rankFD: An R Software Package for Nonparametric Analysis of General Factorial Designs

by Frank Konietschke, Markus Pauly, Arne C. Bathke, Sarah Friedrich and Edgar Brunner

Abstract Many experiments can be modeled by a factorial design which allows statistical analysis of main factors and their interactions. A plethora of parametric inference procedures have been developed, for instance based on normality and additivity of the effects. However, often, it is not reasonable to assume a parametric model, or even normality, and effects may not be expressed well in terms of location shifts. In these situations, the use of a fully nonparametric model may be advisable. Nevertheless, until very recently, the straightforward application of nonparametric methods in complex designs has been hampered by the lack of a comprehensive R package. This gap has now been closed by the novel R-package **rankFD** that implements current state of the art nonparametric ranking methods for the analysis of factorial designs. In this paper, we describe its use, along with detailed interpretations of the results.

1 Introduction

Nonparametric methods and in particular rank-based methods are commonly used for the analysis of experiments when it cannot be assumed that the observations derive from a normal population distribution. In online discussion fora regarding the application of statistical methods one can often find questions such as: "Does anybody know whether there is a nonparametric analog of ANOVA?". The common response is: "You may use rank methods" which usually prompts the next question: "Does anybody know a software package performing the computations for a nonparametric ANOVA / rank ANOVA?". The answers to this question vary: some list more or less popular statistical software packages, others give the heuristic advice of simply replacing the observations by their ranks and then performing regular ANOVA on the ranks. This suggests that there is a lack of clear advice on not just how to implement rank-based methods, but also how to interpret and understand the theoretical background. As such, the goal of the present article is to both explain when and how to use the procedures implemented in **rankFD**, and also provide the reader with enough of the theoretical background so that they can interpret the results correctly.

In order to provide a more precise answer regarding the nonparametric analog of ANOVA, one has to discuss the quantities by which a potential effect in a trial can be intuitively described. Such effects may be the differences or ratios of the means of the observations or of some other parameter or estimand defined in a semi-parametric model. To compare the differences of means in semi-parametric models where the normal distribution cannot be assumed, the so-called studentized permutation procedures (Janssen, 1997; Pauly et al., 2015; Smaga, 2015) are appropriate. These procedures provide quite accurate results even in case of small to moderate sample sizes, depending on the type of the data and the underlying population distribution. However, there are several situations where differences or linear combinations of means may not be appropriate to describe intuitive treatment effects – for example if the data have floor and ceiling effects or if the distributions have completely different shapes. In case of ordinal data, means are not even defined, and using a numerical encoding of the ordered categories as seemingly metric data may lead to incorrect conclusions (Kahler et al., 2008). In such cases, treatment effects can reasonably be described by the so-called *relative effect* which was introduced by Mann and Whitney (1947) and Putter (1955). For independent observations $X \sim F_1$ and $Y \sim F_2$, the relative effect is defined as $\theta = P(X < Y) + \frac{1}{2}P(X = Y)$, which can be equivalently written as $\theta = \int F_1 dF_2$. It may be noted that this effect has been known under many different names in the literature, for example Wilcoxon functional (Janssen, 1999a), Mann-Whitney type effect (Dobler et al., 2019), stochastic superiority (D'Agostino et al., 2006), or probabilistic index (Acion et al., 2006; Thas et al., 2012). We prefer the expression "relative effect" or "nonparametric relative treatment effect" with reference to Birnbaum and Klose (1957).

The relative effect θ can be estimated by replacing the distribution functions F_1 and F_2 with their empirical counterparts, $\widehat{F}_1$ and $\widehat{F}_2$, the so-called empirical distribution functions. This leads to the estimator $\widehat{\theta} = \int \widehat{F}_1 d\widehat{F}_2 = \frac{1}{n_1}\left(\overline{R}_{2\cdot} - \frac{n_2+1}{2}\right)$, where $\overline{R}_{2\cdot} = \frac{1}{n_2}\sum_{k=1}^{n_2} R_{2k}$ denotes the mean of the overall ranks R_{2k} of the observations $X_{21}, \ldots, X_{2n_2}$ among all $N = n_1 + n_2$ observations in the experiment. It is well-known that $\widehat{\theta}$ is an unbiased and L_2-consistent estimator of the relative effect θ and thus, the mean of the ranks provides the basis for estimating θ and for statistical inference regarding θ.

