Mavioğlu, R. N., Matits, L., Gumpp, A. M., Mack, M., Behnke, A., Tumani, V., Karabatsiakis, A., Bongartz, W., & Kolassa, I.-T. (2023). Randomized Controlled Trial Investigating Potential Effects of Relaxation on Mitochondrial Function in Immune Cells: A Pilot Experiment. *Biological Psychology*, *183*, 108656. https://doi.org/10.1016/j.biopsycho.2023.108656

Konietschke, F., Bathke, A. C., Harrar, S. W., & Pauly, M. (2015). Parametric and Nonparametric Bootstrap Methods for General MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301. https://doi.org/10.1016/j.jmva.2015.05.001

Konietschke, F., Bösiger, S., Brunner, E., & Hothorn, L. A. (2013). Are Multiple Contrast Tests Superior to the ANOVA? *The International Journal of Biostatistics*, *9*(1), 63–73. https://doi.org/10.1515/ijb-2012-0020

Konietschke, F., Hothorn, L. A., & Brunner, E. (2012). Rank-Based Multiple Test Procedures and Simultaneous Confidence Intervals. *Electronic Journal of Statistics*, *6*(none), 738–759. https://doi.org/10.1214/12-EJS691

Konietschke, F., & Pauly, M. (2012). A Studentized Permutation Test for the Nonparametric Behrens-Fisher Problem in Paired Data. *Electronic Journal of Statistics*, *6*(none), 1358–1372. https://doi.org/10.1214/12-EJS714

Konietschke, F., & Pauly, M. (2014). Bootstrapping and Permuting Paired t-Test Type Statistics. *Statistics and Computing*, *24*(3), 283–296. https://doi.org/10.1007/s11222-012-9370-4

Konietschke, F., Schwab, K., & Pauly, M. (2021). Small Sample Sizes: A Big Data Problem in High-Dimensional Data Analysis. *Statistical Methods in Medical Research*, *30*(3), 687–701. https://doi.org/10.1177/0962280220970228

Larson, J. (1996a). *Seasons of Love*. US, Dream Works Records.

Larson, J. (1996b). *Seasons of Love B*. US, Dream Works Records.

Lawley, D. N. (1938). A Generalization of Fisher's z-Test. *Biometrika*, *30*(1/2), 180–187. https://doi.org/10.2307/2332232

Lehmann, E., & Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer International Publishing. https://doi.org/10.1007/978-3-030-70578-7

Maronna, R. A., Martin, R. D., Yohai, V. J., & Martin, D. (2006). *Robust Statistics: Theory and Methods* (Reprinted with corr). Wiley.

Mathai, A. M., & Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications*. M. Dekker.

Munko, M., Ditzhaus, M., Dobler, D., & Genuneit, J. (2024). RMST-Based Multiple Contrast Tests in General Factorial Designs. *Statistics in Medicine*, *43*(10), 1849–1866. https://doi.org/10.1002/sim.10017

Munko, M., Ditzhaus, M., Pauly, M., Smaga, Ł., & Zhang, J.-T. (2023, June). General Multiple Tests for Functional Data. https://doi.org/10.48550/arXiv.2306.15259v1

Munzel, U. (1999). Nonparametric Methods for Paired Samples. *Statistica Neerlandica*, *53*(3), 277–286. https://doi.org/10.1111/1467-9574.00112

Nadaraya, É. A. (1965). On Non-Parametric Estimates of Density Functions and Regression Curves. *Theory of Probability & Its Applications*, *10*(1), 186–190. https://doi.org/10.1137/1110024

Neubert, K., & Brunner, E. (2007). A Studentized Permutation Test for the Non-Parametric Behrens-Fisher Problem. *Computational Statistics & Data Analysis*, *51*(10), 5192–5204. https://doi.org/10.1016/j.csda.2006.05.024

Noguchi, K., Abel, R. S., Marmolejo-Ramos, F., & Konietschke, F. (2020). Nonparametric Multiple Comparisons. *Behavior Research Methods*, *52*(2), 489–502. https://doi.org/10.3758/s13428-019-01247-9

Noguchi, K., Konietschke, F., Marmolejo-Ramos, F., & Pauly, M. (2021). Permutation Tests Are Robust and Powerful at 0.5% and 5% Significance Levels. *Behavior Research Methods*, *53*(6), 2712–2724. https://doi.org/10.3758/s13428-021-01595-5

Pauly, M., Brunner, E., & Konietschke, F. (2015). Asymptotic Permutation Tests in General Factorial Designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(2), 461–473. https://doi.org/10.1111/rssb.12073

Pigeot, I. (2000). Basic Concepts of Multiple Tests—A Survey. *Statistical Papers*, *41*(1), 3–36. https://doi.org/10.1007/BF02925674

Pillai, K. C. S. (1955). Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, *26*(1), 117–121.

