Annals of the Institute of Statistical Mathematics (2020) 72:997–1022
https://doi.org/10.1007/s10463-019-00717-3

Nonparametric MANOVA in meaningful effects

Dennis Dobler[1] · **Sarah Friedrich**[2] · **Markus Pauly**[3]

Received: 20 April 2018 / Revised: 13 March 2019 / Published online: 6 April 2019

Abstract

Multivariate analysis of variance (MANOVA) is a powerful and versatile method to infer and quantify main and interaction effects in metric multivariate multi-factor data. It is, however, neither robust against change in units nor meaningful for ordinal data. Thus, we propose a novel nonparametric MANOVA. Contrary to existing rank-based procedures, we infer hypotheses formulated in terms of meaningful Mann–Whitney-type effects in lieu of distribution functions. The tests are based on a quadratic form in multivariate rank effect estimators, and critical values are obtained by bootstrap techniques. The newly developed procedures provide asymptotically exact and consistent inference for general models such as the nonparametric Behrens–Fisher problem and multivariate one-, two-, and higher-way crossed layouts. Computer simulations in small samples confirm the reliability of the developed method for ordinal and metric data with covariance heterogeneity. Finally, an analysis of a real data example illustrates the applicability and correct interpretation of the results.

Authors are in alphabetical order.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-019-00717-3) contains supplementary material, which is available to authorized users.

✉ Dennis Dobler
d.dobler@vu.nl

Sarah Friedrich
safr@sund.ku.dk

Markus Pauly
markus.pauly@tu-dortmund.de

[1] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

[2] Section of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, 1014 Copenhagen, Denmark

[3] Institute for Mathematical Statistics and Industrial Applications, Faculty of Statistics, Technical University of Dortmund, 44221 Dortmund, Germany

Keywords Covariance heteroscedasticity · Multivariate data · Multivariate ordinal data · Multiple samples · Rank-based methods · Wild bootstrap

1 Motivation and introduction

In many experiments, e.g., in the life sciences or in econometrics, observations are obtained in elaborate factorial designs with multiple endpoints. Such data are usually analyzed using MANOVA methods such as Wilk's Λ. These procedures, however, rely on the assumptions of multivariate normality and covariance homogeneity and usually break down if these prerequisites are not fulfilled. In particular, if the observations are not even metric, such applications are no longer possible since means no longer provide adequate effect measures. To this end, several rank-based methods have been proposed for nonparametric MANOVA and repeated measures designs which are usually based on Mann–Whitney-type effects: in the context of a nonparametric *univariate* two-sample problem with independent and continuous observations $Y_{ik} \sim F_i,\ i = 1, 2,\ k = 1, \ldots, n_i$, Mann and Whitney (1947) introduced the effect $w = P(Y_{11} \leq Y_{21}) = \int F_1 dF_2$ also known as ordinal effect size measure (Acion et al. 2006). An estimator of w is easily obtained by replacing the distribution functions with their empirical counterparts. While this effect has several desirable properties and is widely accepted in practice (Brumback et al. 2006; Kieser et al. 2013), generalizations to more than one dimension or higher-way factorial designs are not straightforward.

Concerning the latter, there basically exist two possibilities in the literature to cope with $a \geq 3$ sample groups with independent univariate observations $Y_{ik} \sim F_i,\ i = 1, \ldots, a,\ k = 1, \ldots, n_i$: *First*, considering only the pairwise effects $w_{i\ell} = P(Y_{i1} \leq Y_{\ell 1}),\ 1 \leq i \neq \ell \leq a$ (as proposed by Rust and Filgner 1984) can lead to paradox results in the sense of Efron's Dice. That is, due to the possible situation $w_{12} > \frac{1}{2}$ and $w_{23} > \frac{1}{2}$ the third group appears to be stochastically greater than the first group even though $w_{31} > \frac{1}{2}$ is possible at the same time. See also Thas et al. (2012) and the contributed discussions by M. P. Fay and W. Bergsma and colleagues for pros and cons of the possibly induced intransitivity by certain probabilistic index models. We refer to Brown and Hettmansperger (2002), Thangavelu and Brunner (2007) or Brunner et al. (2017) and the references cited therein for further considerations on this issue. *Second,* in order to circumvent the problem of intransitive effects, the group-wise distribution functions F_i may be compared to the same reference distribution. Usually, this is the pooled distribution function $H = \frac{1}{N}\sum_{i=1}^{a} n_i F_i$ (Kruskal 1952; Kruskal and Wallis 1952), resulting in so-called (e.g., Brunner et al. 2017) relative effects $r_i = \int H dF_i$. Multivariate generalizations of this approach can be found in Puri and Sen (1971), Munzel and Brunner (2000) or Brunner et al. (2002); see also De Neve and Thas (2015) for a related approach. Since these quantities depend on the sample sizes n_i, however, they are no fixed model constants and changing the sample sizes might dramatically alter the results; see again Brunner et al. (2017) for an example in the univariate case. For this reason, Brunner and Puri (2001) proposed a different nonparametric effect $p_i = \int G dF_i$ for univariate factorial designs, where $G = \frac{1}{a}\sum_{i=1}^{a} F_i$ denotes the unweighted mean of all distribution functions. The same approach has also been

extended to other settings by Gao and Alvo (2005, 2008), Gao et al. (2008) and Umlauft et al. (2017). Nevertheless, none of them considered null hypotheses formulated in terms of fixed and meaningful model parameters. For a more intuitive interpretation of the results, however, it is sensible to formulate and test hypotheses in more vivid effect sizes. In particular, it is widely accepted in quantitative research that "effect sizes are the most important outcome of empirical studies" (Lakens 2013). Brunner et al. (2017) therefore infer null hypotheses stated in terms of the unweighted nonparametric effects via $H_0^p : \mathbf{Hp} = \mathbf{0}$ for a suitable hypothesis matrix $\mathbf{H}$ and the pooled vector $\mathbf{p}$ of the effects p_i, see also Konietschke et al. (2012) for the special case of one group repeated measures.

In the present paper, we strive to generalize their models and methods in several directions:

1. We examine generalizations to the more involved context of multivariate data where dependencies between observations from the same unit need to be taken into account. This multivariate case allows for testing hypotheses on the influence of several factors on single or several outcome measurements.
2. More general as in repeated measures designs the outcomes in different components may be measured on different units (such as grams and meters). In particular, they actually need not even be elements of metric spaces; totally ordered sets serve equally well as spaces of outcomes because we develop rank-based methods for our analyses. Query scores are an example of such ordered data without having a unit in general. Differences will be tested with the help of a quadratic form in the rank-based effect estimates.
3. This test statistic is analyzed by means of modern empirical process theory instead of the more classical and sometimes laborious projection-based approaches for rank statistics; that is, the proofs of the asymptotic properties become much shorter with the present technique. Since it is asymptotically non-pivotal, appropriate bootstrap methods for asymptotically reproducing its correct limit null distribution are proposed. As resampling entails several good properties when applied to empirical distribution functions and our rank-based estimates offer a representation as a functional of multiple empirical distribution functions, we obtain reliable inference methods using bootstrap techniques as shown by simulation results. These indicate a good control of the type-I error rate even for small sample setups with ordinal or heteroscedastic metric data.

Our model formulation thereby comprises novel procedures for general multivariate factorial designs with crossed or nested factors and even contains the so-called nonparametric multivariate Behrens–Fisher problem as a special case. Moreover, the methodology also allows for subsequent post hoc tests.

The paper is organized as follows: in Sect. 2 we describe the statistical model and the null hypotheses of interest. Section 3 presents the asymptotic properties of our estimator and, subsequently, states the asymptotic validity of its bootstrap versions. Deduced statistical inference procedures for nonparametric MANOVA designs are discussed in Sect. 4 and opposed to Repeated Measures Analyses. The methods' small sample behaviors are analyzed in extensive simulation studies in Sect. 5. Section 6 contains the real data analysis of the gender influence on education and annual house-

hold income of shopping mall customers in the San Francisco Bay Area. We conclude with some final remarks in Sect. 7. The proofs of all theoretic results are given in "Appendix A" and the derivation of the asymptotic covariance matrices is the content of Section 10 in the supplement.

2 Statistical model

Throughout, let $(\Omega, \mathcal{A}, P)$ be a probability space on which all random variables will be defined. We assume a general factorial design with multivariate data, that is, we consider independent random vectors

$$\mathbf{X}_{ik} = (X_{ijk})_{j=1}^{d} \colon \Omega \longrightarrow \mathbb{R}^d, \quad i = 1, \ldots, a; \quad k = 1, \ldots, n_i \tag{1}$$

of dimension $d \in \mathbb{N}$, where X_{ijk} denotes the jth measurement of individual k in group i. Thus, the total sample size is $N = \sum_{i=1}^{a} n_i$. The distribution of $\mathbf{X}_{ik}$ is assumed to be the same within each group with marginals denoted by

$$X_{ijk} \sim F_{ij}, \quad i = 1, \ldots, a, \quad k = 1, \ldots, n_i, \quad j = 1, \ldots, d.$$

Throughout, we understand all F_{ij} as the so-called normalized distribution functions, i.e. the means of their left- and right-continuous versions (Ruymgaart 1980; Akritas et al. 1997; Munzel 1999). This allows for a unified treatment of metric and ordinal data and will later on lead to statistics formulated in terms of mid-ranks. We denote the single and the pooled samples by

$$\mathcal{X}_i = \{\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}\}, \quad i = 1, \ldots, a, \quad \text{and} \quad \mathcal{X} = \bigcup_{i=1}^{a} \mathcal{X}_i,$$

respectively. Different to the special case of repeated measurements (Konietschke et al. 2012; Brunner et al. 2017) the components are in general not commensurate. Therefore, comparisons between the different groups are performed component-wise. To this end, let $G_j = \frac{1}{a} \sum_{i=1}^{a} F_{ij}$, $j = 1, \ldots, d$ denote the unweighted mean distribution function for the jth component. We consider G_j as a benchmark distribution for comparisons in the jth component. In particular, denote by $Y_j \sim G_j$ a random variable that is independent of $\mathcal{X}$ and define unweighted nonparametric effects for group i and component j by

$$p_{ij} = P(Y_j < X_{ij1}) + \tfrac{1}{2} P(Y_j = X_{ij1}) = \int G_j \, dF_{ij} = \frac{1}{a} \sum_{\ell=1}^{a} w_{\ell ij} = \overline{w}_{\cdot ij}, \tag{2}$$

where $w_{\ell ij} = \int F_{\ell j} \, dF_{ij} = P(X_{\ell j1} < X_{ij1}) + \frac{1}{2} P(X_{\ell j1} = X_{ij1})$ quantifies the Mann–Whitney effect for groups ℓ and i in component j. Note that $w_{\ell ij} = 1/2$ in case of $\ell = i$. This definition naturally extends the univariate effect measure given in Brunner et al. (2017) to our general multivariate setup. Note that, in contrast to their

suggestion for an extension to repeated measures designs, comparisons with respect to the overall mean distribution $G = \frac{1}{ad}\sum_{i=1}^{a}\sum_{j=1}^{d} F_{ij}$ are not appropriate here since we study a more general model that allows for components measured on different units. However, the advantages of an unweighted effect measure as discussed in Brunner et al. (2017) still apply: the p_{ij}'s in (2) are fixed model quantities that do not depend on the sample sizes $n_1, \ldots, n_a$, thus allowing for a transitive ordering. Moreover, interpretation of these effects is rather simple: an effect p_{ij} smaller than 1/2 means that observations from the distribution F_{ij} (i.e. from component j in group i) tend to smaller values than those from the corresponding *benchmark distribution* G_j.