For two random variables X and Y, a relative effect $\theta > 1/2$ indicates a tendency that X takes smaller values than Y, while $\theta < 1/2$ means that X tends to have larger values than Y. No tendency in either direction corresponds to a relative effect of $\theta = \frac{1}{2}$. Crucially, the presence of a relative effect does not translate to a difference in means, and likewise, the absence of a relative effect does not suggest that the means are the same. In other words, if X has a mean μ_x and Y has a mean μ_y, then we may have $\theta \neq 1/2$ when $\mu_x = \mu_y$, or $\theta = \frac{1}{2}$ when $\mu_x \neq \mu_y$. Analogously, for the medians $\widetilde{\mu}_x$ and $\widetilde{\mu}_y$, it is possible that $\theta \neq \frac{1}{2}$ and $\widetilde{\mu}_x = \widetilde{\mu}_y$, or that $\theta = \frac{1}{2}$ and $\widetilde{\mu}_x \neq \widetilde{\mu}_y$. Thus, from a significant result of a rank test it cannot be concluded that $\mu_x \neq \mu_y$ or $\widetilde{\mu}_x \neq \widetilde{\mu}_y$. In this sense, rank tests based on $\widehat{\theta}$ (e.g., the Wilcoxon-Mann-Whitney test, the Fligner-Policello test, or the Brunner-Munzel test) are not tests of the equality of means or medians, and therefore not simply nonparametric analogs of the t-test since the hypotheses and consistency regions of these tests are not identical. Note that the consistency region contains all distribution functions for which the power of the test tends to 1 as sample sizes tend to ∞. In most parametric models, the set of distribution functions contained in the hypothesis and in the consistency region are complementary. In some nonparametric models, however this is in general not the case which may lead to difficulties interpreting "significant" results obtained by rank-based tests (Brunner et al., 2020). Some details will be explained in Section 2.2. Similar remarks apply to rank tests for multiple samples or even in factorial designs. This is ultimately the reason why the heuristic approach of replacing the observations by their ranks may lead to non-valid procedures in general (Conover and Iman, 1981). Especially in factorial designs, linear combinations of means may have different meanings than linear combinations of relative effects. With this in mind, users of the R-package for rank tests described in this paper should know that they might get different results than obtained by using a common ANOVA package.

The second question often read in discussion fora —'what software package should I use' —can be answered more easily. Most statistical software packages provide options for the classical nonparametric rank-based methods, however, these can still be quite limited and more contemporary and/or appropriate methods may not be available. For example, most statistical software packages offer the Wilcoxon-Mann-Whitney and the Kruskal-Wallis test for independent observations, as well as some particular procedures from the literature. However, more modern nonparametric rank-based methods developed during the last decades (Ruymgaart, 1980; Akritas and Arnold, 1994; Akritas et al., 1997; Brunner and Puri, 1996; Konietschke et al., 2012; Brunner et al., 2017, 2019) are not mplemented in most packages. Moreover, in software tools following a more classical paradigm, ties (i.e., two or more different observations with exactly the same value, as frequently is the case in ordinal or count data) are often considered in form of "corrections" that are added to the case of no ties, instead of considering the situation of no ties as a special case of a general model allowing for arbitrary ties (only the trivial case of one-point distributions should generally be excluded). Also, quick algorithms (Streitberg and Röhmel, 1986; Mehta et al., 1988) for the computation of exact p values for permutation-based procedures are rarely used, and general methods for purely nonparametric effects in factorial designs are not provided in standard implementations. However, exactly such procedures are often needed in applications. Researchers are then tempted to use heuristic procedures as described above, although the conclusions drawn from them might be misleading.

Finally, confidence intervals for purely nonparametric effects, such as the relative effect θ, are not provided in standard software, in spite of the fact that appropriate confidence intervals for the effect measures being used in the analysis have been required by the pertinent guidelines for decades. Instead, some software packages offer confidence intervals for location shift effects which in general may be neither compatible to the decisions of the rank tests nor justified regarding the types of alternatives or the scales of the measurements in the experiment. Recall that the relative effect is not a measure of mean or median differences, and therefore confidence intervals for mean or median shifts are not congruent with hypothesis tests based on the relative treatment effect, such as the Wilcoxon-Mann-Whitney and the Kruskal-Wallis tests, among others.

The R package **rankFD** intends to close these gaps. It includes the classical rank tests for continuous observations as special cases, allows for situations with arbitrary ties, and extends these procedures to factorial designs. The hypotheses tested in factorial designs are expressed as linear hypotheses in terms of the distribution functions as introduced in Akritas et al. (1997) or as linear combinations of the relative effects as discussed in Brunner et al. (2017). Ranking procedures for testing equalities of distribution functions in factorial longitudinal data (repeated measures) and multivariate data are implemented in R packages **nparLD** (Noguchi et al., 2012), **npmv** and **nparMD** (Burchett et al., 2017; Kiefel et al., 2022), respectively. Semiparametric methods for testing null hypotheses in general factorial designs in means are implemented in the R packages **GFD** (Friedrich et al., 2017) and **MANOVA.RM** (Friedrich et al., 2019b).

In any case, it must be clearly noted that rank methods, especially in factorial designs, answer different questions than those considered by the ANOVA in common factorial designs. The relations between linear combinations of the expectations of the observations and their respective counterparts

expressed in terms of rank or pseudo-rank means depend on the underlying distribution functions. Questions investigated by parametric factorial designs are related to the expected values of the observations, while questions investigated by using rank- and pseudo-rank-based methods are related to relative effects. The latter compare the distributions in the different treatment groups to an average distribution. Thus, it should not be a surprise to obtain different answers if different questions are posed. This must be kept in mind when responding to the seemingly simple question: "Does anybody know whether there is a nonparametric analog of ANOVA?".