Price, R. M., & Bonett, D. G. (2001). Estimating the Variance of the Sample Median. *Journal of Statistical Computation and Simulation*, *68*(3), 295–305. https://doi.org/10.1080/00949650108812071

Pudelko, H. (1992). *Wenn Ich Will*. Germany, WEA.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898719109

Roy, S. N. (1953). On a Heuristic Method of Test Construction and Its Use in Multivariate Analysis. *The Annals of Mathematical Statistics*, *24*(2), 220–238. https://doi.org/10.1214/aoms/1177729029

Roy, S. N. (1945). The Individual Sampling Distribution of the Maximum, the Minimum and Any Intermediate of the p-Statistics on the Null-Hypothesis. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *7*(2), 133–158.

Rubarth, K., Sattler, P., Zimmermann, H. G., & Konietschke, F. (2022). Estimation and Testing of Wilcoxon–Mann–Whitney Effects in Factorial Clustered Data Designs. *Symmetry*, *14*(2), 244. https://doi.org/10.3390/sym14020244

Sattler, P., & Pauly, M. (2018). Inference for High-Dimensional Split-Plot-Designs: A Unified Approach for Small to Large Numbers of Factor Levels. *Electronic Journal of Statistics*, *12*(2), 2743–2805. https://doi.org/10.1214/18-EJS1465

Sattler, P., Pauly, M., & Munko, M. (2025, June). Quadratic Form based Multiple Contrast Tests for Comparison of Group Means. https://doi.org/10.48550/arXiv.2411.10121v2

Sattler, P., & Rosenbaum, M. (2025). Choice of the Hypothesis Matrix for Using the ANOVA-Type-Statistic. *Statistics & Probability Letters*, *219*, 110356. https://doi.org/10.1016/j.spl.2025.110356

Sattler, P., & Zimmermann, G. (2024). Choice of the Hypothesis Matrix for Using the Wald-Type-Statistic. *Statistics & Probability Letters*, *208*, 110038. https://doi.org/10.1016/j.spl.2024.110038

Segbehoe, L. S., Schaarschmidt, F., & Djira, G. D. (2022). Simultaneous Confidence Intervals for Contrasts of Quantiles. *Biometrical Journal*, *64*(1), 7–19. https://doi.org/10.1002/bimj.202000077

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics* (1st ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316481

Stapelton, J. H. (1995). *Linear Statistical Models* (1st ed.). John Wiley & Sons, Ltd.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1–25. https://doi.org/10.2307/2331554

Thiel, K. E., Matits, L., & Baumeister, M. (2025). HypnoTreatSynth Dataset: Effects of Relaxation Hypnosis on Heart Rate Variability in Chronically Stressed Individuals. https://doi.org/DOI:10.17877/TUDODATA-2025-M9X7Y26X

Thiel, K. E., Sattler, P., Bathke, A. C., & Zimmermann, G. (2024, December). Resampling NANCOVA: Nonparametric Analysis of Covariance in Small Samples. https://doi.org/10.48550/arXiv.2412.17513v1

Tukey, J. W. (1994). The Problem of Multiple Comparisons. Unpublished Manuscript Reprinted. In H. I. Braun (Ed.), *The Collected Works of John W. Tukey, Braun, H. I. (ed.)* (Vol. Volume 8). Chapman & Hall.

Umlauft, M., Placzek, M., Konietschke, F., & Pauly, M. (2019). Wild Bootstrapping Rank-Based Procedures: Multiple Testing in Nonparametric Factorial Repeated Measures Designs. *Journal of Multivariate Analysis*, *171*, 176–192. https://doi.org/10.1016/j.jmva.2018.12.005

van der Vaart, A. W., & Wellner, J. A. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. https://doi.org/10.1090/S0002-9947-1943-0012401-3

Welch, B. L. (1938). The Significance of the Difference Between Two Means When the Population Variances Are Unequal. *Biometrika*, *29*(3/4), 350–362. https://doi.org/10.2307/2332010

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Wilks, S. S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, *24*(3/4), 471–494. https://doi.org/10.2307/2331979

Zimmermann, G., Pauly, M., & Bathke, A. C. (2019). Small-Sample Performance and Underlying Assumptions of a Bootstrap-Based Inference Method for a General Analysis of Covariance Model with Possibly Heteroskedastic and Nonnormal Errors. *Statistical Methods in Medical Research*, *28*(12), 3808–3821. https://doi.org/10.1177/0962280218817796

Zimmermann, G., Pauly, M., & Bathke, A. C. (2020). Multivariate Analysis of Covariance with Potentially Singular Covariance Matrices and Non-Normal Responses. *Journal of Multivariate Analysis*, *177*, 104594. https://doi.org/10.1016/j.jmva.2020.104594