In this setup, we formulate null hypotheses as $H_0^p : \mathbf{Hp} = \mathbf{0}$ where $\mathbf{p} = (p_{11}, p_{12}, \ldots, p_{ad})'$ denotes the vector of the relative effects $p_{ij}, i = 1, \ldots, a,\ j = 1, \ldots, d$ and $\mathbf{H}$ is a suitable hypothesis matrix with ad columns. Instead of $\mathbf{H}$ we may equivalently use the unique projection matrix $\mathbf{T} = \mathbf{H}'(\mathbf{H}\mathbf{H}')^{+}\mathbf{H}$ which is idempotent and symmetric and fulfills $\mathbf{Hp} = \mathbf{0} \Leftrightarrow \mathbf{Tp} = \mathbf{0}$; see e.g., Brunner et al. (1997, 2017) and Brunner and Puri (2001). Henceforth, let $\mathbf{I}_d$ and $\mathbf{J}_d$ denote the d-dimensional identity matrix and the $d \times d$ matrix of 1's, respectively, and define by $\mathbf{P}_d = \mathbf{I}_d - \frac{1}{d}\mathbf{J}_d$ the so-called d-dimensional centering matrix.

In particular, in case of $a = 2$ our approach includes the nonparametric multivariate Behrens–Fisher problem

$$H_0^p(\mathbf{T}) : \{\mathbf{Tp} = \mathbf{0}\} = \{\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{1}_d/2\}$$

with $\mathbf{T} = \mathbf{P}_2 \otimes \mathbf{I}_d = \frac{1}{2}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right) \otimes \mathbf{I}_d$ and $\mathbf{p}_i = (p_{i1}, \ldots, p_{id})',\ i = 1, 2$, where $\otimes$ denotes the Kronecker product. Similarly, one-way layouts are covered by choosing $\mathbf{T} = \mathbf{P}_a \otimes \mathbf{I}_d$, leading to the null hypothesis $H_0^p(\mathbf{T}) : \{\mathbf{p}_1 = \cdots = \mathbf{p}_a\}$. Moreover, more complex factorial designs can be treated as well by splitting up the group index i into sub-indices $i_1, i_2, \ldots$ according to the number of factors considered. For example, consider a two-way layout with crossed factors A and B with levels $i_1 = 1, \ldots, a$ and $i_2 = 1, \ldots, b$, respectively. In this case, the random vectors in (1) become $\mathbf{X}_{i_1 i_2 k}, i_1 = 1, \ldots a,\ i_2 = 1, \ldots, b,\ k = 1, \ldots, n_{i_1 i_2}$. We thus obtain the effect vector $\mathbf{p} = (\mathbf{p}'_{11}, \ldots, \mathbf{p}'_{ab})'$, where all vectors $\mathbf{p}_{i_1 i_2} = (p_{i_1 i_2 1}, \ldots, p_{i_1 i_2 d})', i_1 = 1, \ldots, a, i_2 = 1, \ldots, b$ are d-variate and there are $n_{i_1 i_2} > 0$ subjects observed at each factor level combination. Hypotheses of interest in this context are the hypotheses of no main effects as well as the hypothesis of no interaction effect between the factors. The hypothesis of no main effect of factor A can be written as $H_0^p(\mathbf{T}_A) : \{\mathbf{T}_A\mathbf{p} = \mathbf{0}\} \equiv \{(\mathbf{P}_a \otimes \frac{1}{b}\mathbf{J}_b \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}\}$. Similarly, the hypothesis of no effect of factor B is formulated as $H_0^p(\mathbf{T}_B) : \{\mathbf{T}_B\mathbf{p} = \mathbf{0}\} \equiv \{(\frac{1}{a}\mathbf{J}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}\}$ and the hypothesis of no interaction effect as $H_0^p(\mathbf{T}_{AB}) : \{\mathbf{T}_{AB}\mathbf{p} = \mathbf{0}\} \equiv \{(\mathbf{P}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}\}$. For other covered factorial designs and corresponding contrast matrices we refer to Section 4 in Konietschke et al. (2015). Equivalent formulations of the above null hypotheses in terms of the illustrative but notationally more elaborate decomposition into all factor influences are given in the Supplementary Material of Brunner et al. (2017) for the univariate case, but directly carry over to the present context. We note that in the general multivariate case, null hypotheses like H_0^p have only been considered in the special

case of the nonparametric Behrens–Fisher problem (Brunner et al. 2002). Up to now, multivariate testing procedures for one-, two-, or even higher-way layouts focus on null hypotheses formulated in terms of distribution functions; see, e.g., Bathke et al. (2008) and Harrar and Bathke (2008, 2012) and the references given in Sect. 1.

To estimate the vector of effects, we consider the empirical (normalized) distribution functions $\widehat{F}_{ij}(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c(x - X_{ijk})$ where $c(u) = \mathbb{1}\{u > 0\} + \frac{1}{2}\mathbb{1}\{u = 0\}$. Thus, we obtain estimators for the nonparametric effects p_{ij} by replacing the distribution functions with their empirical counterparts

$$\widehat{p}_{ij} = \int \widehat{G}_j d\widehat{F}_{ij} = \frac{1}{a}\sum_{\ell=1}^{a} \widehat{w}_{\ell ij},$$

where $\widehat{G}_j = \frac{1}{a}\sum_{\ell=1}^{a} \widehat{F}_{\ell j}$ and

$$\widehat{w}_{\ell ij} = \int \widehat{F}_{\ell j} \mathrm{d}\widehat{F}_{ij} = \frac{1}{n_\ell}\frac{1}{n_i}\sum_{k=1}^{n_i}\sum_{r=1}^{n_\ell} c(X_{ijk} - X_{\ell jr}) = \frac{1}{n_\ell}\left(\overline{R}_{ij\cdot}^{(\ell i)} - \frac{n_i + 1}{2}\right).$$

Here, $R_{ijk}^{(\ell i)}$ denotes the (mid-)rank of observation X_{ijk} in dimension j among the $(n_i + n_\ell)$ observations in the pooled sample $X_{\ell j1}, \ldots, X_{\ell jn_\ell}, X_{ij1}, \ldots, X_{ijn_i}$ and $\overline{R}_{ij\cdot}^{(\ell i)} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ijk}^{(\ell i)}$ are the corresponding rank means. We combine all estimated effect sizes into the ad-dimensional vector $\widehat{\mathbf{p}} = (\widehat{p}_{11}, \widehat{p}_{12}, \ldots, \widehat{p}_{ad})'$. To detect deviations from null hypotheses of the form $H_0^p(\mathbf{T}) : \{\mathbf{Tp} = \mathbf{0}\}$ we propose the following ANOVA-type test statistic (ATS)

$$T_N = N\widehat{\mathbf{p}}'\mathbf{T}\widehat{\mathbf{p}}, \tag{3}$$

where again $N = \sum_{i=1}^{a} n_i$ denotes the total sample size in the experiment. The quadratic form (3) has the advantage that it detects any deviations from the null hypothesis. In particular, expanding $T_N = N[(\widehat{\mathbf{p}} - \mathbf{p}) + \mathbf{p}]'\mathbf{T}[(\widehat{\mathbf{p}} - \mathbf{p}) + \mathbf{p}]$, our theoretical results from Sect. 3 will prove that T_N converges to infinity in probability under $H_1^p(\mathbf{T}) : \{\mathbf{Tp} \neq \mathbf{0}\}$ because $N\mathbf{p}'\mathbf{Tp}$ does not vanish. Moreover, we note that a more conventional Wald-type statistic involving some generalized inverse of a consistent covariance matrix estimator of $\widehat{\mathbf{p}}$ would be questionable. In particular, different to the special case $a = 2$ (Brunner et al. 2002) of the multivariate Behrens–Fisher problem, there might be potential rank jumps for general a, see the discussion on page 9 in Brunner et al. (2017) for the univariate case $d = 1$.

3 Asymptotic properties and resampling methods

We now turn to the asymptotic properties of the estimated effect sizes $\widehat{\mathbf{p}}$ and propose bootstrap methods to approximate its unknown limit distribution. For a lucid presentation of the results, we thereby assume the following sample size condition:

Condition 1 $\frac{n_i}{N} \to \lambda_i \in (0,1)$ *for all groups* $i = 1, \ldots, a$ *as* $N \to \infty$.

In other words, no group shall constitute a vanishing fraction of the combined sample. Due to the Glivenko–Cantelli theorem in combination with the continuous mapping theorem, the consistency of $\widehat{\mathbf{p}}$ for $\mathbf{p}$ follows already under the weaker assumption $\min_{1 \le i \le a}(n_i) \to \infty$. Asymptotic normality is established in our main theorem below:

Theorem 1 *Suppose Condition 1 holds. As* $N \to \infty$, *we have*

$$\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} \mathbf{Z} \sim N_{ad}(\mathbf{0}_{ad}, \boldsymbol{\Sigma}), \tag{4}$$

where $\mathbf{0}_{ad} \in \mathbb{R}^{ad}$ *is the zero vector and the covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{ad \times ad}$ *is stated in Section 10 in the supplement.*

Note that this theorem immediately implies the asymptotic normality of $\sqrt{N}\mathbf{T}\widehat{\mathbf{p}}$ under $H_0^p(\mathbf{T}) : \{\mathbf{Tp} = \mathbf{0}\}$. Thus, the continuous mapping theorem yields the corresponding convergence in distribution for the quadratic form T_N defined in (3):

Corollary 1 *Suppose Condition 1 holds. As* $N \to \infty$, *we have under* $H_0^p(\mathbf{T}) : \{\mathbf{Tp} = \mathbf{0}\}$

$$T_N = N\widehat{\mathbf{p}}'\mathbf{T}\widehat{\mathbf{p}} \xrightarrow{d} \mathbf{Z}'\mathbf{T}\mathbf{Z} \overset{d}{=} \sum_{h=1}^{ad} \nu_h Y_h^2, \tag{5}$$

where $Y_1, \ldots, Y_{ad}$ *are independent and standard normally distributed and* $\nu_1, \ldots, \nu_{ad} \ge 0$ *are the eigenvalues of* $\boldsymbol{\Sigma}^{1/2}\mathbf{T}\boldsymbol{\Sigma}^{1/2}$.