The paper is organized as follows. Section 2.2 discusses the statistical models and explains the concepts and methodology underlying the inferential procedures provided by the package **rankFD** while the corresponding test statistics are described in Section 2.3. Section 2.4 lists and explains the different functions used in this package, as well as examples demonstrating the usage of these functions on real-life data. The paper closes with a discussion of the meaning and interpretation of these methods and their relations to some procedures implemented in other R packages

2 Statistical models, effects, and hypotheses

First we consider the simple experimental design involving only one factor A with a levels involving n_i independent observations in each level i. These are modeled as

$$X_{ik} \sim F_i,\ i = 1, \ldots, a;\ k = 1, \ldots, n_i. \tag{1}$$

Throughout, we assume that the observations X_{ik} are measured at least on an ordinal scale, whereas F_i denotes an arbitrary distribution (or its cdf), with the exception of one-point distributions. In total, there are $N = \sum_{i=1}^{a} n_i$ observations in the trial. This statistical model does not involve any explicit parameters or parametrization that could be used to describe appropriate treatment effects. To describe effects in such a general model, we therefore define weighted and unweighted *relative effects*

$$\theta_i = \int H_N dF_i = P(Y < X_{ik}) + \frac{1}{2}P(Y = X_{ik}), \quad i = 1, \ldots, a, \tag{2}$$

$$\psi_i = \int G dF_i = P(Z < X_{ik}) + \frac{1}{2}P(Z = X_{ik}), \quad i = 1, \ldots, a. \tag{3}$$

In this general definition of a relative treatment effect, each distribution function F_i is compared either to a weighted average $H_N = \frac{1}{N}\sum_{i=1}^{a} n_i F_i$ or an unweighted average $G = \frac{1}{a}\sum_{i=1}^{a} F_i$ of the distribution functions. This can be regarded as comparing each observation $X_{ik} \sim F_i$ with either an artificial independent observation $Y \sim H_N$ of the weighted mean distribution or $Z \sim G$ of the unweighted mean distribution. The former leads to the weighted relative effect θ_i, while the latter leads to the unweighted relative effect ψ_i. In case of equal sample sizes, both effects coincide.

The unweighted relative effects ψ_i can be interpreted as follows: If $\psi_i < \frac{1}{2}$, then the observations in group i tend to be smaller than those coming from the average distribution G. If $\psi_i = \psi_j$, then in relation to the average distribution G, the observations coming from distributions F_i and F_j have the exact same tendency towards smaller or larger observations. Thus, it is reasonable to consider the case of $\psi_i = \psi_j$ as *no (relative) treatment effect* between levels i and j. The relations and interpretations for the weighted effects θ_i and θ_j follow analogously. In the following, we collect all distribution functions and relative effects in the vectors $\mathbf{F} = (F_1, \ldots, F_a)^\top$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_a)^\top$ or $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_a)^\top$, respectively.

Estimators of the weighted relative effects θ_i defined in (2) can be obtained using the ranks R_{ik} of the observations X_{ik}. In fact, $\widehat{\theta}_i = \frac{1}{N}\left(\overline{R}_{i\cdot} - \frac{1}{2}\right)$ is an unbiased and consistent estimator of θ_i, where $\overline{R}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik}$, and R_{ik} denotes the rank of X_{ik} among all $N = \sum_{i=1}^{d} n_i$ observations. In case of ties, mid-ranks must be used. Formally, the mid-rank R_{ik} is obtained from the empirical weighted average distribution function $\widehat{H}_N(x) = \frac{1}{N}\sum_{i=1}^{d} n_i \widehat{F}_i(x)$ by $R_{ik} = \frac{1}{2} + N\widehat{H}(X_{ik})$.

In the same way, the unweighted relative effects ψ_i defined in (3) are estimated using the so-called pseudo-ranks $R^{\psi}_{ik} = \frac{1}{2} + N\widehat{G}(X_{ik})$, where $\widehat{G}(x)$ denotes the empirical unweighted average distribution function. An unbiased and consistent estimator $\widehat{\psi}_i$ of ψ_i is given by

$$\widehat{\psi}_i \quad = \quad \frac{1}{N}\left(\overline{R}^{\psi}_{i\cdot} - \frac{1}{2}\right), \tag{4}$$

where $\overline{R}^{\psi}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} R^{\psi}_{ik}$. For details we refer to Brunner et al. (2019), Section 2.3.2 or to Happ et al. (2020), Section 2. Basically, the estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_a)^\top$ and $\widehat{\boldsymbol{\psi}} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_a)^\top$ are vectors whose components are linear functions of the rank means $\overline{R}_{i\cdot}$ or the pseudo-rank means $\overline{R}^{\psi}_{i\cdot}$, respectively.

Thus, rank tests are related to the weighted relative effects θ_i in (2), while pseudo-rank tests are related to the unweighted relative effects ψ_i in (3).

