Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2022, Vol. 31(1) 105–118

DOI: 10.1177/09622802211046389
journals.sagepub.com/home/smm
SAGE

# Ranking procedures for repeated measures designs with missing data: Estimation, testing and asymptotic theory

**Kerstin Rubarth[1,2], Markus Pauly[3], and Frank Konietschke[1,2]**

**Abstract**

We develop purely nonparametric methods for the analysis of repeated measures designs with missing values. Hypotheses are formulated in terms of purely nonparametric treatment effects. In particular, data can have different shapes even under the null hypothesis and therefore, a solution to the nonparametric Behrens-Fisher problem in repeated measures designs will be presented. Moreover, global testing and multiple contrast test procedures as well as simultaneous confidence intervals for the treatment effects of interest will be developed. All methods can be applied for the analysis of metric, discrete, ordinal, and even binary data in a unified way. Extensive simulation studies indicate a satisfactory control of the nominal type-I error rate, even for small sample sizes and a high amount of missing data (up to 30%). We apply the newly developed methodology to a real data set, demonstrating its application and interpretation.

**Keywords**

Rank statistics, nonparametric methods, relative effect, repeated measurements, missing data

## 1 Introduction

Repeated measures (RM) designs are commonly used in various research areas and especially in biomedicine. In such layouts, subjects (e.g. patients) are observed under different time points or experimental conditions allowing for statistical inference within a longitudinal framework. A special example is given by a paired design in which each subject is observed twice. Even though RM designs might be more efficient and a cost saving alternative to general factorial designs involving independent units only, missing values might occur, which aggravate both the statistical modeling and evaluation tremendously. Besides determining the missing value mechanism (missing completely at random (MCAR), missing at random (MAR) or missing not at random), estimation of treatment effects along with testing hypotheses of interest becomes a challenging part. Up to now, many powerful parametric (mean-based) as well as purely nonparametric (rank-based) statistical methods exist for data evaluations. To name a few, RM analysis of variance (RM-ANOVA), linear mixed models, or generalized estimation equations are well-established parametric tools that can be used to analyze RM with missing data (assuming normality and specific covariance matrices). For a detailed overview we refer to Little and Rubin[1]. Brunner et al.[2] and Domhof et al.[3] propose purely nonparametric rank-based methods, which do not rely on any distributional assumption and can be used for analyzing metric, discrete or even ordinal data in a unified way. While these ranking methods assume MCAR data, Akritas et al.[4] propose a generalized approach for bivariate data that is valid under a mixture of MCAR and MAR observations. The methods are known to be powerful and robust (with respect to data distributional shapes) and, in particular, they are invariant

[1]Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, Berlin, Germany
[2]Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Straße 2, Berlin, Germany
[3]Department of Statistics, TU Dortmund University, Dortmund, Germany

**Corresponding author:**
Frank Konietschke, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, Berlin, Germany.
Email: frank.konietschke@charite.de

under any monotone transformation of the data. Therefore, ranking procedures are often preferred for making statistical inference in ordinal data, in general. In RM designs with missing values, however, the application of existing ranking methods has some disadvantages, which are all motivated from a practical point of view:

1. The procedures can only be used to test global null hypotheses formulated in terms of the distribution functions (i.e. all distributions are identical). Thus, they do not allow for *variance heteroscedasticity* under the null hypothesis. But, allowing for different variances makes the statistical method more flexible and robust to model mis-specifications.
2. Testing the *global null hypothesis* usually does not answer the main research question of the practitioners; inferring linear contrasts in means of the effects of interest to detect local and specific differences is of practical importance (controlling the family wise error rate in the strong sense).
3. The methods cannot be inverted into *confidence intervals* for the treatment effects. Confidence intervals, however, are used to display variability in the data and shall complement any decent statistical analysis. In particular, international regulatory authorities (e.g. international conference on harmonization ICH) require the computation of confidence intervals (see ICH E9 for clinical trials).
4. The procedures proposed by Domhof et al.[3] might not even be *computable*, because the estimator of the proposed variance-covariance matrix is not necessarily positive semidefinite[3].

The present paper aims to improve upon these points and thus foster the applicability of nonparametric methods. As a specific example, we propose a solution to the nonparametric Behrens-Fisher problem in RM designs with missing data. All of the proposed methods use all-available data and are valid under the MCAR mechanism. In particular, to tackle point 4, the paper proposes a positive-semidefinite estimator of the variance-covariance matrix of the rank means, that is consistent under arbitrary (but fixed) alternatives. The estimation of the covariance matrix as proposed by Konietschke et al.[5] cannot be easily transferred to the case of incomplete data. Thus, this estimation problem along with the development of statistical methods is the main focus of this paper. As a side note, we also introduce a new approximation of the distribution of the ANOVA-type statistic of Konietschke et al.[5] via the Greenhouse-Gaisser method, first introduced by Box[6]. The remainder of the paper is organized as follows. A motivating example with real data is given in section "Motivating example". In the next section "Nonparametric statistical model and effects", the statistical model is introduced. Existing ranking methods and their limitations are discussed in section "Existing rank methods and their limitations". Point estimators along with their asymptotic distributions are exemplified in section "Estimators and their asymptotic distribution" followed by positive semidefinite estimation of the variance covariance matrix in section "Estimation of the covariance matrix". Test procedures and confidence intervals are provided in Section "Test statistics". Results of extensive simulation studies are presented in Section "Simulation study", where we exemplify the behavior of the methods in case of MCAR and MAR scenarios. It turns out that the methods are also applicable in MAR scenarios. The paper closes with the evaluation of the illustrative example in section "Analysis of the example" and a discussion about the findings in section "Discussion and conclusions". Technical proofs and additional results of the simulation study can be found in the supplementary material.

## 2 Motivating example

As a motivating example we consider a migraine trial data set, which has already been investigated by Kostecki-Dillon et al.[7], Gao[8], and Konietschke et al.[9]. In total 135 patients were enrolled in a non-drug headache program, which consisted of four consecutive sessions. In each session, the headache severity level was measured on an ordinal scale ranging from 0 to 20. The lower the score, the better the clinical outcome. The objective of this study is to investigate whether the scores change during the four consecutive sessions. Boxplots of the scores are displayed in Figure 1. It can be readily seen that scores decrease (on median) until session 3 and slightly increase in the last session. However, the data set contains a large

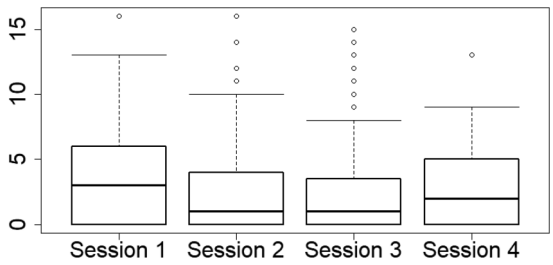


**Figure 1.** Boxplots of the migraine severity level for all four sessions—all available cases.

**Table 1.** Missing pattern in migraine trial, O = observed, M = missing.

| Session 1 | Session 2 | Session 3 | Session 4 | Total | % |
|---|---|---|---|---|---|
| O | O | O | O | 33 | 24.44 |
| O | O | O | M | 48 | 35.56 |
| O | O | M | M | 42 | 31.11 |
| O | M | M | M | 2 | 1.48 |
| M | O | O | O | 2 | 1.48 |
| M | M | O | O | 4 | 2.96 |
| M | M | M | O | 1 | 0.74 |
| M | M | M | M | 2 | 1.48 |
| O | M | O | O | 1 | 0.74 |
| | | | | 135 | 100 |

amount of missing values, out of the 135 patients, only 33 could be observed in each of the four sessions. Table 1 displays the exact missing pattern. Gao[8] performed correlation analyses of the missing proportions versus the headache severity level and concluded that assuming MCAR mechanism is reasonable for the study. Note that the data is measured on an ordinal scale and is highly skewed. Therefore, calculating means and applying mean-based inference methods for analyzing this data set is inappropriate. As a remedy, a purely nonparametric statistical model will be introduced in the next section.