The limit Theorem 1 raises the question how to find adequate critical values for tests in T_N. A first naive idea is the approximation of the right-hand side of (5) by combining the representation as a weighted sum of independent χ^2-variables with estimators for the involved eigenvalues ν_h or via estimating the covariance matrix $\boldsymbol{\Sigma}$. However, such choices usually result in too liberal inference methods as already observed by Brunner et al. (2017) in the univariate case. Another idea is to generalize the F-approximation proposed in Brunner et al. (2017) to the present situation. But since this will in general not lead to asymptotically correct level α tests (even in the simplest univariate two-sample setting with $a = 2$ and $d = 1$, see Brunner et al. 2017), we instead focus on resampling the test statistic T_N. In particular, we think that this is the only reasonable way to end up with an asymptotically exact testing procedure. To this end, we study two bootstrap approaches for recovering the unknown limit distribution of T_N under H_0^p: a group-wise as well as a wild bootstrap.

First, a variant of the classical bootstrap (Efron 1979) applied to the present context is considered: for each group $i = 1, \ldots, a$, draw n_i times randomly with replacement from the d-dimensional data vectors $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}$ to obtain the ith *bootstrap sample* $\mathbf{X}_{i1}^*, \ldots, \mathbf{X}_{in_i}^*$. Denote their marginal empirical distribution functions as F_{ij}^*, $j = 1, \ldots, d$. These are the bootstrap counterparts of all $\widehat{F}_{ij}$. We denote by ${p_{ij}}^*$ and $\mathbf{p}^*$ the bootstrap versions of $\widehat{p}_{ij}$ and $\widehat{\mathbf{p}}$, respectively. They are based on the bootstrapped empirical distribution functions. Finally, the bootstrap version of the test

statistic is $T_N^* = N(\mathbf{p}^* - \widehat{\mathbf{p}})'\mathbf{T}(\mathbf{p}^* - \widehat{\mathbf{p}})$. The following two limit theorems reveal that it always approaches the null distribution of T_N. The theorems hold under both the null hypothesis $H_0^p(\mathbf{T}) : \{\mathbf{Tp} = \mathbf{0}\}$ and the alternative hypothesis $H_1^p(\mathbf{T}) : \{\mathbf{Tp} \neq \mathbf{0}\}$.

Theorem 2 *Suppose Condition 1 holds. As $N \to \infty$, we have, conditionally on $\mathcal{X}$,*

$$\sqrt{N}(\mathbf{p}^* - \widehat{\mathbf{p}}) \xrightarrow{d} \mathbf{Z} \sim N_{ad}(\mathbf{0}_{ad}, \mathbf{\Sigma}) \tag{6}$$

in outer probability, where $\mathbf{\Sigma}$ is as in Theorem 1.

Corollary 2 *Suppose Condition 1 holds. As $N \to \infty$, we have, conditionally on $\mathcal{X}$,*

$$T_N^\star = N(\mathbf{p}^* - \widehat{\mathbf{p}})'\mathbf{T}(\mathbf{p}^* - \widehat{\mathbf{p}}) \xrightarrow{d} \mathbf{Z}'\mathbf{T}\mathbf{Z} \overset{d}{=} \sum_{h=1}^{ad} \nu_h Y_h^2 \tag{7}$$

in outer probability, i.e. the same limit distribution as in Corollary 1.

The second proposed resampling procedure is the wild bootstrap which recently has been proposed for the analysis of nonparametric repeated measures designs (Friedrich et al. 2017; Umlauft et al. 2019). To transfer it to the present MANOVA context, we first notice that $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$ has an asymptotically linear representation in $\sqrt{N}((\widehat{F}_{11}, \ldots, \widehat{F}_{ad})' - (F_{11}, \ldots, F_{ad})')$. Indeed, if we define $\phi_i(f_1, \ldots, f_a) = \int(\frac{1}{a}\sum_{\ell=1}^a f_\ell)\mathrm{d}f_i$ for functions $f_1, \ldots, f_a$ such that the integral is well defined, then

$$\begin{aligned}
\sqrt{N}(\widehat{p}_{ij} - p_{ij}) &= \sqrt{N}(\phi_i(\widehat{F}_{1j}, \ldots, \widehat{F}_{aj}) - \phi_i(F_{1j}, \ldots, F_{aj})) \\
&= \sqrt{N}\left[\int(\widehat{G}_j - G_j)\mathrm{d}F_{ij} - \int(\widehat{F}_{ij} - F_{ij})\mathrm{d}G_j + \int(\widehat{G}_j - G_j)\mathrm{d}(\widehat{F}_{ij} - F_{ij})\right] \\
&= \sqrt{N}\int(\widehat{G}_j - G_j)\mathrm{d}F_{ij} - \sqrt{N}\int(\widehat{F}_{ij} - F_{ij})\mathrm{d}G_j + o_p(1).
\end{aligned} \tag{8}$$

The last equality follows from the functional delta-method applied to ϕ_i; cf. Dobler and Pauly (2018) for the two-sample case. It is sometimes also called the asymptotic equivalence theorem (Brunner et al. 2017). Now, in the proposed application of the wild bootstrap the residuals $\varepsilon_{ijk}(x) = c(x - X_{ijk}) - F_{ij}(x), k = 1, \ldots, n_i$, in the above asymptotic expansion are resampled. In particular, the residuals are replaced with

$$\widehat{\varepsilon}_{ijk}(x) = D_{ik} \cdot [c(x - X_{ijk}) - \widehat{F}_{ij}(x)], \quad k = 1, \ldots, n_i.$$

Here, $D_{ik}, i = 1, \ldots, a, \quad k = 1, \ldots, n_i$, are so-called *wild bootstrap multipliers* which are i.i.d. with zero-mean and unit variance. These resampled residuals are also centered and their conditional variance can be considered as the empirical counterpart of $var(\varepsilon_{ijk}(x)) = F_{ij}(x)(1 - F_{ij}(x))$:

$$var(\widehat{\varepsilon}_{ijk}(x) \mid \mathcal{X}) = [c(x - X_{ijk}) - \widehat{F}_{ij}(x)]^2,$$

the expectation of which equals $var(\varepsilon_{ijk}(x))$. Lastly, we require that the multipliers fulfill $\int_0^\infty \sqrt{P(|D_{11}| > x)}\mathrm{d}x < \infty$ which is implied by $E|D_{11}|^{2+\eta} < \infty$ for any $\eta > 0$. Hence, it is a weak assumption on the tail heaviness; cf. p. 177 in van der Vaart and Wellner (1996). Note that, for each i, k, our wild bootstrap implementation uses the same multiplier D_{ik} for every component j in order to ensure an appropriate dependence structure across the components.

Additionally, apart from estimating the residuals ε_{ijk} by $\widehat{\varepsilon}_{ijk}$, the unknown distribution functions $F_{\ell j}$ in the integrators in (8) need to be estimated by their empirical counterparts. Denote by $F_{ij}^\star = \frac{1}{n_i}\sum_{k=1}^{n_i}\widehat{\varepsilon}_{ijk}$ and $G_j^\star = \frac{1}{a}\sum_{\ell=1}^{a} F_{\ell j}^\star$ the wild bootstrap versions of $\widehat{F}_{ij} - F_{ij}$ and $\widehat{G}_j - G_j$, respectively. We obtain the following wild bootstrap counterpart of $\widehat{p}_{ij} - p_{ij}$:

$$p_{ij}^\star = \int G_j^\star(x)\mathrm{d}\widehat{F}_{ij}(x) - \int F_{ij}^\star(x)\mathrm{d}\widehat{G}_j(x). \tag{9}$$

A representation of $p_{ij}^\star$ in terms of mid-ranks is given in Section 8 of the supplement. This is particularly useful for a computationally efficient implementation of the wild bootstrap procedure. We find the following conditional central limit theorems for $\mathbf{p}^\star = \sqrt{N}(p_{11}^\star, p_{12}^\star, \ldots, p_{ad}^\star)'$ and $T_N^\star = \mathbf{p}^{\star\prime}\mathbf{T}\mathbf{p}^\star$, which again hold under both $H_0^p(\mathbf{T})$ and $H_1^p(\mathbf{T})$:

Theorem 3 *Suppose Condition 1 holds. As $N \to \infty$, we have, conditionally on $\mathcal{X}$,*

$$\mathbf{p}^\star \xrightarrow{d} \mathbf{Z} \sim N_{ad}(\mathbf{0}_{ad}, \boldsymbol{\Sigma}) \tag{10}$$

in outer probability, where $\boldsymbol{\Sigma}$ is as in Theorem 1.

Corollary 3 *Suppose Condition 1 holds. As $N \to \infty$, we have, conditionally on $\mathcal{X}$,*

$$T_N^\star = \mathbf{p}^{\star\prime}\mathbf{T}\mathbf{p}^\star \xrightarrow{d} \mathbf{Z}'\mathbf{T}\mathbf{Z} \overset{d}{=} \sum_{h=1}^{ad} \nu_h Y_h^2 \tag{11}$$

in outer probability, i.e. the same limit distribution as in Corollary 1.

Each of the conditional central limit theorems (7) and (11) gives a theoretic justification of the use of the random quantiles of T_N^* and $T_N^\star$ conditional on $\mathcal{X}$. It follows that these quantiles converge in outer probability to the quantiles of the asymptotic distribution of the ATS under $H_0^p(\mathbf{T})$, i.e. of $\sum_{h=1}^{ad} \nu_h Y_h^2$. For a practical implementation, such critical values are obtained for given $\mathcal{X}$ via Monte-Carlo simulations. To this end, numerous independent realizations of T_N^* or $T_N^\star$ are realized. Their empirical quantiles then serve as critical values for the hypothesis tests.

Comparing both proposed resampling procedures, we see that the classical has a theoretical advantage over the wild bootstrap: the $o_p(1)$ term in (8) is implicitly resampled as well. In principle, a wild bootstrap counterpart $\sqrt{N}\int F_{ij}^\star \mathrm{d}G_j^\star$ could also be added to the definition $p_{ij}^\star$ in order to resample the remainder term. However, the massively increased computational load makes such a modification impractical. Beyond

these theoretical points, we will evaluate the final comparison of both resampling schemes in extensive simulations in Sect. 5.

4 Deduced inference procedures

Nonparametric MANOVA The previous considerations directly imply that consistent and asymptotic level α tests for $H_0^p(\mathbf{T}) : \{\mathbf{Tp} = \mathbf{0}\}$ are given by

$$\varphi_N^\star = \mathbb{1}\{T_N > c^\star(\alpha)\} \text{ and } \varphi_N^* = \mathbb{1}\{T_N > c^*(\alpha)\},$$

where $c^*(\alpha)$ and $c^\star(\alpha)$ denote the $(1-\alpha)$ quantile of the group-wise bootstrap and wild bootstrap versions of T_N given $\mathcal{X}$, i.e. of $T_N^\star$ in case of the wild bootstrap. Their finite sample performance will be studied in Sect. 5. As described in Sect. 2, these tests can be used to infer various global null hypotheses of interest about (nonparametric) main and interaction effects of interest which can straightforwardly be inverted to construct confidence regions for these nonparametric effects.