Hypotheses formulated in terms of distribution functions

Classical rank-based methods for a one-way layout, (e.g., Kruskal-Wallis test, Kruskal (1952); Kruskal and Wallis (1952); or Hettmansperger-Norton test, Hettmansperger and Norton (1987)) can be used to test null hypotheses formulated in terms of the distribution functions, such as

$$H_0^F : \quad F_1 = \ldots = F_a, \tag{5}$$

where obviously, equal distribution functions imply equal variances if H_0^F in (5) is true (if second moments exist).

Two- and higher way layouts are covered within model (1) by sub-indexing the index i, similar to the theory of linear models. For instance, a two-way design involving a factor A with a levels and a factor B with b levels, respectively, can be written as

$$X_{ijk} \quad \sim \quad F_{ij}, i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, n_{ij}, \tag{6}$$

and the distribution functions and relative effects are then collected in the structured vectors $\mathbf{F} = (F_{11}, \ldots, F_{ab})^\top$ and $\boldsymbol{\psi} = (\psi_{11}, \ldots, \psi_{ab})^\top$ or $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{ab})^\top$, respectively.

Consequently, Akritas and Arnold (1994), Brunner and Puri (1996), and Akritas et al. (1997) suggested to formulate null hypotheses in two- and higher-way layouts in a similar way as in linear models, with the expected values being replaced by the corresponding distribution functions. In a two-way layout, for example, hypotheses of no (distribution-)main effects A or B and no (distribution-)interaction (AB) are written as

$$\begin{aligned}
H_0^F(A): &\qquad \overline{F}_{1\cdot} = \cdots = \overline{F}_{a\cdot}, &&\qquad \overline{F}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} F_{ij}, &&\qquad i = 1, \ldots, a, \\
H_0^F(B): &\qquad \overline{F}_{\cdot 1} = \cdots = \overline{F}_{\cdot b}, &&\qquad \overline{F}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} F_{ij}, &&\qquad j = 1, \ldots, b, \\
H_0^F(AB): &\qquad F_{ij} = \overline{F}_{i\cdot} + \overline{F}_{\cdot j} - \overline{F}_{\cdot\cdot}, &&\qquad \overline{F}_{\cdot\cdot} = \frac{1}{ab}\sum_{r=1}^{a}\sum_{s=1}^{b} F_{rs}, &&\qquad i = 1, \ldots, a; j = 1, \ldots, b\,.
\end{aligned} \tag{7}$$

In order to extend the hypotheses in (5) or (7) to higher-way layouts, general hypotheses are written using matrix notation as

$$H_0^F(\mathbf{C}) : \mathbf{CF} \quad = \quad \mathbf{0}, \tag{8}$$

where $\mathbf{C}$ denotes an appropriate hypothesis matrix, in the same way as in linear models, only replacing means with the respective distribution functions. Note that $\mathbf{0}$ is here understood to be a vector of functions which are identically 0. Testing these hypotheses H_0^F of no distribution effects can be performed using the argument `hypothesis="H0F"` in the `rankFD` function. More details are provided in Section 2.4.

Hypotheses formulated in terms of relative effects

In general, researchers may not be interested in detecting the somewhat abstract alternative $H_1^F : \mathbf{CF} \neq \mathbf{0}$ that H_0^F in (8) is not true, but instead they want to detect whether a tendency to smaller or larger values exists between treatment levels. In a one-way layout, for example, the latter corresponds to the testing problem

$$H_0^p : \psi_1 = \ldots = \psi_a, \tag{9}$$

formulated in terms of the relative effects ψ_i. Here, the symbol H_0^p refers to the probabilities ψ_i in (3).

Remark: Of course, one can also state the hypothesis

$$H_0^p : \theta_1 = \ldots = \theta_a, \tag{10}$$

but it must be kept in mind that the hypothesis (10) depends on the relative sample sizes n_i/N in groups $i = 1, \ldots, a$. Thus, the rejection region of such a test is not invariant, but it changes with the ratios n_i/N of the sample sizes. In extreme cases, this might lead to surprising results when compared

to the results obtained in designs with equal sample sizes. For details we refer to Brunner et al. (2020) and Brunner et al. (2019). The unweighted mean distribution is, however, one reference distribution of choice that helps in reducing the issues obtained with the weighted version. Whether the unweighted version is the "best" one, can not be answered and guaranteed, in general (Zimmermann et al., 2022).

In a two-way layout, for example, the hypotheses of no main effects or no interactions in terms of the relative effects $\psi_{ij} = \int GdF_{ij}$ are written as

$$\begin{aligned} H_0^P(A): &\quad \overline{\psi}_{1\cdot} = \cdots = \overline{\psi}_{a\cdot}, && i = 1,\ldots,a, \\ H_0^P(B): &\quad \overline{\psi}_{\cdot 1} = \cdots = \overline{\psi}_{\cdot b}, && j = 1,\ldots,b, \\ H_0^P(AB): &\quad \psi_{ij} = \overline{\psi}_{i\cdot} + \overline{\psi}_{\cdot j} - \overline{\psi}_{\cdot\cdot}, && i = 1,\ldots,a;\ j = 1,\ldots,b, \end{aligned} \tag{11}$$

where $\overline{\psi}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\psi_{ij}$, $\overline{\psi}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\psi_{ij}$, and $\overline{\psi}_{\cdot\cdot} = \frac{1}{2}$. The matrix notation of these hypotheses is, analogously to (7) and (8),

$$H_0^P(\mathbf{C}): \mathbf{C}\boldsymbol{\psi} \quad = \quad \mathbf{0}, \tag{12}$$

where $\boldsymbol{\psi}$ denotes the vector of unweighted relative effects. For a detailed explanation of using matrix notation in factorial designs we refer to, e.g., Brunner et al. (2017) or Brunner et al. (2019), Sect. 5.2 and Sect. 8.7.1.