## 3 Nonparametric Statistical Model and Effects

The data example from Section "Motivating example" can be described by $n$ independent and identically distributed $d$-dimensional random vectors

$$\boldsymbol{X}_k = \big((\lambda_{1k}, X_{1k}), \ \ldots, (\lambda_{dk}, X_{dk})\big)', \ k = 1, \ \ldots, n, \ \text{with} \tag{1}$$

$$\lambda_{ik} = \begin{cases} 1, & X_{ik} \text{ is observed} \\ 0, & X_{ik} \text{ is missing,} \end{cases} \tag{2}$$

and marginal distributions $X_{ik} \sim F_i(x)$. The indicators $\lambda_{ik}$ are known constants and used for convenient notation. In order to account for metric, discrete, ordered categorical, and even dichotomous data in a unified way, we use the normalized version of the distribution function

$$F_i(x) = P(X_{ik} < x) + \frac{1}{2}P(X_{ik} = x), \ i = 1, \ \ldots, d; \ k = 1, \ \ldots, n, \tag{3}$$

which is the average of the left- and the right continuous versions $F_i^-(x) = P(X_{i1} < x)$ and $F_i^+(x) = P(X_{i1} \leq x)$ of the distribution function, respectively. The normalized version of the distribution function was first mentioned by Ruymgaart[10] and was later used by several authors developing rank statistics in the case of ties. A detailed overview is provided by Brunner et al.[11].

In model (1), the numbers of non-missing observations under condition $i$ and in total are given by

$$\lambda_i = \sum_{k=1}^{n} \lambda_{ik} \ \text{and} \ N = \sum_{i=1}^{d} \lambda_i, \tag{4}$$

respectively. In order to derive asymptotic results, we consider the following general framework:

$$\textbf{(A1)} \ \lambda_i \to \infty, \ i = 1, \ \ldots, d, \tag{5}$$

$$\textbf{(A2)} \ n \to \infty \ \text{such that} \ \frac{n}{\lambda_i} \leq N_0 < \infty, \ N_0 \ \text{being an arbitrary constant.} \tag{6}$$

These assumptions imply that asymptotic results hold even when the numbers of missing values are bounded, which is the most realistic case. Beyond that, other assumptions as, e.g. specific pattern of missing values, are not required. Model (3), however, does not contain any parameters which could be used to define an appropriate treatment effect. To accomplish this, Domhof et al.[3] propose to use the marginal distributions within the *weighted relative marginal effects*

$$r_{i,N} = \int H_N dF_i = P(Y < X_{i1}) + \frac{1}{2}P(Y = X_{i1}), \ i = 1, \ \ldots, d. \tag{7}$$

Here, $H_N = \frac{1}{N}\sum_{i=1}^{d} \lambda_i F_i(x)$ denotes a *weighted* mean distribution function and $Y \sim H_N$. Since $H_N$ depends on the amount of missing data, $r_{i,N}$ is not a model constant and cannot be used to quantify causal differences between the distributions a priori. Testing hypotheses formulated in terms of the weighted relative effects as well as computing confidence intervals for them would be dubious. Following Konietschke et al.[12] and Brunner et al.[13] for the complete case setting, we therefore propose to use *unweighted* relative effects

$$p_i = \int G dF_i = P(Z < X_{i1}) + \frac{1}{2}P(Z = X_{i1}),\ i = 1,\ \ldots,\ d, \tag{8}$$

see also Umlauft et al.[14]. Here, $G = \frac{1}{d}\sum_{s=1}^{d} F_s$ denotes the unweighted mean distribution function and $Z \sim G$, independent of $X_{i1}$. Thus, $p_i$ relates the distribution $F_i$ *relatively* to the mean distribution $G$ and models whether observations coming from $F_i$ tend to result in larger values than those from $G$. If $p_i < p_j$, then data coming from $F_i$ tend to be smaller than those coming from $F_j$. If $p_i = p_j$, then none of the observations tend to be smaller or larger. *No treatment effect* is therefore indicated as $\boldsymbol{Cp} = \boldsymbol{0}$, where $\boldsymbol{C}$ denotes an appropriate contrast matrix and $\boldsymbol{p} = (p_1,\ \ldots,\ p_d)'$ denotes the vector of all relative marginal effects. Note that the null hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$ implies $H_0^p : \boldsymbol{Cp} = \boldsymbol{0}$, whereas the reverse does not hold in general, as can be easily seen in normal distribution models. Therefore, testing $H_0^p$ is known as the nonparametric Behrens-Fisher problem[15]. We note that Domhof[16] develops univariate confidence intervals for the effects $p_i^* = \frac{1}{d-1}\sum_{s\neq i}\int F_s dF_i$, which, however, might result in paradox conclusions, for example, non-transitive relative effects and we therefore do not follow this approach further. The asymptotic results however, are similar. First, existing rank methods for testing the null hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$ will be discussed.

## 4 Existing rank methods and their limitations

It has been shown in (7) that the weighted relative marginal effect $r_{i,N}$ is a summary measure of the marginal distribution functions $F_i(x)$ and $H_N(x)$. A consistent estimator of $r_{i,N}$ is now obtained by replacing each marginal distribution function $F_i(x)$ and $H_N(x)$ with their empirical counterpart within the integral representation of $r_{i,N}$ in (7). In order to account for possibly missing values, we define the empirical distribution function of the data under condition $i$ as the average of the all-available data by

$$\widehat{F}_{ik}(x) = \begin{cases} c(x - X_{ik}), & \lambda_{ik} = 1 \\ 0, & \lambda_{ik} = 0 \end{cases} \text{ resulting in } \widehat{F}_i(x) = \frac{1}{\lambda_i}\sum_{k=1}^{n} \widehat{F}_{ik}(x). \tag{9}$$

Here, $c(u) = 0,\ 1/2,\ 1$, according as $u <,\ =,\ > 0$, denotes the normalized version of the count function. Furthermore, let $\widehat{H}_N(x) = \frac{1}{N}\sum_{i=1}^{d} \lambda_i \widehat{F}_i(x)$ denote the empirical counterpart of $H_N$ and note that $R_{ik} = \lambda_{ik}(N\widehat{H}_N(X_{ik}) + \frac{1}{2})$ is the mid-rank of $X_{ik}$ among all $N$ observed values. If $\lambda_{ik} = 0$ and thus $X_{ik}$ is missing, $R_{ik} = 0$, for convenience. Plugging-in $\widehat{H}_N$ and $\widehat{F}_i$ into (7) leads to a rank estimator of $r_{i,N}$ as

$$\widehat{r}_{i,N} = \int \widehat{H}_N d\widehat{F}_i = \frac{1}{\lambda_i}\sum_{k=1}^{n} \lambda_{ik}\widehat{H}_N(X_{ik}) = \frac{1}{N}\left(\overline{R}_{i\cdot} - \frac{1}{2}\right). \tag{10}$$

Here, $\overline{R}_{i\cdot} = \lambda_i^{-1}\sum_{k=1}^{n} R_{ik}$ denotes the mean of the ranks under condition $i$. For convenient representation of asymptotic results, the point estimators are collected in the vector $\widehat{\mathbf{r}}_N = (\widehat{r}_{1,N},\ \ldots,\ \widehat{r}_{d,N})'$.