Moreover, the results derived in Sect. 3 also allow post hoc analyses, i.e. subsequent multiple comparisons. To exemplify the typical paths of action we consider the one-way situation with a independent groups and nonparametric effect size vectors $\mathbf{p}_i = (p_{i1}, \ldots, p_{id})'$ in group $i, i = 1, \ldots, a$. If the global null hypothesis

$$H_0^p(\mathbf{P}_a \otimes \mathbf{I}_d) : \{\mathbf{p}_1 = \cdots = \mathbf{p}_a\}$$

of equal effect size vectors is rejected, one is usually interested in inferring

(i) the (univariate) endpoints that caused the rejection, as well as
(ii) the groups showing significant differences (all pairs comparisons).

The above questions directly translate to testing the univariate hypotheses

$$H_{0j}^p : \{p_{1j} = \cdots = p_{aj}\}, \quad j = 1, \ldots, d \tag{12}$$

in case of (i) and to an all pairs comparison given by multivariate hypotheses

$$H_{0i\ell}^p : \{\mathbf{p}_i = \mathbf{p}_\ell\}, \quad 1 \le i < \ell \le a \tag{13}$$

in case of (ii). Note that our derived methodology allows for testing these hypotheses in a unified way by performing tests on all univariate endpoints for (12) and by selecting pairwise comparison contrast matrices for (13). Therefore, a first naive approach would be to adjust the individual tests accordingly (e.g. by Bonferroni or Holm corrections) to ensure control of the family-wise error rate. However, note that the effect size vectors are defined via component-wise comparisons. This implies that the intersection of all H_{0j}^p as well as the intersection of all $H_{0i\ell}^p$ is exactly given by the global null hypothesis $H_0^p(\mathbf{P}_a \otimes \mathbf{I}_d)$. Moreover, we can even test all subset intersections of H_{0j}^p, $j = 1, \ldots, d$ (or $H_{0i\ell}^p$, $1 \le i < \ell \le a$) by choosing adequate contrast matrices and performing the corresponding bootstrap procedures. Thus, both questions can even be treated (separately) by applying the closed testing principle of Marcus et al.

(1976). This is a major advantage over existing inference procedures that are developed for testing null hypotheses formulated in terms of distribution functions (Ellis et al. 2017). In particular, since equality of marginals does not imply equality of multivariate distributions, the closed testing principle cannot be applied to the latter to answer question (i).

To ensure a reasonable computation time, the above approach is only applicable for small or moderate d and a. However, some computation time can be saved by formulating a hierarchy on the questions (either for study-specific reasons or by weighing up the sizes of a and p). For example, assume that (i) is more important than (ii). In this case we may start by applying the closed testing algorithm to test hypotheses H_{0j}^{p} and subsequently only infer pairwise comparisons on the significant univariate endpoints (instead of testing all multivariate $H_{0i\ell}^{p}$). Contrary, assume that d is much larger than a. Then it may be reasonable to first infer (ii) and subsequently consider (i) for the significant pairs.

Confidence intervals and regions The bootstrap critical values of the tests $\varphi_N^\star$ and φ_N^* immediately give rise to confidence regions for contrasts in the effects vector $\mathbf{p}$: let $T_N(\mathbf{p}_0) = N(\widehat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{T}(\widehat{\mathbf{p}} - \mathbf{p}_0)$ be a test statistic for testing the null hypothesis $\{\mathbf{Hp} = \mathbf{Hp}_0\}$ for a given effects vector $\mathbf{p}_0$ and a fixed contrast matrix $\mathbf{H}$, where again $\mathbf{T} = \mathbf{H}'(\mathbf{HH}')^{+}\mathbf{H}$ denotes the corresponding unique projection matrix. Then, Corollary 3 guarantees that

$$\sup_x |P_{\mathbf{p}_0}(T_N(\mathbf{p}_0) \leq x) - P(T_N^\star \leq x|\mathbf{X})| \to 0$$

in outer probability, i.e. $\mathbb{1}\{T_N(\mathbf{p}_0) > c^\star(\alpha)\}$ is an asymptotic level α wild bootstrap test for $\{\mathbf{Hp} = \mathbf{Hp}_0\}$. Here, $P_{\mathbf{p}_0}(T_N(\cdot) \leq x)$ denotes a distribution function of $T_N(\cdot)$ under the assumption that $\mathbf{p} = \mathbf{p}_0$ holds. Similarly, $\mathbb{1}\{T_N(\mathbf{p}_0) > c^*(\alpha)\}$ is a groupwise bootstrap test of asymptotic level α for the same null hypothesis. Inverting these tests thus leads to the asymptotic $(1-\alpha)$ confidence ellipsoids

$$C_{1-\alpha} = \{\mathbf{Hp} : T_N(\mathbf{p}) \leq c(\alpha)\} = \{\mathbf{Hp} : N(\widehat{\mathbf{p}} - \mathbf{p})'\mathbf{T}(\widehat{\mathbf{p}} - \mathbf{p}) \leq c(\alpha)\} \tag{14}$$

for $\mathbf{Hp}$, where $c(\alpha)$ denotes any of the bootstrap critical values $c^\star(\alpha)$ or $c^*(\alpha)$. In case of a contrast vector $\mathbf{H}' = \mathbf{h} \in \mathbb{R}^{ad}$, this leads to confidence intervals for linear contrasts $\mathbf{h}'\mathbf{p}$.

Comparison with repeated measures analysis In mean-based MANOVA, classical profile and repeated measures analyses are already incorporated by choosing adequate hypotheses matrices to test for certain time effects or specific profiles, see, e.g. Friedrich and Pauly (2018) and the discussion in Bathke et al. (2018). However, in our nonparametric case, this would not be the best option as our MANOVA approach does not make use of the repeated measures' commensurate nature. For the latter, Brunner et al. (2017) already outlined extensions to general repeated measures designs in Sect. 5 of their paper which have been picked up in more detail by Umlauft et al. (2019). Here, the effects are not defined as probabilities with respect to a componentwise mean distribution G_j as in (2) but with respect to an overall mean distribution. The resulting effects are thus more meaningful in a repeated measures analysis but

also completely meaningless in a general MANOVA setting with non-commensurate entries.

5 Simulations

In the previous section, we analyzed the large sample properties of the proposed inference procedures. Here, we additionally analyze their finite sample properties. In particular, we oppose the type I error rate control of both bootstrap procedures for several designs, covering

- homo- and heteroscedastic situations with continuous data and
- situations with ordinal data in
- one- and two-factorial MANOVA designs.

In case of the multivariate Behrens–Fisher problem with $a = 2$, we additionally compared their results with the ANOVA-type testing procedure of Brunner et al. (2002). This approach is based on approximating the unknown distribution of the test statistic by an F-distribution with estimated degrees of freedom and thus does not provide an asymptotically valid procedure. The results of these simulations are therefore only presented in Section 9 of the supplementary material. In summary, this test provided a more conservative behavior compared to both bootstrap methods; especially in all heteroscedastic settings under consideration.

In all simulation setups, we chose the significance level $\alpha = 5\%$. Moreover, we also compare the power of the two bootstrap procedures. All simulations were conducted using the R-computing environment (Core Team 2016), version 3.2.3, each with 5000 simulation runs and 5000 bootstrap iterations.

On the two bootstraps From existing results on both bootstrap procedures in linear models with metric observations (Wu 1986; Davidson and Flachaire 2008) one would expect that the wild bootstrap should outperform the group-wise bootstrap; particularly in heteroscedastic settings. However, as discussed in Sect. 3, the approximation with the wild bootstrap neglects a small $o_P(1)$-term in (8), whereas the group-wise bootstrap does not. Moreover, we are here dealing with a nonparametric situation with potentially ordinal data and dependent (mid)ranks. As a result, the wild bootstrap is faced with a completely different situation, in which the consequences of a possible heteroscedasticity may be additionally mitigated due to the robust nature of the rank-based procedures. For these reasons, it is ad hoc not clear which procedure behaves beneficial in which situation. Our empirical studies below will shed some light on this.

5.1 Simulations under the null hypothesis

5.1.1 Continuous data

For the one-way layout, data were generated similarly to the simulation study in Konietschke et al. (2015). We considered $a = 2$ treatment groups and $d \in \{4, 8\}$ endpoints as well as the following covariance settings:

$$(S1)\ \text{Setting 1:}\ \mathbf{V}_1 = \mathbf{I}_d + 0.5(\mathbf{J}_d - \mathbf{I}_d) = \mathbf{V}_2,$$
$$(S2)\ \text{Setting 2:}\ \mathbf{V}_1 = \left((0.6)^{|r-s|}\right)_{r,s=1}^{d} = \mathbf{V}_2.$$

Setting 1 represents a compound symmetry structure, while Setting 2 is an autoregressive covariance structure. Data were generated as

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

where $\mathbf{V}_i^{1/2}$ denotes a square root of the matrix $\mathbf{V}_i$, i.e., $\mathbf{V}_i = \mathbf{V}_i^{1/2} \cdot \mathbf{V}_i^{1/2}$. The i.i.d. random errors $\boldsymbol{\epsilon}_{ik} = (\epsilon_{i1k}, \ldots, \epsilon_{idk})'$ with mean $E(\boldsymbol{\epsilon}_{ik}) = \mathbf{0}_d$ and $Cov(\boldsymbol{\epsilon}_{ik}) = \mathbf{I}_{d\times d}$ were generated by simulating independent standardized components $\epsilon_{ijk} = (Y_{ijk} - E(Y_{ijk}))/(Var(Y_{ijk}))^{1/2}$ for various distributions of Y_{ijk}. In particular, we simulated standard normal and standard lognormal distributed random variables. We investigated balanced as well as unbalanced designs with sample size vectors $\mathbf{n}^{(1)} = (10, 10)$ and $\mathbf{n}^{(2)} = (10, 20)$, and increased sample sizes by multiplying each element of the respective vector $\mathbf{n}^{(h)}$, $h = 1, 2$, with a factor $m \in \{1, 2, 5\}$. In this setting, we tested the null hypothesis of no treatment effect $H_0^p : \{(\mathbf{P}_a \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}_{2d}\} = \{\mathbf{p}_1 = \mathbf{p}_2\}$, where $\mathbf{p}_i = (p_{i1}, \ldots, p_{id})'$, $i = 1, 2$, and $\mathbf{p} = (\mathbf{p}_1', \mathbf{p}_2')'$.