In a similar way as in the one-way layout, the hypotheses involving the weighted relative effects θ_{ij} in the two-way layout can be stated by replacing ψ_{ij}, $\overline{\psi}_{i\cdot}$, and $\overline{\psi}_{\cdot j}$ in (11) with θ_{ij}, $\overline{\theta}_{i\cdot}$, and $\overline{\theta}_{\cdot j}$, respectively. It may be noted, however, that – unlike in the one-way layout – in two- or higher-way layouts surprising results may already be obtained in case of moderate unequal samples sizes in simple shift-effect models. These basic models cannot be considered "extreme cases". This means that unequal sample sizes in two- or higher-way layouts constitute a serious challenge for rank tests while this is not the case for pseudo-rank tests. For more details we refer to Brunner et al. (2019), Chapter 5 and Brunner et al. (2020), Section 4.

Note that H_0^P in (12) neither implies variance homogeneity nor equal shapes of the distributions. In the case of two samples, this situation is also known as the *nonparametric Behrens-Fisher* problem (Fligner and Policello, 1981; Brunner and Munzel, 2000; Konietschke et al., 2012). In general, it is easier to estimate the covariance matrix of the empirical relative effects under the stronger null hypothesis H_0^F than under H_0^P. Therefore, statements about the sampling distribution of test statistics based on ranks have traditionally been formulated under H_0^F, even though it is well-known that those test statistics can only detect alternatives of the form $H_1^P: \mathbf{C}\boldsymbol{\theta} \neq \mathbf{0}$ or $H_1^P: \mathbf{C}\boldsymbol{\psi} \neq \mathbf{0}$.

Remark: The **rankFD** package implements the current state-of-the-art methods for testing H_0^P (using ranks as well as pseudo-ranks) in general factorial designs (Konietschke et al., 2012; Brunner et al., 2017), and it allows for the computation of a wide range of nonparametric test statistics. It explicitly also includes the classical tests based on weighted relative effects θ_i (using ranks) and on unweighted relative effects ψ_i (using pseudo-ranks). Both types of ranking procedures are included in **rankFD**. A reason for including the former tests is that it allows users to reproduce findings that have been obtained by other researchers using rank tests. Also, it offers the possibility to directly compare procedures which may facilitate a transparent discussion in that regard.

Multiple comparisons

So far, both null hypotheses H_0^F and H_0^P have been written as global null hypotheses. If they get rejected, one may only conclude that *some* factor level differs from the others (at corresponding significance level α). However, it still remains unknown specifically *which one* differs. Therefore, testing global null hypotheses often does not answer the particular research question of interest to scientists applying statistical methods, namely the specific localization of those treatment groups that are "driving" the significant results. In order to accomplish this goal, testing linear contrasts using a $q \times a$ contrast matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^\top \\ \vdots \\ \mathbf{c}_q^\top \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1a} \\ c_{21} & c_{22} & \cdots & c_{2a} \\ \vdots & \vdots & \vdots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qa} \end{pmatrix};\ \sum_{i=1}^{a} c_{\ell i} = 0,\ \ell = 1,\ldots,q,$$

in terms of multiple null hypotheses $H_0^{(\ell)}: \mathbf{c}_\ell^\top\boldsymbol{\psi} = 0$ (or $H_0^{(\ell)}: \mathbf{c}_\ell^\top\boldsymbol{F} = 0$) is the key. Here, each row vector $\mathbf{c}_\ell^\top$ describes one of q different contrasts reflecting the researcher's particular question. For

instance, in a one-way layout with $a = 4$ levels, many-to-one (Dunnett-type) (Dunnett, 1955) or all pairwise (Tukey-type) comparisons are performed with the contrast matrices

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix} \qquad \text{or} \qquad \mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