Akritas and Brunner[17] have shown that $\sqrt{n}\boldsymbol{C}\widehat{\mathbf{r}}_N$ follows, asymptotically, as $n \to \infty$, a multivariate normal distribution with expectation $\boldsymbol{0}$ and covariance matrix $\boldsymbol{CV}_n\boldsymbol{C}$ under the special null hypothesis $H_0^F : \boldsymbol{CF} = \boldsymbol{0}$, where $\boldsymbol{V}_n = Cov(\lambda_{11}H_N(X_{11}),\ \ldots,\ \lambda_{d1}H_N(X_{d1}))'$. The asymptotic distribution of the estimators under (any) alternative has not yet been developed. Moreover, since $\boldsymbol{V}_n$ is unknown in practical applications and must be estimated from the data, Domhof et al.[3] propose to use $\widehat{\mathbf{V}}_n = (\hat{v}(i, i'))_{i,i'=1,\ldots,d}$ with

$$\begin{aligned} \hat{v}(i, i) &= \frac{n}{N^2\lambda_i(\lambda_i - 1)}\sum_{k=1}^{n} \lambda_{ik}(R_{ik} - \bar{R}_{i\cdot})^2, \\ \hat{v}(i, i') &= \frac{n}{NK(i, i')}\sum_{k=1}^{n} \lambda_{ik}\lambda_{i'k}(R_{ik} - \bar{R}_{i\cdot})(R_{i'k} - \bar{R}_{i'\cdot}). \end{aligned} \tag{11}$$

Here $K(i, i') = (\lambda_i - 1)(\lambda_{i'} - 1) + \Lambda(i, i') - 1$ with $\Lambda(i, i') = \sum_{k=1}^{n} \lambda_{ik}\lambda_{i'k}$, see Brunner et al.[18] and Domhof et al.[3]. The estimator $\widehat{\mathbf{V}}_n$, however, is not necessarily positive semidefinite in case of missing values and thus might result in a negative

variance estimator of a linear combination of the estimators $\widehat{r}_{i,N}$. In addition, since the distribution of the estimators is only known under the null hypothesis $H_0^F$, confidence intervals for the effects of interest—even without missing values—cannot be computed. In the next section, consistent estimators of the unweighted relative effects $p_i$ along with a positive semidefinite estimator of its variance-covariance matrix will be proposed. The estimator is even consistent under general alternatives formulated in terms of $\boldsymbol{p}$.

## 5 Estimators and their asymptotic distribution

Following the above ideas, a point estimator for $p_i$ in (8) is readily available by replacing each of the unknown distribution functions $F_i(x)$ and $G(x)$ with their empirical counterparts $\widehat{F}_i(x)$ in the integral representation of $p_i$. Plugging-in the empirical counterpart

$$\widehat{G}(x) = \frac{1}{d}\sum_{s=1}^{d}\widehat{F}_s(x)$$

of $G(x)$ in (8) leads to the point esimator

$$\widehat{p}_i = \int \widehat{G}d\widehat{F}_i = \frac{1}{\lambda_i}\sum_{k=1}^{n}\frac{\lambda_{ik}}{d}\sum_{s=1}^{d}\frac{1}{\lambda_s}\sum_{\ell=1}^{n}\lambda_{s\ell}c(X_{ik} - X_{s\ell}). \tag{12}$$

In general, the estimator cannot be computed using ranks of the data; instead it is a sum of indicators (values of the count functions) and therefore its numerical computation is slightly more involved than that of $\widehat{r}_{i,N}$. If no missing values are apparent, then $\widehat{r}_{i,N} = \widehat{p}_i$, $i = 1, \ldots, d$. First its asymptotic properties will be studied in the following proposition.

**Proposition 1.** The estimator $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_d)'$ is asymptotically unbiased and strongly consistent under **(A1)** in (5), i.e.

1. $E(\widehat{\boldsymbol{p}}) = \boldsymbol{p} + \mathcal{O}(\frac{1}{n})$
2. $\widehat{\boldsymbol{p}} - \boldsymbol{p} \overset{\text{a.s.}}{\to} \mathbf{0}$, $\min\{\lambda_1, \ldots, \lambda_d\} \to \infty$.

Next, the asymptotic distribution of the statistic $\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ will be derived. The following Theorem 1 shows that $\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ has, asymptotically, under **(A1)** and **(A2)**, the same distribution as the random vector $\sqrt{n}\mathbf{B} = \sqrt{n}(B_1, \ldots, B_d)'$, whose components are sums of independent random variables:

$$\begin{aligned}
\sqrt{n}B_i &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\Psi_{ik} - E(\Psi_{ik})), \text{ where,}\\
\Psi_{ik} &= \frac{n\lambda_{ik}}{\lambda_i}\left(G(X_{ik}) - \frac{1}{d}F_i(X_{ik})\right) - \frac{1}{d}\sum_{s\neq i}\frac{n\lambda_{sk}}{\lambda_s}F_i(X_{sk}), \text{ and}\\
E(\Psi_{ik}) &= \frac{n\lambda_{ik}}{\lambda_i}\left(p_i - \frac{1}{d}p^{(ii)}\right) - \frac{1}{d}\sum_{s\neq i}\frac{n\lambda_{sk}}{\lambda_s}p^{(is)}.
\end{aligned} \tag{13}$$

Here, $p^{(is)} = \int F_i dF_s$ and $p^{(ii)} = \int F_i dF_i = \frac{1}{2}$ denote pairwise defined relative marginal effects between time points $i$ and $s$ and $i$ and $i$, respectively.

**Theorem 1.** Let $\sqrt{n}\mathbf{B} = \sqrt{n}(B_1, \ldots, B_d)'$ be the vector of the random variables $\sqrt{n}B_i$, $i = 1, \ldots, d$, as defined in (13). If $n \to \infty$ such that **(A1)** and **(A2)** in (5) and (6) hold, then,

$$\left|\left|\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) - \sqrt{n}\mathbf{B}\right|\right|_2^2 = \mathcal{O}\left(\frac{1}{n}\right),$$

with $\|\mathbf{x}\|_2^2$ denoting the $L_2$-norm.

It follows from Theorem 1, that the asymptotic covariance matrix of the linear rank statistic $\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ is given by

$$\mathbf{V}_n = Cov(\sqrt{n}\mathbf{B}). \tag{14}$$

The asymptotic multivariate normality of the linear rank statistic $\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ is given in the next Theorem.

**Theorem 2.** Under the assumptions **(A1)** and **(A2)**, the statistic $\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ follows asymptotically, as $n \to \infty$, a multivariate normal distribution with expectation $\mathbf{0}$ and covariance matrix $\mathbf{V}_n$.

The covariance matrix $\mathbf{V}_n$, however, is unknown in practical applications and must be estimated from the data for making statistical inferences. We will derive a consistent and positive-semidefinite estimator in the next section.

## 6 Estimation of the covariance matrix

If the random variables $\Psi_{ik}$ in (13) were observable, then an estimator of $\mathbf{V}_n$ would be given by

$$\widetilde{\mathbf{V}}_n = \frac{1}{n-1}\sum_{k=1}^{n}\big(\boldsymbol{\Psi}_k - E(\boldsymbol{\Psi}_k)\big)\big(\boldsymbol{\Psi}_k - E(\boldsymbol{\Psi}_k)\big)'.$$

Note that in the definition of $\widetilde{V}_N$ each vector $\boldsymbol{\Psi}_k = (\Psi_{1k}, \ldots, \Psi_{dk})'$ is centered with its own specific expectation and therefore, $\widetilde{\mathbf{V}}_n$ is not the empirical covariance matrix of the vectors $\boldsymbol{\Psi}_k$. However, the variables $\boldsymbol{\Psi}_k$ are not observable and thus, $\widetilde{\mathbf{V}}_n$ cannot be computed in practical applications. Therefore, we replace them with observable random variables, which are 'close enough' to $\boldsymbol{\Psi}_k$ in an appropriate norm. Define the vectors $\widehat{\boldsymbol{\Psi}}_k = (\widehat{\Psi}_{1k}, \ldots, \widehat{\Psi}_{dk})'$, the components of which are the empirical counterparts

$$\widehat{\Psi}_{ik} = \frac{n\lambda_{ik}}{\lambda_i}\left(\widehat{G}(X_{ik}) - \frac{1}{d}\widehat{F}_i(X_{ik})\right) - \frac{1}{d}\sum_{s\neq i}\frac{n\lambda_{sk}}{\lambda_s}\widehat{F}_i(X_{sk}), \text{ and}$$

$$\widehat{\beta}_{ik} = \frac{n\lambda_{ik}}{\lambda_i}\left(\widehat{p}_i - \frac{1}{d}\widehat{p}^{(ii)}\right) - \frac{1}{d}\sum_{s\neq i}\frac{n\lambda_{sk}}{\lambda_s}\widehat{p}^{(is)},$$

with $\widehat{p}^{(is)} = \int \widehat{F}_i d\widehat{F}_s$ and $\widehat{p}^{(ii)} = \int \widehat{F}_i d\widehat{F}_i = \frac{1}{2}$. Furthermore, consider the estimator

$$\widehat{\mathbf{V}}_n = \frac{1}{n-1}\sum_{k=1}^{n}\Big(\widehat{\boldsymbol{\Psi}}_k - \widehat{\boldsymbol{\beta}}_k\Big)\Big(\widehat{\boldsymbol{\Psi}}_k - \widehat{\boldsymbol{\beta}}_k\Big)', \ \widehat{\boldsymbol{\beta}}_k = (\widehat{\beta}_{1k}, \ldots, \widehat{\beta}_{dk})'. \tag{15}$$

Its properties are listed in the next Theorem.