The results for the normal and lognormal distribution are displayed in Table 1. The group-wise bootstrap approach shows a slightly more liberal behavior across most scenarios. The wild bootstrap, on the other hand, apparently keeps the nominal significance level much better. Both tests approach the nominal level with an increasing sample size. In general, we did not find a big impact of the samples' balanced- or unbalancedness on their type-I error rates.

5.1.2 A heteroscedastic setting

We simulated a heteroscedastic setting, where $H_0^p : \mathbf{Tp} = \mathbf{0}_{2d}$ is satisfied. To this end, we took

$$\mathbf{X}_{ik} \sim N(\mathbf{0}_d, \sigma_i^2 \mathbf{I}_d)$$

for different choices of $\sigma_i^2 \in \{1, 1.2, 2\}$ as well as sample sizes $n_i \in \{10, 20\}$. In contrast to the scenario above, we now distinguish between $\mathbf{n}^{(2)} = (10, 20)$ and $\mathbf{n}^{(3)} = (20, 10)$, since the heteroscedastic setting is not symmetric anymore (except for the case $\sigma_1^2 = \sigma_2^2$, where we again only consider $\boldsymbol{n}^{(2)}$). Sample sizes were again increased as described above. The results are displayed in Table 2. In this case, we observe a very conservative behavior of the wild bootstrap across all scenarios, which improves with growing sample sizes. But even for the sample size factor $m = 5$, the simulated type-I error rates for the wild bootstrap-based tests are as small as 2.3–3.7%. In this heteroscedastic setting, the classical, group-wise bootstrap yields better results in all scenarios, especially for the higher dimension $d = 8$. In particular, it maintains the 5% level very well in this case across varying sample sizes and variances. In fact,

Table 1 Type-I error results in % for the homoscedastic setting with normal and lognormal distributed data with $d = 4$ and $d = 8$ dimensions, varying sample sizes and different covariance settings

Distr	Cov	$\boldsymbol{n}$	Wild bootstrap			Group-wise bootstrap		
			$m = 1$	2	5	1	2	5
$d = 4$								
Normal	S1	(10, 10)	5.9	5.5	5.6	7.3	6.4	5.9
		(10, 20)	5.6	5.5	5.2	6.6	6.3	5.5
	S2	(10, 10)	5.1	5.3	5.6	7.0	6.5	5.9
		(10, 20)	5.2	5.0	5.1	6.8	6.3	5.3
Lognormal	S1	(10, 10)	7.2	6.3	5.8	7.3	6.7	5.8
		(10, 20)	6.8	6.2	5.6	7.4	6.3	5.5
	S2	(10, 10)	6.5	6.1	5.6	7.2	6.6	5.7
		(10, 20)	6.6	5.9	5.5	7.0	6.5	5.7
$d = 8$								
Normal	S1	(10, 10)	6.6	5.7	5.2	7.6	6.2	5.5
		(10, 20)	6.1	5.9	5.1	6.9	6.5	5.4
	S2	(10, 10)	3.7	3.9	4.2	6.8	6.1	5.3
		(10, 20)	3.8	3.9	4.6	6.5	5.7	5.1
Lognormal	S1	(10, 10)	7.8	5.9	5.4	8.3	6.2	5.5
		(10, 20)	7.7	6.3	5.3	8.1	6.4	5.5
	S2	(10, 10)	5.4	4.6	4.7	7.8	6.0	5.5
		(10, 20)	5.3	5.2	4.6	7.1	6.2	5.1

in the lower-dimensional case ($d = 4$) its type-I error control for small sample sizes is even better than in the homoscedastic setting. Thus, different to bootstrapping in linear models with metric data, the group-wise bootstrap is preferred to the wild bootstrap in case of heteroscedasticity.

5.1.3 Ordinal data

We simulated ordinal data using the function *ordsample* from the R package **GenOrd** (Barbiero and Ferrari 2015; Ferrari and Barbiero 2012). The package **GenOrd** allows for simulation of discrete random variables with a given correlation structure and given marginal distributions. The latter are linked via a Gaussian copula function in order to achieve the desired correlation structure on the discrete components. We simulated uniform marginal distributions, such that the outcomes in the jth dimension are uniformly distributed on $j+1$ categories, $1 \leq j \leq d$. For the correlation structure, we used the same underlying covariance matrices as in the homoscedastic setting above. Again, we considered $d \in \{4, 8\}$ dimensions and the same sample sizes as in the homoscedastic setting. The results, which are similar to the ones obtained for continuous data in the homoscedastic setting (Table 1) are displayed in Table 3. We find that the type-I error control for the wild bootstrap-based test is the best. Only in some few scenarios with small sample sizes the test is slightly conservative. On

Table 2 Type-I error results in % for the heteroscedastic setting with $d = 4$ and $d = 8$ dimensions and varying sample sizes

σ_i^2	$\boldsymbol{n}$	Wild bootstrap			Group-wise bootstrap		
		m =1	2	5	1	2	5
$d = 4$							
(1, 2)	(10, 10)	0.6	1.3	3.1	6.3	5.2	5.5
	(10, 20)	0.7	1.7	3.5	5.6	5.2	5.1
	(20, 10)	1.3	2.0	3.3	6.5	5.4	5.2
(1, 1)	(10, 10)	0.5	1.2	3.2	6.0	5.3	5.3
	(10, 20)	0.8	1.7	3.7	5.9	5.8	5.4
(1.2, 1)	(10, 10)	0.6	1.2	3.2	6.2	5.3	5.5
	(10, 20)	0.9	1.9	3.7	6.2	5.9	5.3
	(20, 10)	0.9	1.9	3.5	5.9	5.6	5.2
$d = 8$							
(1, 2)	(10, 10)	0.1	0.4	2.5	4.8	4.7	4.6
	(10, 20)	0.1	0.9	3.2	4.3	4.3	5.3
	(20, 10)	0.2	0.9	3.4	4.3	4.8	5.0
(1, 1)	(10, 10)	<0.01	0.4	2.5	5.1	4.9	4.6
	(10, 20)	0.1	1.1	3.2	4.7	4.4	5.4
(1.2, 1)	(10, 10)	<0.01	0.3	2.3	4.8	4.9	4.8
	(10, 20)	0.2	1.1	3.4	4.9	4.4	5.4
	(20, 10)	0.1	0.7	3.4	4.3	4.8	5.6

the other hand, the group-wise bootstrap-based test is too liberal in most situations with small sample sizes. Moreover, the difference between the number of dimensions considered is not as pronounced as in the heteroscedastic setup.

5.1.4 A two-way layout

In order to see the performance of the proposed methods in a more complicated setup, where paradoxical results as mentioned before can actually occur, we have simulated a 2×2-design similar to Brunner et al. (2018). More precisely, we considered observation vectors $\mathbf{X}_{i\ell k}, i = 1, 2,\ \ell = 1, 2,\ k = 1, \ldots, n_{i\ell}$ following d-variate normal distributions $N_d(\mu_{i\ell}\mathbf{1}_d, \sigma_{i\ell}^2\mathbf{I}_d)$. We chose $d = 4$ and $\boldsymbol{\mu}_{i\ell} = (\mu_{i\ell 1}, \ldots, \mu_{i\ell d})' = (1, 2, 3, 4)'$ for $i, \ell = 1, 2$, such that the null hypothesis of no interaction between the two factors is fulfilled. We considered balanced as well as unbalanced settings with $\mathbf{n} = (25, 25, 25, 25)'$ and $\mathbf{n} = (10, 20, 20, 50)'$, respectively. Furthermore, we distinguished between a homoscedastic scenario setting $\sigma_{i\ell} = 0.4$ and a heteroscedastic scenario with $\boldsymbol{\sigma} = (\sigma_{i\ell})_{i,\ell=1,2} = (0.4, 1.2, 1.2, 4)'$.

In order to keep the design unbalanced while increasing the sample size, sample sizes were again multiplied by a factor m. The results are displayed in Table 4.

We find that whether the design is homoscedastic or does not have a big impact on the results. The wild bootstrap shows a quite conservative behavior which, however,

Table 3 Type-I error rates in % for ordinal data with different sample sizes and different covariance structures

Cov. setting	$\boldsymbol{n}$	Wild bootstrap			Group-wise bootstrap		
		m =1	2	5	1	2	5
$d = 4$							
S1	(10, 10)	6.3	5.5	5.0	8.1	6.6	5.4
	(10, 20)	6.7	5.5	5.1	7.7	6.2	5.6
S2	(10, 10)	5.9	5.1	5.2	7.8	6.8	5.5
	(10, 20)	6.3	5.2	5.2	7.8	6.4	5.6
$d = 8$							
S1	(10, 10)	6.4	5.6	5.9	7.8	6.4	6.1
	(10, 20)	7.2	5.3	5.2	8.2	5.9	5.5
S2	(10, 10)	3.6	3.8	5.1	7.6	6.2	6.2
	(10, 20)	4.5	3.8	4.4	7.4	5.6	4.8

Table 4 Type-I error rates in % for the interaction hypothesis in a 2×2-design with normally distributed data for different sample sizes and variances

Cov.setting	**n**	Wild bootstrap			Group-wise bootstrap		
		$m = 1$	2	5	1	2	5
Homoscedastic	(25, 25, 25, 25)	2.88	3.92	3.94	5.80	5.36	4.30
Heteroscedastic	(25, 25, 25, 25)	2.92	3.84	3.86	5.82	5.46	4.42
Homoscedastic	(10, 20, 20, 50)	1.96	3.40	3.66	5.90	5.54	4.92
Heteroscedastic	(10, 20, 20, 50)	1.90	3.28	3.82	6.06	5.56	4.94

is not as conservative as in the heteroscedastic setup underlying Table 2 which may be explained by the larger total sample size N in this 2×2 design. The group-wise bootstrap keeps the pre-assigned level very well in almost all scenarios. We also note that the wild bootstrap version leads to worse results for unbalanced compared to balanced designs: it becomes much more conservative in the small sample size case $m = 1$.

Finally, we also simulated ordinal data in a 2×2-design. We chose an autoregressive covariance structure as above and binomially distributed marginals, i.e. , for every dimension $j = 1, \ldots, d$ the marginals were binomially distributed with size $j + 2$ and success probability $1/(j + 1)$. Both balanced and unbalanced sample sizes were chosen as above. The results again show a slight superiority of the wild bootstrap in terms of type-I error rate control; see Table 5. The group-wise bootstrap appears to be a bit too liberal for the smallest sample sizes.