Note: left shows many-to-one (Dunnett-type); right shows all-pairwise (Tukey-type) contrast matrix

respectively. Which contrast to use depends on the respective research question of interest. Bretz et al. (2001) provide a broad overview of different contrast matrices, which are numerically available within the `contrMat` function of the **multcomp** package in R (Hothorn et al., 2008). In general factorial designs involving more than one factor, multiple comparisons in terms of means of the levels of the main effects are a meaningful and valuable asset of a fundamental data analysis. For instance, in a 2×4 two-way design, many-to-one comparisons to the control group ($j = 1$ of factor B) are expressed as

$$\begin{aligned} H_0^{P(1)} &: \overline{\psi}_{\cdot 1} = \overline{\psi}_{\cdot 2} \\ H_0^{P(2)} &: \overline{\psi}_{\cdot 1} = \overline{\psi}_{\cdot 3} \\ H_0^{P(3)} &: \overline{\psi}_{\cdot 1} = \overline{\psi}_{\cdot 4} \end{aligned} \qquad \mathbf{C} = \begin{array}{c} \begin{matrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} & \psi_{21} & \psi_{22} & \psi_{23} & \psi_{24} \end{matrix} \\ \begin{pmatrix} 1/2 & -1/2 & 0 & 0 & 1/2 & -1/2 & 0 & 0 \\ 1/2 & 0 & -1/2 & 0 & 1/2 & 0 & -1/2 & 0 \\ 1/2 & 0 & 0 & -1/2 & 1/2 & 0 & 0 & -1/2 \end{pmatrix} \end{array}.$$

The `rankFD` function implements a broad list of pre-defined contrasts as well as flexible options allowing for user-defined contrast matrices for making multiple comparisons of the levels of the main or interaction effects. We provide computational details in Section 2.4.

Confidence intervals

To comply with the basic principle "*no test without a confidence interval*", the **rankFD** package also provides confidence intervals for the nonparametric quantities upon which the test is based. Two-sided $(1-\alpha)$-confidence intervals for ψ_i and $\theta = \psi_2 - \psi_1$ are obtained from the asymptotic distribution of the estimators $\widehat{\psi}_i$ in (4) by

$$CI = \left[\widehat{\psi}_i \mp z_{1-\alpha/2} \frac{\widehat{s}_i}{\sqrt{N}} \right], \tag{13}$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution. Here, the variance estimator $\widehat{s}_i^2$ is a quite involved linear combination of different quadratic forms obtained from different rankings of the observations X_{ik}. For details we refer to Brunner et al. (2019), Sect. 4.6.1.

The confidence intervals in (13) may suffer from poor coverage probability if ψ_i is close to the limits 0 or 1 and, moreover, the limits of the confidence interval may exceed the boundaries 0 or 1. In this case, so-called *range preserving* intervals can be obtained by using the *logit*-transformation. The limits thus obtained are then "back-transformed" using the *expit*-transformation. For details we refer to Brunner et al. (2019), Sect. 4.6.2.

In **rankFD**, these confidence intervals are computed by the function `rankFD()` using the options CI.method = "normal" for the limits in (13) or CI.method = "logit" for the range preserving confidence intervals obtained by the *logit*-transformation. By default, `rankFD()` provides confidence intervals for both, ψ_i and θ_i. Regarding the confidence intervals for θ_i the same remarks as in Sect. 2.2.2 apply. Furthermore, since the Wilcoxon-Mann-Whitney test (and relative methods) use variance estimators that are only consistent under the respective null hypothesis H_0^F formulated in terms of the distribution functions, the tests cannot be inverted into confidence intervals for ψ_i.

3 Test statistics

The **rankFD** package implements a broad class of different test statistics for testing the general null hypotheses $H_0^F : \mathbf{CF} = \mathbf{0}$, $H_0^P : \mathbf{C}\boldsymbol{\psi} = \mathbf{0}$, and $H_0^P : \mathbf{C}\boldsymbol{\theta} = \mathbf{0}$, respectively. They include global test procedures (quadratic forms) and multiple contrast tests (linear statistics) for the analysis of data

from general factorial designs, as well as methods specifically designed for the evaluation of two independent samples including the classical rank tests.

In the following, we will briefly explain these procedures. They are all based on the (asymptotic) distribution of standardized vectors of point estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_d)^\top$ or $\widehat{\boldsymbol{\psi}} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_d)^\top$ of the *weighted* or *unweighted* relative effects as defined in (2) and (3), respectively. Since both of them denote the probabilities (appropriately weighted) of data being smaller in group i than in the joint sample, estimators can be constructed using the (usual) ranks R_{ik} or the so-called *pseudo-ranks* R_{ik}^{ψ} (Happ et al., 2020). In **rankFD** these point estimators are obtained by

`effect=weighted`	(scaled) mean of ranks R_{ik}
`effect=unweighted`	(scaled) mean of pseudo-ranks R_{ik}^{ψ}

For more details, we refer to (Brunner et al., 2019, Section 2.3.2). Besides the vectors of point estimators $\widehat{\boldsymbol{\theta}}$ or $\widehat{\boldsymbol{\psi}}$, their (estimated) covariance matrices are needed for the computation of test statistics. In the general nonparametric setup considered here, we can take advantage of the type of hypothesis we aim to test. Assuming H_0^F to hold, then the covariance matrices of $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ and of $\sqrt{N}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi})$ have (much) simpler structures than under H_0^P (Konietschke et al., 2012). This property carries over to its estimation and therefore the estimators used in the statistics for testing H_0^F or H_0^P are different. However, for the ease of notation, we denote with $\widehat{\mathbf{V}}_N$ their estimators in a general way having both versions in mind. In the following, we therefore provide the statistics using $\widehat{\boldsymbol{\psi}}$ (and in turn the pseudo-ranks) for the ease of convenience only. For more details we refer to Brunner et al. (2020).