**Theorem 3.** Let $\mathbf{V}_n = Cov(\sqrt{n}\mathbf{B})$ and let $\widehat{\mathbf{V}}_n$ as given in (15). Then,

1. $\widehat{\mathbf{V}}_n$ is positive semidefinite.
2. If $n \to \infty$ such that **(A1)** and **(A2)** hold, then $\widehat{\mathbf{V}}_n - \mathbf{V}_n \xrightarrow{\text{a.s.}} \mathbf{0}$.

## 7 Test statistics

In this section, we introduce different test procedures (global and multiple) to test the null hypothesis $H_0^p : \mathbf{C}\boldsymbol{p} = \mathbf{0}$. First, quadratic test procedures will be explained and a linear multiple contrast test procedure (MCTP) will be introduced afterwards. One advantage of the MCTP over the quadratic test procedures is, that it can be used for testing multiple hypotheses as well as for constructing simultaneous confidence intervals for the relative marginal effects $p_i$ and linear combinations thereof. In particular, the adjusted $p$-values and the corresponding confidence intervals are compatible, that is, it is not possible that the null hypothesis is rejected by the multiple comparison procedure, but the corresponding confidence interval includes $\frac{1}{2}$ - the hypothetical value of no treatment under $H_0^p$.

### 7.1 Quadratic tests

Domhof et al.[3] introduce two different types of test statistics to infer the global null hypothesis $H_0^F : \mathbf{CF} = \mathbf{0}$ formulated in terms of the distribution functions. Since both of the methods only depend on the vector of point estimators, their estimated variance-covariance and hypothesis (contrast) matrices, they can be generalized to test $H_0^p : \mathbf{Cp} = \mathbf{0}$ using the newly developed estimators. Consider the Wald-type statistic (WTS)

$$Q_n = n\widehat{\mathbf{p}}'\boldsymbol{C}'[\boldsymbol{C}\widehat{\boldsymbol{V}}_n\boldsymbol{C}']^{+}\boldsymbol{C}\widehat{\mathbf{p}}, \tag{16}$$

which is a quadratic form in the point estimators $\widehat{\mathbf{p}}$. Under $H_0^p$, the distribution of $Q_n$ can be approximated by a $\chi^2_{\widehat{f}}$ distribution with $\widehat{f} = rank(\boldsymbol{C}\widehat{\boldsymbol{V}}_n\boldsymbol{C}')$ degrees of freedom. Here, $[\,\cdot\,]^+$ denotes the Moore-Penrose inverse of a matrix. Simulation studies indicate, however, that the test upon $Q_n$ behaves liberal and highly over-rejects the null hypothesis when sample sizes are small. Therefore, Domhof et al.[3] propose to approximate its distribution by an F-distribution resulting in the so-called ANOVA-type statistic. Let $\boldsymbol{M} = \boldsymbol{C}'[\boldsymbol{C}\boldsymbol{C}']^-\boldsymbol{C}$ be a projection matrix and let $[\boldsymbol{C}\boldsymbol{C}']^-$ be a generalized inverse of $\boldsymbol{C}\boldsymbol{C}'$. Then, the null hypotheses $H_0^p(\boldsymbol{C}): \boldsymbol{C}\boldsymbol{p} = \boldsymbol{0}$ and $H_0^p(\boldsymbol{M}): \boldsymbol{M}\boldsymbol{p} = \boldsymbol{0}$ are equivalent. It holds under $H_0^p(\boldsymbol{M})$ that the distribution of the ATS

$$A_n = \frac{ntr(\boldsymbol{M}\widehat{\boldsymbol{V}}_n)}{tr(\boldsymbol{M}\widehat{\boldsymbol{V}}_n\boldsymbol{M}\widehat{\boldsymbol{V}}_n)}\widehat{\mathbf{p}}'\boldsymbol{M}\widehat{\mathbf{p}} \tag{17}$$

can be approximated by a $\chi^2_{\widehat{f}}$ distribution with

$$\widehat{f} = \frac{[tr(\boldsymbol{M}\widehat{\boldsymbol{V}}_n)]^2}{tr(\boldsymbol{M}\widehat{\boldsymbol{V}}_n\boldsymbol{M}\widehat{\boldsymbol{V}}_n)} \tag{18}$$

degrees of freedom. Here, $tr(\mathbf{A})$ denotes the trace of the matrix $\mathbf{A}$. However, simulation studies indicate a liberal behavior of the ATS in some settings[5]. We therefore propose to apply the Greenhouse-Gaisser method introduced by Box[6] resulting in

$$A_{n,2} = \frac{n\widehat{\mathbf{p}}'\boldsymbol{M}\widehat{\mathbf{p}}}{tr(\boldsymbol{M}\widehat{\boldsymbol{V}}_n)} \tag{19}$$

which can be approximated by an $\mathcal{F}_{\widehat{f}_1,\widehat{f}_2}$ distribution with

$$\widehat{f}_1 = \widehat{f} \text{ and } \widehat{f}_2 = (n-1)\widehat{f} \tag{20}$$

degrees of freedom. Note that we will refer to (17) as the first version of the ATS or ATS (1) and to (19) as the second version of the ATS or ATS (2) for convenience. Next, MCTPs along with simultaneous confidence intervals will be introduced.

## 7.2 Multiple contrast test procedure

As already outlined above, the quadratic test procedures can only be used to test the global null hypothesis of no effect. Therefore, Konietschke et al.[5] propose a MCTP for making local statistical inference. In order to test the individual null hypothesis $H_0^{(\ell)}: \mathbf{c}_\ell'\mathbf{p} = 0$, consider the test statistic

$$T_\ell = \sqrt{n}\frac{\mathbf{c}_\ell'(\widehat{\mathbf{p}} - \mathbf{p})}{\sqrt{\mathbf{c}_\ell'\widehat{\boldsymbol{V}}_n\mathbf{c}_\ell}}.$$

Here, $\mathbf{c}_\ell'$ denotes the $\ell$th row vector of $\mathbf{C}$. Even though the exact distribution of $T_\ell$ for finite sample size $n$ remains unknown, it is asymptotically standard normal. Furthermore, the test statistics $T_\ell$ and $T_m$ are not necessarily independent due to the chosen contrasts $\mathbf{c}_\ell'$ and $\mathbf{c}_m'$ ($\ell \neq m$) and/or the RM data. In order to take the correlations across the $q$ test statistics within the multiplicity adjustment into account, we collect them in the vector

$$\mathbf{T} = \left(T_1, \ldots, T_q\right)'.$$

It follows from Theorem 2 and Slutzky's Theorem, that $\mathbf{T}$ follows, asymptotically, a multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{C}\mathbf{V}_n\mathbf{C}'\mathbf{D}^{-1/2},$$

where $\mathbf{D}$ is a diagonal matrix of the diagonal elements of $\mathbf{C}\mathbf{V}_n\mathbf{C}'$. For large sample sizes, the individual null hypothesis $H_0^{(\ell)}: \mathbf{c}_\ell'\mathbf{p} = 0$ will be rejected at multiple level $\alpha$, that is, the procedure controls the familywise type-I error rate, if