5.2 Power simulations

In order to compare the power behavior of the two bootstrap methods, we have considered a shift alternative, i.e., we simulated data in a two-sample setting as

Table 5 Type-I error rates in % for the interaction hypothesis in a 2×2-design with ordinal data and balanced as well as unbalanced sample sizes

n	Wild bootstrap			Group-wise bootstrap		
	$m = 1$	2	5	1	2	5
Balanced	5.42	5.10	5.40	6.26	5.36	5.38
Unbalanced	5.26	4.88	4.74	6.62	5.68	4.88

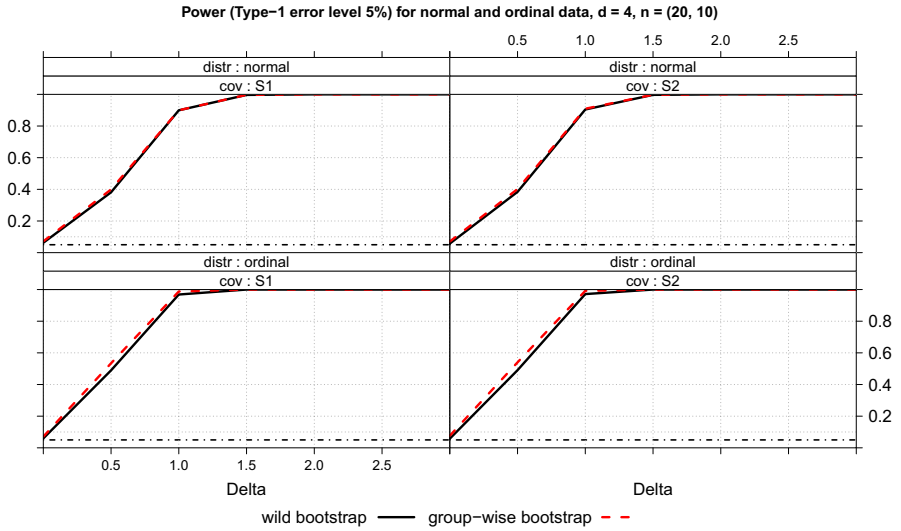

Fig. 1 Power simulation results for continuous (normally distributed) and ordinal data with $d = 4$ dimensions and sample sizes $\mathbf{n} = (20, 10)'$

$$\tilde{\mathbf{X}}_{ik} = \boldsymbol{\mu}_i + \mathbf{X}_{ik}, \quad i = 1, 2; \quad k = 1, \ldots, n_i, \tag{15}$$

where $\boldsymbol{\mu}_1 = \mathbf{0}_d$ and $\boldsymbol{\mu}_2 = (\delta, \ldots, \delta)'$ for $\delta \in \{0, 0.5, 1, 1.5, 2, 3\}$ and $\mathbf{X}_{ik}$ corresponds to the respective random vectors simulated in the one-way setting as specified in Sect. 5.1 above. In fact, we first considered a homoscedastic situation with ordinal or normally distributed random vectors with covariance structure given in Setting 1 respectively 2 in both groups, sample size vector $\mathbf{n} = (20, 10)'$ and dimension $d = 4$.

The results are shown in Fig. 1. As the group-wise bootstrap was slightly liberal compared to the wild bootstrap method in these situations, its power function is marginally larger. Moreover, we also investigated the power behavior for a heteroscedastic situation (Fig. 2). Simulating data as in (15) with $\mathbf{X}_{ik}$ as in Sect. 5.1.2 we again observe that the group-wise bootstrap has slightly larger power than its wild bootstrap counterpart. This can be explained by the rather conservative behavior of the wild bootstrap method seen in Table 2. However, all in all the power of both methods are more or less identical.

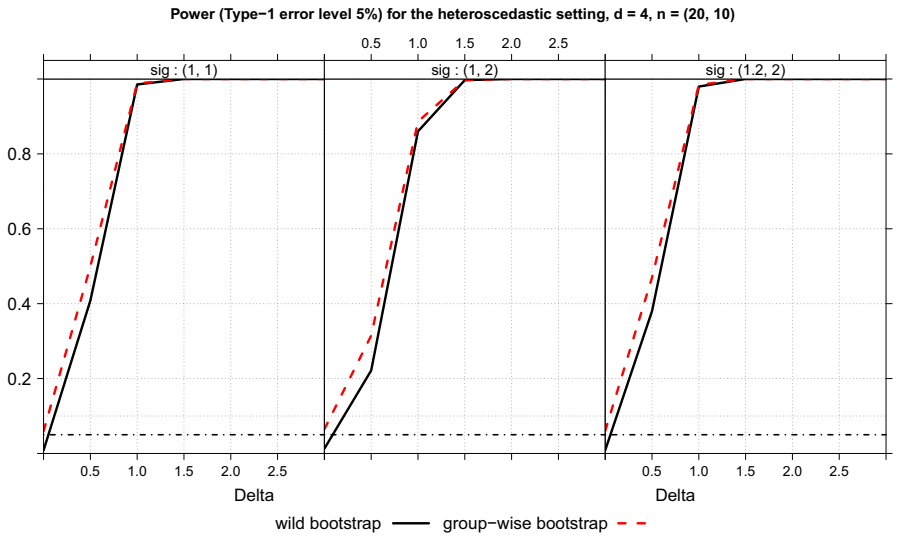

Fig. 2 Power simulation results for heteroscedastic data with $d = 4$ dimensions and $\mathbf{n} = (20, 10)'$

5.3 Runtime comparisons

To further investigate the advantages and disadvantages of both bootstrap approaches, we compared their computational cost in an additional simulation study for a 2×2 design. The results, presented in Section 9.3 in the supplement, show a clear advantage for the wild bootstrap which was most apparent for larger dimensions ($d = 8$). All in all, the group-wise bootstrap increased the running time by factors between 1.2 and 5.7 in the considered scenarios. However, these findings only alter its practical applicability for sample sizes or dimensions larger than the choices considered here, as the longest observed computation time (in the mean) resulted in 2.25 s based on 5000 bootstrap replications.

6 Data example

As a data example, we consider the data set 'marketing' in the R-package **ElemStatLearn** (Halvorsen 2015). This data set contains information on the annual household income along with 13 other demographic factors of shopping mall customers in the San Francisco Bay Area. Most of the variables in this data set are measured on an ordinal scale, rendering mean-based approaches unfeasible. For our example, we consider the influence of sex and language on annual household income and educational status. The annual household income is categorized in 9 categories ranging from 'less than $10,000' to '$75,000 and more', while education ranges from 'Grade 8 or less' to 'Grad Study' (6 categories). This two-dimensional outcome is to be analyzed with respect to the influence factors sex (with levels Male vs. Female) and language (with levels English, Spanish, and Other).

The original data set consists of 8993 observations. After removing those observations with missing values in one of the variables considered here, 8561 observations remain. Correlation of the two outcome variables was assessed by Kendall's Tau ($\tau = 0.37$) and Spearman's correlation coefficient ($\rho = 0.45$), both indicating a moderate positive correlation. Figure 3 shows the empirical distribution functions for the two dimensions for male and female participants while the estimated nonparametric effects are displayed in Table 6. The effects suggest main effects of both factors. To investigate this and a potential nonparametric interaction effect, we have to test the null hypotheses

$$
\begin{aligned}
H_0^p(\mathbf{T}_A) &: \{\mathbf{T}_A\mathbf{p} = \mathbf{0}\} \equiv \{(\mathbf{P}_a \otimes \mathbf{J}_b/b \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}\} && \text{(No Effect of Sex)} \\
H_0^p(\mathbf{T}_B) &: \{\mathbf{T}_B\mathbf{p} = \mathbf{0}\} \equiv \{(\mathbf{J}_a/a \otimes \mathbf{P}_b \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}\} && \text{(No Effect of Language)} \\
H_0^p(\mathbf{T}_{AB}) &: \{\mathbf{T}_{AB}\mathbf{p} = \mathbf{0}\} \equiv \{(\mathbf{P}_a \otimes \mathbf{P}_b \otimes \mathbf{I}_d)\mathbf{p} = \mathbf{0}\} && \text{(No Interaction Effect) .}
\end{aligned}
$$

Here, the $a = 2$ levels of the factor Sex are 1 (male) and 2 (female), and the $b = 3$ levels of the factor Language are 1 (English), 2 (Spanish), and 3 (Others). The dimension of the measurements is $d = 2$ (household income and educational status) and $\mathbf{p} = (\mathbf{p}'_{11}, \ldots, \mathbf{p}'_{23})'$. In fact, the multivariate bootstrap procedures both lead to highly significant p values (all < 0.0001) for the two main as well as the interaction effect. Note, that here and in what follows only p values based on the wild bootstrap approach are reported as both bootstrap approaches yielded the same results.

Since the test for the interaction hypothesis yields a significant result, we continue by analyzing male and female participants separately. In order to further interpret the results, we also apply the post hoc comparisons described in Sect. 4. In particular, since the global null hypotheses $\{\mathbf{p}_{i_a1} = \mathbf{p}_{i_a2} = \mathbf{p}_{i_a3}\} = \{\mathbf{P}_3 \otimes \mathbf{I}_2(\mathbf{p}'_{i_a1}, \mathbf{p}'_{i_a2}, \mathbf{p}'_{i_a3})' = \mathbf{0}\}$, $i_a = 1, 2$, are rejected in both groups, we continue with the pairwise comparisons of the languages by testing the hypotheses $\{\mathbf{p}_{i_a i_{b,1}} = \mathbf{p}_{i_a i_{b,2}}\} = \{\mathbf{P}_2 \otimes \mathbf{I}_2(\mathbf{p}'_{i_a i_{b,1}}, \mathbf{p}'_{i_a i_{b,2}})' = \mathbf{0}\}$, $i_a = 1, 2,\ 1 \le i_{b,1} < i_{b,2} \le 3$. Since again all results are significant at the 5% level, we finally consider the univariate outcomes. The results are displayed in Table 7. This reveals some interesting aspects of the data and demonstrates the power of the multivariate approach. For example, the significant difference between 'English' and 'Other' in the male group cannot be detected when considering income and education separately, i.e., a simple univariate analysis would not reveal any difference. This shows a clear advantage of the multivariate approach, which—additionally to the effect contributions of each response—also considers their correlation, thus being able to take advantage of the information added by each response. Furthermore, the significant difference between 'Spanish' vs. 'Other' in the female group is driven by education, while income does not have a significant effect here.

In order to make these methods easily available for users, both bootstrap procedures have been implemented by Sarah Friedrich in an R package **rankMANOVA**, which is available from GitHub (https://github.com/smn74/rankMANOVA).

All analyses discussed in this section can be conducted with **rankMANOVA** by splitting the data accordingly. The implementation of a routine for these post hoc calculations is part of future research.