Global test procedures

In order to test the null hypothesis H_0^F as given in (8), the **rankFD** package implements the *Wald-type* statistic

$$W_N(\mathbf{C}) \quad = \quad N\widehat{\boldsymbol{\psi}}^\top \mathbf{C}^\top \left[\mathbf{C}\widehat{\mathbf{V}}_N\mathbf{C}^\top\right]^+ \mathbf{C}\widehat{\boldsymbol{\psi}}. \tag{14}$$

Here, the matrix $[\mathbf{A}]^+$ denotes the Moore-Penrose inverse of the matrix $\mathbf{A}$. Under the hypothesis H_0^F, the statistic $W_N(\mathbf{C})$ follows, for large sample sizes, a χ^2_w-distribution with $w = \text{rank}(\mathbf{C}\widehat{\mathbf{V}}_N\mathbf{C}^\top)$ degrees of freedom. Since the statistic involves the estimators and the known contrast matrix only, its numerical computation is feasible. However, very large sample sizes ($n_i \geq 50$; depending on the actual design) are necessary for an accurate type-1 error rate control. Therefore, Akritas et al. (1997) and Brunner et al. (2017) propose the so-called *ANOVA-type* statistic

$$A_N(\mathbf{C}) = N \cdot \frac{\widehat{\boldsymbol{\psi}}^\top \mathbf{A} \widehat{\boldsymbol{\psi}}}{\text{trace}(\mathbf{A}\widehat{\mathbf{V}}_N)}, \; \mathbf{A} = \mathbf{C}^\top \left[\mathbf{C}\mathbf{C}^\top\right]^+ \mathbf{C}, \tag{15}$$

and approximate its distribution by an F-distribution with $\widehat{f}_1$ and $\widehat{f}_2$ degrees of freedom (obtained via Box-type approximation as derived by Brunner et al. (1997)). In comparison with the *Wald-type* statistic $W_N(\mathbf{C})$ in (14), the *ANOVA-type* statistic $A_N(\mathbf{C})$ controls the type-I error much better in small sample sizes; $n_i \geq 15$ depending on the design and hypothesis of interest.

Moreover, the approximation of the distribution of $A_N(\mathbf{C})$ is also valid under the more general hypothesis H_0^P. We note that, basically, both statistics can also be computed using the ranks R_{ik} instead of the pseudo-ranks R_{ik}^{ψ}. But the general remarks in Sections 2.2.2 regarding the usual ranks R_{ik} must be carefully considered. We also note that the asymptotic distribution of the *Wald-type* statistic $W_N(\mathbf{C})$ under the more general hypothesis H_0^P is not the χ^2_w-distribution with $w = \text{rank}(\mathbf{C}\widehat{\mathbf{V}}_N\mathbf{C}^\top)$ in general. This would require an additional assumption on the sequence of the empirical covariance matrices $\widehat{\mathbf{V}}_N$ which cannot be verified in practice.

The preceeding comments and discussion might appear somewhat difficult to understand but they are necessary to explain the different options in the printout of **rankFD**. At this point, it becomes evident that the question "Does anybody know whether there is a nonparametric analog of ANOVA?" cannot be answered by some simple statements and that the heuristic technique replacing observations by their ranks and then performing an 'ANOVA on the ranks' may lead to non valid procedures and incorrect conclusions in general.

Multiple contrast test procedures

Both the Wald-type and ANOVA-type statistics are *global* tests, i.e. if the respective hypothesis H_0^F or H_0^P is rejected, the only available information is that *any* of the factor levels (or their combinations) differ at pre-assigned significance level α. The identification of the factor levels which are responsible for the difference is, however, often of major interest and a key research question. Local test decisions in terms of adjusted p-values and simultaneous confidence intervals are of primary importance and key elements of a complete data evaluation. These can be exposed using *Multiple Contrast Test Procedures* (MCTP) (Bretz et al., 2001; Hothorn et al., 2008; Konietschke et al., 2012), which are also known as *max-t-test* type procedures in parametric models (Konietschke et al., 2021). In order to test the local null hypothesis $H_0^{(\ell)} : \mathbf{c}_\ell^\top \boldsymbol{\psi} = 0$, we use the test statistic

$$T_\ell = \sqrt{N} \frac{\mathbf{c}_\ell^\top \widehat{\boldsymbol{\psi}}}{\mathbf{c}_\ell^\top \widehat{\mathbf{V}}_N \mathbf{c}_\ell}, \tag{16}$$

where the contrast vector $\mathbf{c}_\ell$ reflects the researcher's particular question. Typical contrast vectors are discussed by Bretz et al. (2001).