$$|T_\ell| \geq z_{1-\alpha,2,\boldsymbol{R}}, \tag{21}$$

where $z_{1-\alpha,2,\boldsymbol{R}}$ denotes the two-sided $(1-\alpha)$-equicoordinate quantile of the $N(\mathbf{0}, \boldsymbol{R})$ distribution[12]. The two-sided $(1-\alpha)$-equicoordinate quantile satisfies the condition $P(|T_1| < z_{1-\alpha,2,\boldsymbol{R}}, \ldots, |T_q| < z_{1-\alpha,2,\boldsymbol{R}}) = 1-\alpha$ for $(T_1, \ldots, T_q)' \sim N(\mathbf{0}, \boldsymbol{R})$. Compatible $(1-\alpha)$-simultaneous confidence intervals for the effects $\delta_\ell = \mathbf{c}_\ell'\mathbf{p}$ are given by

$$CI_\ell = \left[\mathbf{c}_\ell'\widehat{\mathbf{p}} \mp \frac{z_{1-\alpha,2,\boldsymbol{R}}}{\sqrt{n}}\sqrt{\mathbf{c}_\ell'\widehat{\boldsymbol{V}}_n\mathbf{c}_\ell}\right]. \tag{22}$$

We refer to Bretz et al.[19] for the numerical derivation of the equicoordinate quantile. Finally, for large sample sizes, the global null hypothesis $H_0^p : \boldsymbol{Cp} = \mathbf{0}$ will be rejected at two-sided multiple level $\alpha$, if

$$T_0 = \max\{|T_1|, \ldots, |T_q|\} \geq z_{1-\alpha,2,\boldsymbol{R}}. \tag{23}$$

The correlation matrix, however, is unknown in practical applications. We recommend to replace $\mathbf{R}$ with its consistent estimator

$$\widehat{\mathbf{R}}_n = \widehat{\mathbf{D}}^{-1/2}\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{C}'\widehat{\mathbf{D}}^{-1/2},$$

in (21), (22), and (23), respectively. Here, $\widehat{\mathbf{D}}$ denotes the diagonal matrix obtained from the diagonal elements of $\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{C}$. For small sample sizes, we follow Konietschke et al.[5], who proposed in the case of complete observations to approximate the distribution of $\mathbf{T}$ by a multivariate central $t_{n-1}(\mathbf{0}, \widehat{\boldsymbol{R}}_n)$ distribution. It follows from sections "Estimators and their asymptotic distribution" and "Estimation of the covariance matrix" that both choices lead to asymptotic correct multiple contrast tests.

## 8 Simulation study

All of the procedures developed in the previous sections are of asymptotic nature and thus, investigating their finite sample behavior with respect to their control of the type-I error rate (at nominal 5% level) and power to detect alternatives within extensive simulation studies is mandatory. We considered all introduced tests, namely

1. the WTS in (16) with a critical value from a $\chi^2_{\widehat{f}}$ distribution,
2. the ATS (1) in (17) with the proposed $F$-approximation,
3. the ATS (2) in (19) with the proposed $F$-approximation,
4. the MCTP $T_0$ in (23) with a $t_{n-1}(\mathbf{0}, \widehat{\boldsymbol{R}}_n)$ approximation,

and compared them with

5. the WTS and ATS for testing $H_0^F$ as proposed by Domhof et al.[3]

in different homo- and heteroscedastic RM designs with different rates of missing values. Even though Domhof et al.[3] reported a liberal behavior of the WTS (for testing $H_0^F$), we added the method as a competing procedure for completeness. We thus also investigated their robustness to variance heteroscedasticity. Since all of the methods above use all-available data, we additionally compared them with two MCTP-based approaches: a complete case analysis and a naive imputation approach, in which we either

6. deleted the whole observation vector $\boldsymbol{X}_k$ of subject $k$ if any $X_{ik}$ was missing ($\lambda_{ik} = 0$), or
7. if $X_{ik}$ was missing ($\lambda_{ik} = 0$), we calculated $median(\lambda_{i1}X_{i1}, \ldots, \lambda_{in}X_{in})$, and assigned it to $X_{ik}$ and set $\lambda_{ik} = 1$.

Data have been generated using discretized, by rounding to integers, normal and log-normal distributions with varying numbers of time points $d \in \{3, 4\}$, sample sizes $n \in \{20, 30, 50\}$, amount of missing values $r \in \{0\%, 10\%, 30\%\}$, and six different types of covariance matrices

$$\begin{aligned}
\boldsymbol{\Sigma}_1 &= \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}, \\
\boldsymbol{\Sigma}_3 &= \begin{pmatrix} 1 & 0.3 & 0.6 \\ 0.3 & 1.2 & 0.9 \\ 0.6 & 0.9 & 1.5 \end{pmatrix}, \boldsymbol{\Sigma}_4 = \begin{pmatrix} 1 & 0.2 & 0.4 & 0.6 \\ 0.2 & 2 & 0.7 & 0.5 \\ 0.4 & 0.7 & 2.5 & 0.6 \\ 0.6 & 0.5 & 0.6 & 3 \end{pmatrix}, \\
\boldsymbol{\Sigma}_5 &= \begin{pmatrix} 1 & 0.6 & 0.36 & 0.216 \\ 0.6 & 1 & 0.6 & 0.36 \\ 0.36 & 0.6 & 1 & 0.6 \\ 0.216 & 0.36 & 0.6 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_6 = \begin{pmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 0.8 & 1.5 & 0.8 & 0.64 \\ 0.64 & 0.8 & 2 & 0.8 \\ 0.512 & 0.64 & 0.8 & 2.5 \end{pmatrix}.
\end{aligned} \tag{24}$$

The covariance matrices were chosen to model a broad selection of dependency patterns, including homoscedastic ($\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$) as well as heteroscedastic marginals. Note that $H_0^F$ holds only under $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. We furthermore investigated the

methods' sensitivity to both MCAR and MAR data to cover realistic scenarios. In order to generate the former, we multiplied the observations with randomly chosen indicators $\lambda_{ik} \sim B(1-r)$, with a zero entry being interpreted as a missing observation, whereas we followed Santos et. al.[20] for the latter. Hereby we defined pairs of observations $\{X_{obs}, X_{miss}\}$, where $X_{obs}$ determines the probability that $X_{miss}$ was actually observed. For instance, in case of $d = 4$ we defined the pairs $\{X_{1k}, X_{2k}\}$ and $\{X_{3k}, X_{4k}\}$. Following the idea of Amro et al.[21], we investigated two different types of MAR scenarios, MAR (1) and MAR (2). First, for the MAR (1) scenario, we divided $X_{i,obs}$ into three groups: (1) $\{\boldsymbol{X}_{ik} = X_{i,obs} \in (-\infty, -\sigma_i), k = 1, \ldots, n\}$, (2) $\{\boldsymbol{X}_{ik} = X_{i,obs} \in (-\sigma_i, \sigma_i), k = 1, \ldots, n\}$, and (3) $\{\boldsymbol{X}_{ik} = X_{i,obs} \in (\sigma_i, \infty), k = 1, \ldots, n\}$, where $\sigma_i^2$ is the variance of $X_{i,obs}$. Then, we assigned a missing rate of 10% to the first and third group and a missing rate of 30% to the second group. Second, in the MAR (2) scenario, data was divided into two groups using the median, following the idea of Zhu et al[22]. Specifically, we defined (1) $\{\boldsymbol{X}_{ik} = X_{i,obs} \in (-\infty, median(X_{i,obs}), k = 1, \ldots, n\}$ and (2) $\{\boldsymbol{X}_{ik} = X_{i,obs} \in (median(X_{i,obs}), \infty), k = 1, \ldots, n\}$. Here, we assigned a missing rate of 0% to the first group and a missing rate of 10% to the second group.

For each design, 10,000 simulation runs were performed using the *R* software package of statistical computing, version R 3.6.4[23]. The complete simulation code is available on https://github.com/KerstinRubarth/RM_Miss.