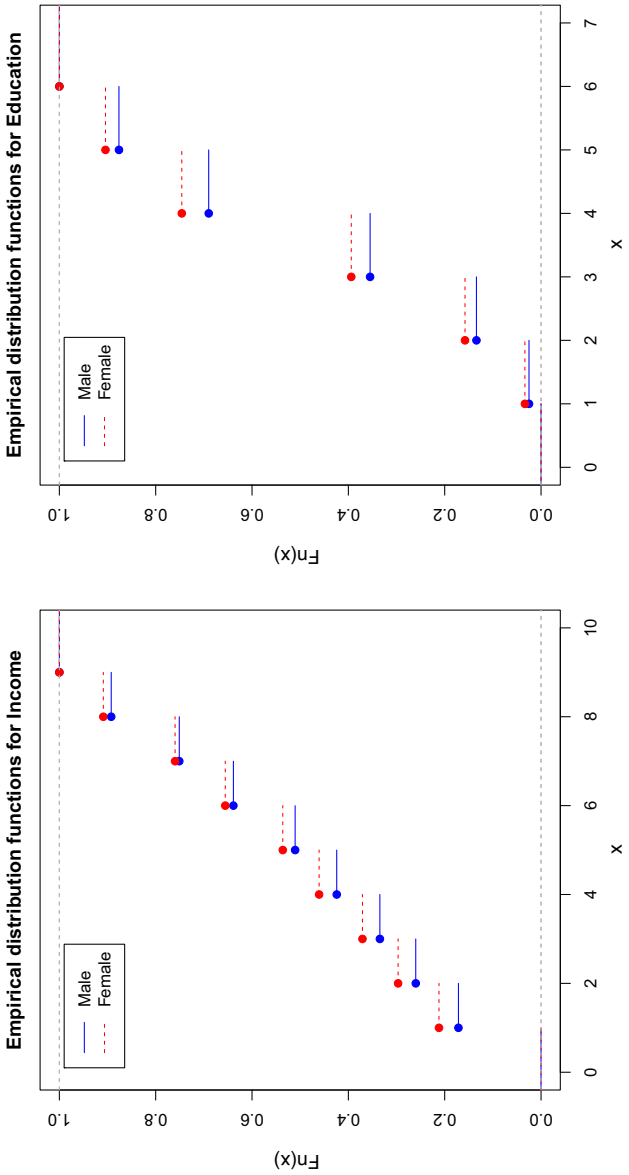

Fig. 3 Empirical distribution functions for male and female participants in the dimensions income and education, respectively

Table 6 Estimated nonparametric effects for the two dimensions income and education

Sex	Language	Income	Education
Male	English	0.586	0.604
	Spanish	0.464	0.405
	Other	0.529	0.559
Female	English	0.561	0.568
	Spanish	0.403	0.366
	Other	0.458	0.498

Table 7 p values for pairwise comparisons of the different groups based on the multivariate (first p value column) and separate univariate (last two columns) approaches

Sex	Language	p value		
		Multivariate	Income	Education
Male	Global null hypothesis	< 0.0001	< 0.0001	< 0.0001
	English versus Spanish	< 0.0001	< 0.0001	< 0.0001
	English versus Other	0.045	0.059	0.078
	Spanish versus Other	< 0.0001	0.044	< 0.0001
Female	Global null hypothesis	< 0.0001	< 0.0001	< 0.0001
	English versus Spanish	< 0.0001	< 0.0001	< 0.0001
	English versus Other	< 0.0001	< 0.0001	0.015
	Spanish versus Other	< 0.0001	0.076	< 0.0001

7 Conclusions and discussion

We have considered an extension of the unweighted treatment effects recently proposed by Brunner et al. (2017) to general multivariate data. These effects do not depend on the sample sizes and allow for transitive ordering. We have rigorously analyzed the asymptotic behavior of the vector of unweighted treatment effects $\widehat{\mathbf{p}}$ and proposed two bootstrap approaches to derive data-driven critical values for global and multiple test decisions: a wild and a group-wise bootstrap. We proved their asymptotic validity using empirical process arguments and analyzed their behavior in a large simulation study, where we considered continuous and ordinal distributions with different covariance settings and sample sizes. In the special situation of the multivariate nonparametric Behrens–Fisher problem we additionally compared both methods to the well-established Brunner et al. (2002) test.

The results were diverse and showed no clear overall preference for any method. This was even true for the Behrens-Fisher problem: In homoscedastic settings for metric or ordinal data, the group-wise bootstrap was too liberal for smaller sample sizes $N \leq 60$ while the wild bootstrap and the Brunner et al. (2002) approach showed better control of the type-I error with a partially too conservative behavior of the latter. In heteroscedastic settings, however, the group-wise bootstrap showed the best results while the other two methods were extremely conservative; even for sample sizes $N \in \{100, 150\}$. In a more general 2×2 design our simulations again indicated

a conservative behavior of the wild bootstrap method for metric data while the group-wise bootstrap kept the nominal level satisfactory. Only in case of small sample sizes ($N \leq 30$) and ordinal data a slight liberality was observed. Here, the wild bootstrap method showed better results. Judging from these findings and its slight theoretical advantage (due to incorporating the $o_p(1)$-term in (8) in the resampling approach), the group-wise bootstrap can be recommended for studies with larger sample sizes ($N \geq 100$) and heteroscedastic situations. In all others we recommend the wild bootstrap.

In future work we will consider extensions of the present setup to censored multivariate data as well as address the question "Which resampling method remains valid and performs preferably?". Here, a challenge will be the correct treatment of ties: The wild bootstrap ceases to reproduce the correct limit distribution in case of right-censored and tied data if it is not adjusted accordingly (Dobler 2017). On the other hand, Akritas (1986) has verified that Efron's bootstrap for right-censored data (Efron 1981) still works in the presence of ties. The planned future paper may also be considered an extension of the article by Dobler and Pauly (2018) to the multi-sample and multivariate case.

Acknowledgements This work was supported by the German Research Foundation (Grant No. PA-2409 4-1).

Appendix

A Proofs

Throughout, let $\mathbb{P}_{1,n_1}, \ldots, \mathbb{P}_{a,n_a}$ be the empirical processes based on the samples $\mathcal{X}_1, \ldots, \mathcal{X}_a$, respectively, which are indexed by the class of functions $\mathcal{G} = \mathcal{F} \circ \Pi$, where

$$\mathcal{F} = \{\mathbb{1}_{(-\infty,x]}(\cdot), \mathbb{1}_{(-\infty,x)}(\cdot) : x \in \mathbb{R}\},$$

and $\Pi = \{\pi_j : j = 1, \ldots, d\}$ is the class of all canonical coordinate projections $\pi_j : \mathbb{R}^d \to \mathbb{R}, (x_1, \ldots, x_d) \mapsto x_j$. Using this indexation, it is easily possible to derive the normalized empirical distribution functions $\widehat{F}_{ij}$ from $\mathbb{P}_{i,n_i}$. In particular, $\widehat{F}_{ij}(x) = \mathbb{P}_{i,n_i}[\frac{1}{2}(\mathbb{1}_{(-\infty,x]}(\cdot) + \mathbb{1}_{(-\infty,x)}) \circ \pi_j]$, where we used the definition $Pf = \int f \mathrm{d}P$ for a suitable function $f \in \mathcal{G}$ and a probability measure P. We also see that every group-specific empirical process $\mathbb{P}_{i,n_i}$ can be considered as an element of $\ell^\infty(\mathcal{G})$ which contains all bounded sequences with indices in $\mathcal{G}$: Based on the definition $\mathbb{P}_{i,n_i} \in \ell^\infty(\mathcal{G})$ if $\sup_{f\in\mathcal{G}} |Pf| < \infty$.

It is important to note that the class $\mathcal{G}$ is obtained from the Vapnik–Červonenkis subgraph class $\mathcal{F}$ concatenated with the class of all canonical coordinate projections Π. This conserves the Vapnik–Červonenkis subgraph property as argued in Lemma 2.6.17(iii) and 2.6.18(vii) of van der Vaart and Wellner (1996). Hence, $\mathcal{G}$ is a Donsker class.

Proof of Theorem 1 Clearly, $\widehat{\mathbf{p}}$ is a multivariate version ϕ of the Wilcoxon functional $\tilde{\phi}(f, g) = \int f(u)\mathrm{d}g(u)$ which is applied to all of the normalized empirical distribu-

tion functions $\widehat{F}_{ij}(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c(x - X_{ijk})$. The Hadamard differentiability of the Wilcoxon functional $\tilde{\phi}$ for normalized distribution functions has been argued in the proof of Theorem 2.1 in Dobler and Pauly (2018), and a similar result for the multivariate ϕ follows immediately. Hence, asymptotic normality follows from an application of the functional delta-method (Theorem 3.9.4 in van der Vaart and Wellner, 1996): it follows that $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$ is asymptotically equal to $\phi'_{F_{11},\ldots,F_{ad}}(\sqrt{N}(\widehat{F}_{ij} - F_{ij})_{i,j})$, where $\phi'_{F_{11},\ldots,F_{ad}}$ is a continuous and linear functional. Hence, the Donsker theorem yields that $\sqrt{N}(\widehat{F}_{ij} - F_{ij})_{i,j}$ converges in distribution to a Gaussian process as $N \to \infty$ and the application of $\phi'_{F_{11},\ldots,F_{ad}}$ proves the asymptotic multivariate normality of $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$.

The asymptotic covariance structure of the resulting multivariate normal distribution is derived in detail in Section 10 of in the supplement, where the asymptotic linear expansion of $\widehat{\mathbf{p}}$ in all empirical distribution functions is utilized. □

Proof of Theorem 2 Similarly, as argued in the proof of Theorem 1, $\mathbf{p}^*$ is obtained as a Hadamard-differentiable functional of all bootstrapped (normalized) empirical distribution functions $F^*_{ij}(t) = \frac{1}{n_i}\sum_{k=1}^{n_i} c(t - X^*_{ijk})$, $j = 1, \ldots, d$, $i = 1, \ldots, a$. As the conditional central limit theorem holds in outer probability for each bootstrapped empirical distribution function, i.e. for each

$$\sqrt{n_i}(F^*_{ij}(t) - \widehat{F}_{ij}(t)) = \frac{1}{\sqrt{n_i}}\left(\sum_{k=1}^{n_i} c(t - X^*_{ijk}) - \sum_{k=1}^{n_i} c(t - X_{ijk})\right),$$

cf. Theorem 3.6.1 in van der Vaart and Wellner (1996), the convergence is transferred to $\sqrt{N}(\mathbf{p}^* - \widehat{\mathbf{p}})$ by means of the functional delta-method for the bootstrap; cf. Theorem 3.9.11 in van der Vaart and Wellner (1996). □

Proof of Theorem 3 First note that, given $\mathcal{X}$, we have conditional convergence in distribution of $\mathbf{F}^\star_N = \sqrt{N}(F^\star_{11}, F^\star_{12}, \ldots, F^\star_{ad})'$ to an (ad)-variate Brownian bridge process $(\mathbf{U}_t)_{t\in\mathbb{R}}$ in outer probability: indeed, each $F^\star_{ij}$ can be written as

$$\begin{aligned} F^\star_{ij}(x) &= \frac{1}{n_i}\sum_{k=1}^{n_i} \widehat{\varepsilon}_{ijk} = \frac{1}{n_i}\sum_{k=1}^{n_i} D_{ik}\cdot[c(x - X_{ijk}) - \widehat{F}_{ij}(x)] \\ &= \frac{1}{n_i}\sum_{k=1}^{n_i}(D_{ik} - \bar{D}_{i\cdot})\cdot c(x - X_{ijk}) \end{aligned}$$

which is due to $\sum_{k=1}^{n_i}[c(x - X_{ijk}) - \widehat{F}_{ij}(x)] = 0$. Here we let $\bar{D}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} D_{ik}$. If we further define $\delta_{\mathbf{X}_{ik}}$ to be the Dirac measure in $\mathbf{X}_{ik}$ and $f_{j,x} = \frac{1}{2}(\mathbb{1}_{(-\infty,x]}(\cdot) + \mathbb{1}_{(-\infty,x)})\circ\pi_j$, we see that $c(x - X_{ijk}) = \delta_{\mathbf{X}_{ik}} f_{j,x}$. Consequently, $F^\star_{ij}(x)$ is a marginal distribution of $\frac{1}{n_i}\sum_{k=1}^{n_i}(D_{ik} - \bar{D}_{i\cdot})\cdot\delta_{\mathbf{X}_{ik}}$. This proves that $F^\star_{ij}(x)$ has the required structure for an application of the conditional Donsker Theorem 3.6.13 in van der Vaart and Wellner (1996). Finally, Example 3.6.12 shows that this Donsker theorem still holds for our choice of wild bootstrap multipliers D_{ik}.

Next, recall the asymptotic linear representation (8) of

$$\sqrt{N}(\widehat{p}_{ij} - p_{ij}) = \sqrt{N}\int (\widehat{G}_j - G_j)\mathrm{d}F_{ij} - \sqrt{N}\int (\widehat{F}_{ij} - F_{ij})\mathrm{d}G_j + o_p(1)$$

which followed from the functional delta-method and which motivated the wild bootstrap version (9). That presentation motivates that the Hadamard-derivative of the multivariate Wilcoxon-type functional ϕ which depends on unknown quantities should be estimated by

$$\phi'_{ij;\widehat{F}} : (\ell^\infty(\mathcal{G}))^a \to \mathbb{R}, \quad (\mathrm{P}_1, \ldots, \mathrm{P}_a) \mapsto \int \left(\frac{1}{a}\sum_{\ell=1}^{a} \mathrm{F}_{\ell j}\right) \mathrm{d}\widehat{F}_{ij} - \int \left(\frac{1}{a}\sum_{\ell=1}^{a} \widehat{F}_{\ell j}\right) \mathrm{dF}_{ij}.$$

Here each $\mathrm{P}_\ell \in \ell^\infty(\mathcal{G})$, $\ell = 1, \ldots, a$, is a distribution on $\mathbb{R}^d$ with marginal normalized distribution functions $\mathrm{F}_{\ell j}$.

We apply the extended continuous mapping theorem (Theorem 1.11.1 in van der Vaart and Wellner, 1996) to the (random) functional $\phi'_{\widehat{F}} = (\phi'_{11;\widehat{F}}, \phi'_{12;\widehat{F}}, \ldots, \phi'_{ad;\widehat{F}}) : \ell^\infty(\mathcal{G}) \to \mathbb{R}^{ad}$. Due to the subsequence principle (Lemma 1.9.2 in van der Vaart and Wellner, 1996) convergence in outer probability is equivalent to outer almost sure convergence along subsequences given a realization of $\mathcal{X}$.

The actual requirement for an application of the extended continuous mapping theorem is satisfied as well: note that $\phi'_{\widehat{F}}$ basically consists of integral mappings of the form

$$\psi : D(\mathbb{R}) \times BV_1(\mathbb{R}) \to \mathbb{R}, \quad (f, g) \mapsto \int f(u)\mathrm{d}g(u)$$

where $D(\mathbb{R})$ is the space of right- (or left-)continuous functions on $\mathbb{R}$ with existing left- (or right-)sided limits and $BV_1(\mathbb{R})$ is the subspace of functions with total variation bounded by 1. Lemma 3.9.17 in van der Vaart and Wellner (1996) states that ψ is Hadamard-differentiable, hence continuous. We conclude that for all sequences of functions $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$, which converge to f_0 in $D(\mathbb{R})$ and to g_0 in $BV_1(\mathbb{R})$, respectively, the sequence of functionals $\psi_n : f \mapsto \int f \mathrm{d}g_n$ satisfies $\psi_n(f_n) \to \int f_0 \mathrm{d}g_0$ as $n \to \infty$.

All in all, the extended continuous mapping theorem, combined with the conditional central limit theorem for the wild bootstrapped empirical distribution functions as stated at the beginning of this proof, concludes this proof: $\phi'_{\widehat{F}}(\mathbf{F}^\star_N)$ converges in distribution to $\phi(\mathbf{U})$ for almost every sample $\mathcal{X}$ which is the same limit in distribution as in Theorem 1. Another application of the subsequence principle yields the desired convergence result in outer probability. □

References

Acion, L., Peterson, J. J., Temple, S., Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, *25*(4), 591–602.

Akritas, M. G. (1986). Bootstrapping the Kaplan–Meier estimator. *Journal of the American Statistical Association*, *81*(396), 1032–1038.
Akritas, M. G., Arnold, S. F., Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, *92*(437), 258–265.
Barbiero, A., Ferrari, P. A. (2015). *GenOrd*: *Simulation of discrete random variables with given correlation matrix and marginal distributions*. R package version 1.4.0.
Bathke, A. C., Harrar, S. W., Madden, L. V. (2008). How to compare small multivariate samples using nonparametric tests. *Computational Statistics and Data Analysis*, *52*(11), 4951–4965.
Bathke, A. C., Friedrich, S., Pauly, M., Konietschke, F., Staffen, W., Strobl, N., Höller, Y. (2018). Testing mean differences among groups: Multivariate and repeated measures analysis with minimal assumptions. *Multivariate Behavioral Research*, *53*(3), 348–359.
Brown, B. M., Hettmansperger, T. P. (2002). Kruskal–Wallis, multiple comparisons and Efron dice. *Australian and New Zealand Journal of Statistics*, *44*(4), 427–438.
Brumback, L. C., Pepe, M. S., Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, *25*(4), 575–590.
Brunner, E., Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers*, *42*(1), 1–52.
Brunner, E., Dette, H., Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, *92*(440), 1494–1502.
Brunner, E., Munzel, U., Puri, M. L. (2002). The multivariate nonparametric Behrens–Fisher problem. *Journal of Statistical Planning and Inference*, *108*(1), 37–53.
Brunner, E., Konietschke, F., Pauly, M., Puri, M. L. (2017). Rank-based procedures in factorial designs: Hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *79*(5), 1463–1485.
Brunner, E., Konietschke, F., Bathke, A. C., Pauly, M. (2018). Ranks and pseudo-ranks: Paradoxical results of rank tests. arXiv preprint arXiv:1802.05650.
Davidson, R., Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, *146*(1), 162–169.
De Neve, J., Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, *110*(511), 1276–1283.
Dobler, D. (2017). A discontinuity adjustment for subdistribution function confidence bands applied to right-censored competing risks data. *Electronic Journal of Statistics*, *11*(2), 3673–3702.
Dobler, D., Pauly, M. (2018). Inference for the Mann–Whitney effect for right-censored and tied data. *TEST*, *27*(3), 639–658.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.
Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, *76*(374), 312–319.
Ellis, A. R., Burchett, W. W., Harrar, S. W., Bathke, A. C. (2017). Nonparametric inference for multivariate data: The R package npmv. *Journal of Statistical Software*, *76*(4), 1–18.
Ferrari, P. A., Barbiero, A. (2012). Simulating ordinal data. *Multivariate Behavioral Research*, *47*(4), 566–589.
Friedrich, S., Pauly, M. (2018). Mats: Inference for potentially singular and heteroscedastic manova. *Journal of Multivariate Analysis*, *165*, 166–179.
Friedrich, S., Konietschke, F., Pauly, M. (2017). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics and Data Analysis*, *113*, 38–52.
Gao, X., Alvo, M. (2005). A unified nonparametric approach for unbalanced factorial designs. *Journal of the American Statistical Association*, *100*(471), 926–941.
Gao, X., Alvo, M. (2008). Nonparametric multiple comparison procedures for unbalanced two-way layouts. *Journal of Statistical Planning and Inference*, *138*(12), 3674–3686.
Gao, X., Alvo, M., Chen, J., Li, G. (2008). Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *Journal of Statistical Planning and Inference*, *138*(8), 2574–2591.
Halvorsen, K. B. (2015). *ElemStatLearn*: *Data sets, functions and examples from the book*: *The elements of statistical learning, data mining, inference, and prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman*. R package version 2015.6.26.
Harrar, S. W., Bathke, A. C. (2008). Nonparametric methods for unbalanced multivariate data and many factor levels. *Journal of Multivariate Analysis*, *99*(8), 1635–1664.

Harrar, S. W., Bathke, A. C. (2012). A modified two-factor multivariate analysis of variance: Asymptotics and small sample approximations (and erratum). *Annals of the Institute of Statistical Mathematics*, *64*(1), 135–165.

Kieser, M., Friede, T., Gondan, M. (2013). Assessment of statistical significance and clinical relevance. *Statistics in Medicine*, *32*(10), 1707–1719.

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., Bathke, A. C., Harrar, S. W., Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, *23*(4), 525–540.

Kruskal, W. H., Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Mann, H. B., Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Marcus, R., Eric, P., Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, *63*(3), 655–660.

Munzel, U. (1999). Linear rank score statistics when ties are present. *Statistics and Probability Letters*, *41*(4), 389–395.

Munzel, U., Brunner, E. (2000). Nonparametric methods in multivariate factorial designs. *Journal of Statistical Planning and Inference*, *88*(1), 117–132.

Puri, M. L., Sen, P. K. (1971). *Nonparametric methods in multivariate analysis*. New York: Wiley.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rust, S. W., Filgner, M. A. (1984). A modification of the Kruskal–Wallis statistic for the generalized Behrens–Fisher problem. *Communications in Statistics-Theory and Methods*, *13*(16), 2013–2027.

Ruymgaart, F. H. (1980). A unified approach to the asymptotic distribution theory of certain midrank statistics. In J.-P. Raoult (Ed.), *Statistique non Parametrique Asymptotique*, pp. 1–18. Berlin: Springer.

Thangavelu, K., Brunner, E. (2007). Wilcoxon–Mann–Whitney test for stratified samples and Efron's paradox dice. *Journal of Statistical Planning and Inference, 137*(3), 720–737. Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri.

Thas, O., De Neve, J., Clement, L., Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *74*(4), 623–671.

Umlauft, M., Konietschke, F., Pauly, M. (2017). Rank-based permutation approaches for nonparametric factorial designs. *British Journal of Mathematical and Statistical Psychology*, *70*, 368–390.

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.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), 1261–1295.