First, we will discuss simulation results when $H_0^F$ holds and data is MCAR (Table S. 1). It appears that the ATS for testing $H_0^F$ tends to be slightly liberal when samples are rather small ($n = 20$). Otherwise, it exhibits an almost accurate type-I error control. In fact, the amount of missing values impact its behavior only by slightly increasing or decreasing the type-I error rate. A similar behavior of the newly developed ATS (1) for testing $H_0^p$ can be detected. The ATS (2) for testing $H_0^p$ controls the type-I error even more accurate while sometimes being slightly conservative. Moreover, the new MCTP tends to be quite

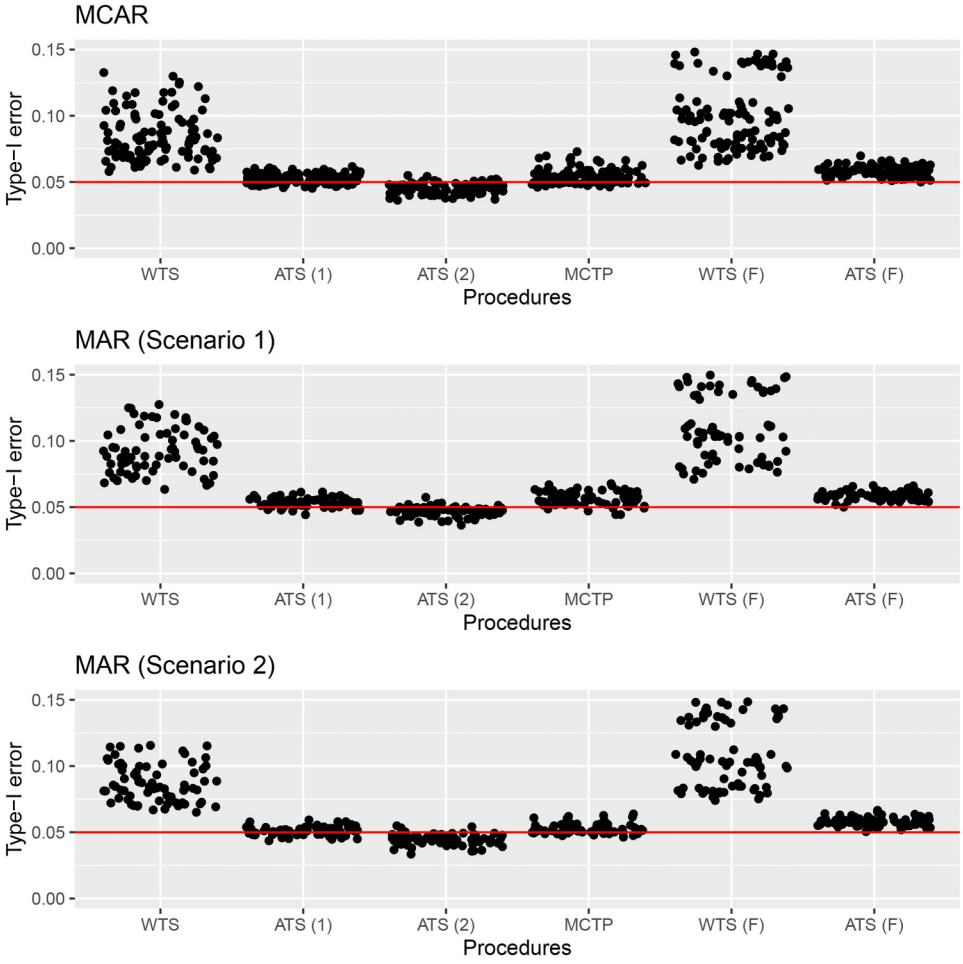


**Figure 2.** Type-I error rates of the newly proposed Wald- (WTS), ANOVA-type (ATS (1) and ATS (2)) and MCT procedures and the procedures of Domhof et al.[3] under the three different missing mechanisms MCAR, MAR (1) and MAR (2). MCT: multiple contrast test; MCAR: missing completely at random; MAR: missing at random.

accurate in most settings and gets slightly liberal if the probability of missingness increases. Contrary, both WTS methods are too liberal for all settings and cannot be recommended. Overall, the simulations for $H_0^F$ indicate that both versions of the ATS and the MCTP for testing $H_0^p$ control the type-I error rate quite accurately when $n \geq 20$.

Next, we will explore the methods' behavior under variance heteroscedasticity (Table S. 2 and Table S. 3). It turns out that the methods tend to be quite accurate in most scenarios and tend to over-reject the null hypothesis when samples are small and missing probabilities are high. However, some scenarios also indicate a fairly robust behavior of the methods for testing $H_0^F$. Overall, the methods seem not be too sensitive towards variance heteroscedasticity. In general, the procedures work equally well in case of normally and log-normally distributed data. However, in most scenarios, both versions of the ATS for testing $H_0^p$ control the type-I error more accurate than the ATS for testing $H_0^F$. As already pointed out, both WTS procedures do not control the type-I-error rate in case of smaller sample sizes. Again, for small sample sizes, e.g. $n = 20$, the ATS (2) under $H_0^p$ performs better than the MCTP. However, all test statistics show a conservative behavior in case of high correlations and small sample sizes. This was already mentioned in Konietschke et al.[5] and Friedrich et al.[24] for the RM design without missing data and Munzel[25] as well as Harrar et al.[26] and Amro et al.[27] for the paired two sample case. Overall, the simulation studies indicate that the newly developed methods based upon the ATS, especially the second version, and (to some extend) the MCTP control the type-I error well when $n \geq 20$. Next, we compare the results of the type-I error simulation for data under MCAR and MAR mechanisms. A graphical overview of the type-I error rates of the procedures for testing $H_0^F$ and $H_0^p$ in various settings can be found in Figure 2. Since the missing rates in the MCAR and MAR scenarios in these simulations were similiar and no difference in terms of the type-I error rates is apparent, the procedures seem to be fairly robust to the missing value mechanism, though the theory was developed under the MCAR assumption.

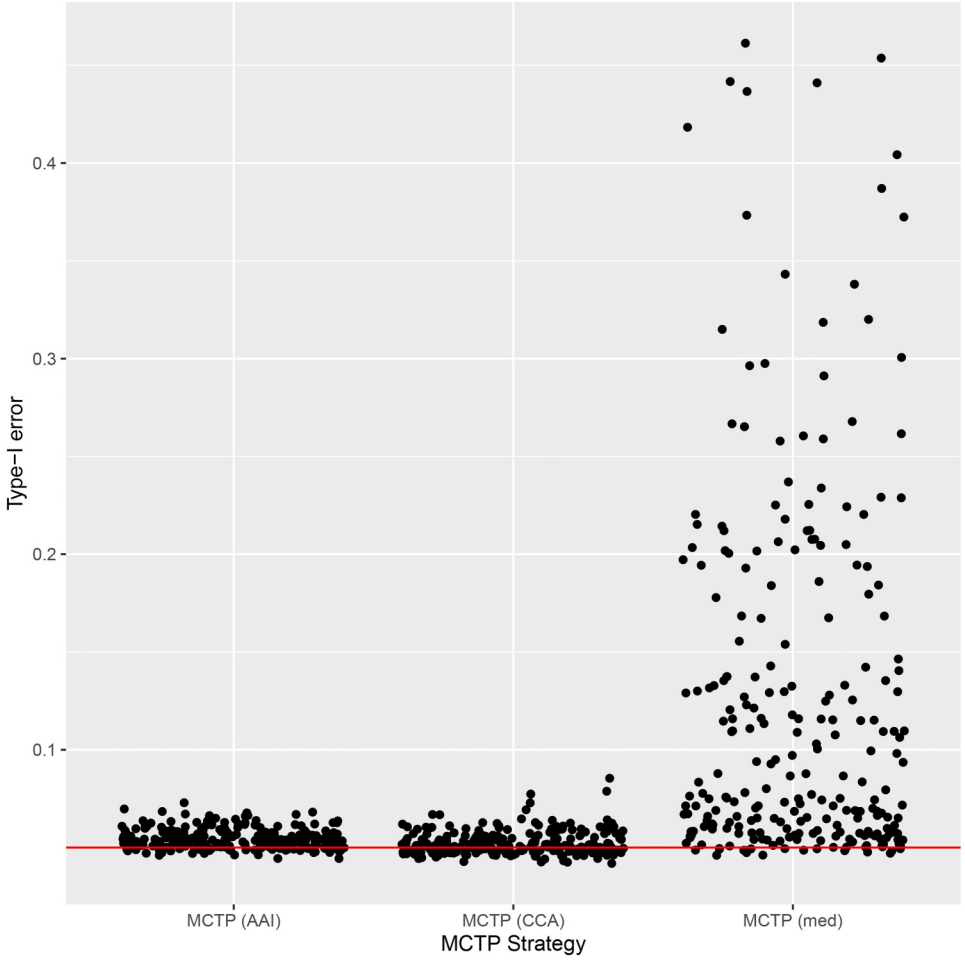


**Figure 3.** Type-I error rates of the MCTP, either using AAI, only complete cases and an imputed data set, using the median for each repeated measurement. MCTP: multiple contrast test procedure; AAI: all-available information; CCA: Complete Case Analysis; med: median.

Since we advocate to use the MCTP as it additionally allows the simultaneous testing of the corresponding single contrast hypotheses, we want to compare its behavior in terms of type-I-error rates in comparison to two "naive" procedures for handling missing data: median imputation and complete case analysis. A graphical presentation of the results can be found in Figure 3. The results for the newly proposed MCTP which uses all-available information and the MCTP using only complete cases are comparable. As expected (van Buuren[28], Ramosaj et al.[29]), the simple median imputation yields in some scenarios with many missings extremly inflated type-I error rates and is therefore not recommend. Note that data has been generated under MCAR and MAR assumptions for this comparison.

In order to investigate the power of the procedures, a simulation study was conducted using four-dimensional normal and log-normal distributions with $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and covariance matrices $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_4$, $\boldsymbol{\Sigma}_5$, $\boldsymbol{\Sigma}_6$. In particular, three different types of shift-alternatives were considered

$$\begin{aligned} &\textbf{Alternative 1} \\ &\boldsymbol{\mu} = (0, 0, 0, \delta)' \\ &\textbf{Alternative 2} \\ &\boldsymbol{\mu} = (0, 0, \delta, \delta)' \\ &\textbf{Alternative 3} \\ &\boldsymbol{\mu} = (0, 1\delta, 2\delta, 3\delta)', \end{aligned} \tag{25}$$

with ranging $\delta = (0.2, 0.4, 0.6, 0.8, 1, 1.5)$ and different amount of missing values. As the WTS turned out to be inappropriate for small sample sizes, it was not included into the power analysis. Moreover, since the second version of the ATS for testing $H_0^p$ showed a more accurate behavior than the first version, we only present results for the second version. The results for covariance matrix $\boldsymbol{\Sigma}_4$ and $n = 30$ under the MCAR assumption are displayed in Table S. 4 (normal distribution) and S. 5 (log-normal distribution). A graphical overview for $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_4$, $\boldsymbol{\Sigma}_5$, $\boldsymbol{\Sigma}_6$ under the MCAR assumption is presented in Figures 4 (normal distribution) and 5 (log-normal distribution) for $n = 20$ and $r = 0.3$. Additionally, we investigated the power of the MCTP using only complete cases, which has a low power compared to the approaches using all-available information due to the decreased sample size. It follows that none of the procedures which use all-availble information is superior in terms of their powers to detect alternatives. However, the MCTP using all available information, provides more information by local test decisions, simultaneous confidence intervals and adjusted $p$-values and is therefore recommended for practical applications despite beeing slightly liberal in some cases.

Next, we investigate the power of the procedures under both MAR scenarios. The results can be found in Figures S. 2 to S. 5. Similar to the results under MCAR, none of the procedures which use all-available information is superior. However, the MCTP using only complete cases exhibits a comparable power in the second MAR scenario compared to the competing

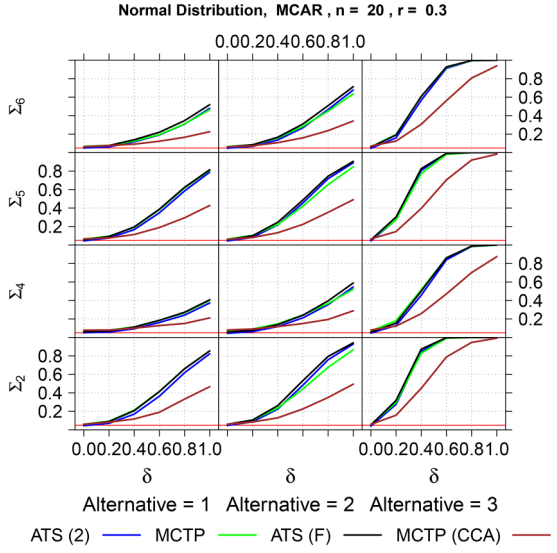


**Figure 4.** Power simulation of the second version of the ATS and the MCTP for testing $H_0^p$, the ATS for testing $H_0^F$ and the MCTP for testing $H_0^p$ using only complete cases, data is MCAR. MCTP: multiple contrast test procedure; MCAR: missing completely at random; ATS: ANOVA-type statistic.

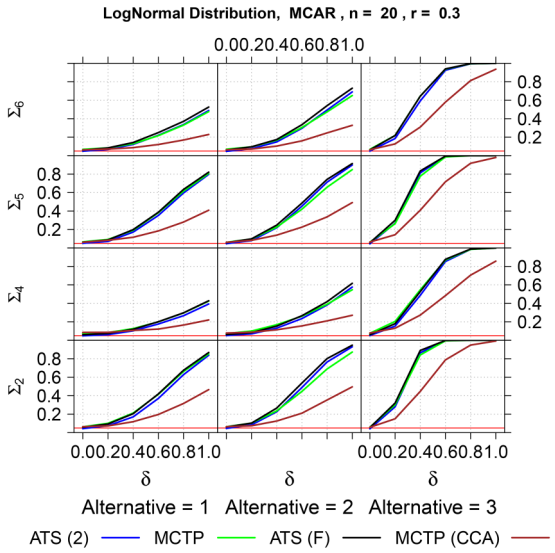


**Figure 5.** Power simulation of the second version of the ATS and the MCTP for testing $H_0^p$, the ATS for testing $H_0^F$ and the MCTP for testing $H_0^p$ using only complete cases, data is MCAR. MCTP: multiple contrast test procedure; MCAR: missing completely at random; ATS: ANOVA-type statistic.

procedures using all-available information. To summarize, we still recommend the novel MCTP for testing $H_0^p$ using all-availabe information if data is MAR.

## 9 Analysis of the example

The headache severity level migraine trial presented in Section "Motivating example" was analyzed using the MCTP, since this procedure can be used not only for testing the global hypothesis of no effect over all time points but also for testing pairwise comparisons. Thus, we chose a Tukey-type contrast matrix

$$\mathbf{C} = \begin{pmatrix} \boldsymbol{c}_1' \\ \boldsymbol{c}_2' \\ \boldsymbol{c}_3' \\ \boldsymbol{c}_4' \\ \boldsymbol{c}_5' \\ \boldsymbol{c}_6' \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

for testing the null hypotheses $H_0^p : \mathbf{Cp} = \mathbf{0}$. The estimated unweighted relative effects are given by $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_4)' = (0.55, 0.47, 0.45, 0.53)'$, which indicate that the scores obtained under session 3 are smallest, followed by sessions 2 and 4, whereas the scores obtained under session 1 are the largest. These computations match the visual impression attained by the boxplots in Figure 1. As the direction of the trend was unknown, we calculated two-sided simultaneous confidence intervals (22) at 95% confidence level. The results along with the values of test statistics $T_\ell$ and $p$-values are displayed in Table 2. Note, that no multiplicity adjustment was necessary as we used the critical value obtained

**Table 2.** Point estimators, simultaneous confidence intervals, $t$-values, and $p$-values for Tukey-type contrasts in relative effects in the migraine trial.

| Comparison | Estimator | 95 %- Confidence interval | $t$-value | $p$-value |
|---|---|---|---|---|
| $\widehat{p}_2 - \widehat{p}_1$ | −0.087 | [−0.165, −0.016] | 2.867 | 0.023 |
| $\widehat{p}_3 - \widehat{p}_1$ | −0.103 | [−0.196, −0.020] | 2.882 | 0.022 |
| $\widehat{p}_4 - \widehat{p}_1$ | −0.028 | [−0.148, 0.080] | 0.610 | 0.927 |
| $\widehat{p}_3 - \widehat{p}_2$ | −0.017 | [−0.099, 0.058] | 0.522 | 0.952 |
| $\widehat{p}_4 - \widehat{p}_2$ | 0.058 | [−0.063 0.168] | 1.241 | 0.594 |
| $\widehat{p}_4 - \widehat{p}_3$ | 0.075 | [−0.049, 0.187] | 1.566 | 0.393 |

from the MCTP in each comparison.

It follows from Table 2 that the data provides the evidence to reject the global null hypothesis $H_0^p$ (p-val. = 0.022), indicating that the treatment has an effect on the migraine of the patients over the course of time. However, session 4 does not indicate an improvement upon the first session. Applying the ANOVA-type procedure yields a higher global p-value of 0.047 (first version of the ATS) and 0.049 (second version of the ATS), respectively.

## 10 Discussion and conclusions

Missing values appear naturally in RM designs. Beyond proper statistical modeling, estimation problems of (model) parameters exist and typically complicate the analysis. In this paper, we discussed purely nonparametric methods and their limitations. In particular, we extended the methods proposed by Konietschke et al.[5] to allow for missing data. The simulation study demonstrates that the proposed methodology controls the type-I-error satisfactorily even for small sample sizes and a high missing rate. The already existing method for RM data with missing values from Domhof et al.[3] can only be used for testing the null hypothesis $H_0^F : F_1 = \cdots = F_d$ with regard to the marginal distribution functions, which is difficult to interpret and does not allow for calculating confidence intervals. The newly proposed method can be used for testing the less strict hypothesis $H_0^p$ and for calculating confidence intervals. Note, that the procedures are not limited to metric data; even ordered categorical data and binary data can be examined in a unified way. All results achieved in this paper are valid under the MCAR mechanism and the results of our simulation study indicate that all proposed methods are not sensitive to MAR scenarios. Further simulation studies indicated that in "extreme" scenarios, for example, smaller sample sizes, high missing probabilities, heteroscedasticity, and high correlations, the MCTP tends to be liberal. Therefore, we plan to explore resampling techniques for making the MCTP more accurate in these scenarios. Moreover, extensions to split plot designs and clustered data will be part of future research.

### Acknowledgments

The authors are grateful to the Associate Editor and the two anonymous referees for helpful comments which considerably improved the paper.

### Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work of Kerstin Rubarth and Frank Konietschke was funded by Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG KO 4680/3-2. The work of Markus Pauly was funded by Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG PA 2409/3-2.

### ORCID iDs

Kerstin Rubarth https://orcid.org/0000-0002-6174-6346
Frank Konietschke https://orcid.org/0000-0002-5674-2076

### Supplemental Material

Supplementary material for this article is available online.

### References

1. Little R and Rubin D. Statistical analysis with missing data. 2nd edn. hoboken. *NJ: John Wiley & Sons, Inc* 2002; **4**.
2. Brunner E, Domhof S and Langer F. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. 2002.
3. Domhof S, Brunner E and Osgood DW. Rank procedures for repeated measures with missing values. *Sociol Methods Res* 2002; **30**: 367–393.
4. Akritas M, Antoniou E and Kuha J. Nonparametric analysis of factorial designs with random missingness: Bivariate data. *J Am Stat Assoc* 2006; **101**: 1513–1526.
5. Konietschke F, Bathke A, Hothorn L, et al. Testing and estimation of purely nonparametric effects in repeated measures designs. *Comput Stat Data Anal* 2010; **54**: 1895–1905.
6. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann Math Stat* 1954; **25**: 290–302.
7. Kostecki-Dillon T, Monette G and Wong P. Pine trees, comas and migraines. *Newsletter York University Institute for Social Research* 1999, **14**.

8. Gao X. A nonparametric procedure for the two-factor mixed model with missing data. *Biometrical J Biometrische Zeitschrift* 2007; **49**: 774–788.
9. Konietschke F, Harrar S, Lange K, et al. Ranking procedures for matched pairs with missing data – asymptotic theory and a small sample approximation. *Comput Stat Data Anal* 2012; **56**: 1090–1102. Second Issue for COMPUTATIONAL STATISTICS FOR CLINICAL RESEARCH.
10. Ruymgaart F. *A unified approach to the asymptotic distribution theory of certain midrank statistics*, volume 821. ISBN 978-3-540-10239-7, 2006. pp. 1–18.
11. Brunner E, Bathke A and Konietschke F. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS*. 2018. ISBN 978-3-030-02912-8.
12. Konietschke F, Hothorn L and Brunner E. Rank-based multiple test procedures and simultaneous confidence intervals. *Electron J Stat* 2012; **6**: 738–759.
13. Brunner E, Konietschke F, Pauly M, et al. Rank-based procedures in factorial designs: Hypotheses about nonparametric treatment effects. *J R Stat Soc, Series B (Stat Methodol)* 2016; **79**: 1463–1485.
14. Umlauft M, Placzek M, Konietschke F, et al. Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *J Multivar Anal* 2019; **171**: 176–192.
15. Brunner E and Munzel U. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biom J* 2000; **42**: 17–25.
16. Domhof S. *Nichtparametrische relative Effekte (Doctoral dissertation, Niedersächsische Staats-und Universitätsbibliothek Göttingen)*. 2001. https://ediss.uni-goettingen.de/bitstream/handle/11858/00-1735-0000-000D-F284-4/domhof.pdf?sequence=1.
17. Akritas MG and Brunner E. A unified approach to rank tests for mixed models. *J Stat Plan Inference* 1997; **61**: 249–277.
18. Brunner E, Munzel U and Puri ML. Rank-score tests in factorial designs with repeated measures. *J Multivar Anal* 1999; **70**: 286–317.
19. Bretz F, Genz A and A Hothorn L. On the numerical availability of multiple comparison procedures. *Biom J* 2001; **43**: 645–656.
20. Santos MS, Pereira RC, Costa AF, et al. Generating synthetic missing data: A review by missing mechanism. *IEEE Access* 2019; **7**: 11651–11667.
21. Amro L, Konietschke F and Pauly M. Incompletely observed nonparametric factorial designs with repeated measurements: A wild bootstrap approach, 2021. 2102.02871.
22. Zhu B, He C and Liatsis P. A robust missing value imputation method for noisy data. *Appl Intell* 2012; **36**: 61–74.
23. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.
24. Friedrich S, Konietschke F and Pauly M. A wild bootstrap approach for nonparametric repeated measurements. *Comput Stat Data Anal* 2017; **113**: 38–52.
25. Munzel U. Nonparametric methods for paired samples. *Stat Neerl* 1999; **53**: 277–286.
26. Harrar SW, Feyasa MB and Wencheko E. Nonparametric procedures for partially paired data in two groups. *Comput Stat Data Anal* 2020; **144**: 106903.
27. Amro L, Konietschke F and Pauly M. Multiplication-combination tests for incomplete paired data. *Stat Med* 2018; **38**: 3243–3255.
28. Van Buuren S. *Flexible imputation of missing data*. Boca Raton: CRC press, 2018.
29. Ramosaj B, Amro L and Pauly M. A cautionary tale on using imputation methods for inference in matched pairs design. *Bioinformatics (Oxford, England)* 2020, **36**: 3099–3106. doi: 10.1093/bioinformatics/btaa082.