Since the statistics T_ℓ and $T_{\ell'}$ are not necessarily independent when $\ell \neq \ell'$, we collect them in the vector $\mathbf{T} = (T_1, \ldots, T_q)^\top$, which follows, asymptotically, as $N \to \infty$, a multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix $\mathbf{R}$. Since $\mathbf{R}$ is unknown, we replace it with the estimator $\widehat{\mathbf{R}}$ obtained from standardizing $\mathbf{C}^\top \widehat{\mathbf{V}}_N \mathbf{C}$, see Konietschke et al. (2012). For large sample sizes, we reject the individual null hypothesis $H_0^{(\ell)}$ at significance level α, if $|T_\ell| \geq z_{1-\alpha}(\widehat{\mathbf{R}})$, where $z_{1-\alpha}(\widehat{\mathbf{R}})$ denotes the two-sided $(1-\alpha)$-equicoordinate quantile from the $N(\mathbf{0}, \widehat{\mathbf{R}})$ distribution. For details we refer to Konietschke et al. (2012); Umlauft et al. (2019). Compatible $(1-\alpha) \times 100\%$ simultaneous confidence intervals are obtained by $CI_\ell = \left[\mathbf{c}_\ell^\top \widehat{\boldsymbol{\psi}} \mp \frac{z_{1-\alpha}(\widehat{\mathbf{R}})}{\sqrt{N}} \sqrt{\mathbf{c}_\ell^\top \widehat{\mathbf{V}}_N \mathbf{c}_\ell} \right]$. Finally, the global null hypothesis H_0^P (or H_0^F) is rejected, if

$$T_0 = \max\{|T_1|, \ldots, |T_q|\} \geq z_{1-\alpha}(\widehat{\mathbf{R}}). \tag{17}$$

For small sample sizes, Konietschke et al. (2012) suggest to use t quantiles rather than normal and the Fisher-transformation for the computation of range-preserving confidence intervals. The `rankFD` function implements all of the different procedures.

4 Software and examples

In the following, we will analyze different data sets to illustrate the application of the implemented functions in **rankFD**. They differ in their complexity and cover two- and several samples as well as a factorial design, respectively. We note that the wrapper function `rankFD()` realizes the actual statistical design from the given `formula` argument. However, few of the statistical methods are available for two independent samples only and we therefore implemented the function `rank.two.samples` for their exclusive analysis. First, we will explain the syntax of the two functions and then illustrate their application using real data sets.

Syntax

Two samples: The `rank.two.samples()` function implements current state of the art methods for testing the null hypothesis $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \neq \frac{1}{2}$ along with the computation of $(1-\alpha) \times 100\%$ confidence intervals for θ. Its most important arguments are

```
rank.two.samples(formula, data, method = c("t.app", "logit", "probit","normal"),
permu = TRUE, alternative = c("two.sided", "less", "greater"),
wilcoxon = c("asymptotic","exact"),shift.int = TRUE,
nperm = 10000,conf.level = 0.95, info = TRUE,rounds = 4)
```

- `formula` plus `data`
  is the standard way of specifying regression relationships in R/S introduced in Chambers and Hastie (1992).
- `method`
  specifies the approximate method, where `t.app` computes the Brunner-Munzel test (Brunner

and Munzel, 2000) with t-approximation, `normal` uses the standard normal quantiles and range-preserving confidence intervals are obtained by `logit` or `probit` tranformation functions (Pauly et al., 2016).

- `permu`
  indicates whether additional studentized permutation tests shall be computed (Janssen, 1999b; Neubert and Brunner, 2007; Pauly et al., 2016)
- `alternative`
  Two-sided and one-sided tests and confidence intervals are available using the argument `alternative`.
- `wilcoxon`
  gives the option to compute additional Wilcoxon-Mann-Whitney tests for testing the equality of the two distributions $H_0^F : F_1 = F_2$ of the two samples. We use the **coin** package for these computations (Zeileis et al., 2008). Both the asymptotic as well as exact distribution of the test is available.
- `shift.int`
  can be used for the computation of a confidence interval for the shift-effect (Hodges-Lehmann).
- `nperm`, `conf.level`, `info` and `rounds`
  list optional arguments specifying the numbers of permutation, coverage probability, output explanation and decimals.

The use of the `plot()` function to a `rank.two.samples` object displays a plot of the confidence interval for θ.

Several samples and factorial designs: In addition, `rankFD()` implements statistical methods for the analysis of general nonparametric factorial designs. Its most important arguments are:

```
rankFD(formula, data, CI.method = c("logit", "normal"),
       effect = c("unweighted", "weighted"), hypothesis = c("H0F", "H0P"),
       contrast = NULL, sci.method = c("fisher", "multi.t"),
       info = TRUE, rounds=4)
```

- `formula` plus `data`
  is the standard way of specifying regression relationships in R/S introduced in Chambers and Hastie (1992).
- `CI.method`
  specifies the computational method of the confidence intervals, either using the normal approximation or the logit transformation function.
- `effect`
  defines the effect to be estimated, in particular,
  `effect = "weighted"` or `effect = "unweighted"`
  estimate the weighted or unweighted relative effect, respectively. As explained above, this choice either leads to using traditional ranks (weighted) or pseudo-ranks (unweighted).
- `hypothesis`
  defines the null hypothesis of interest (either H_0^F or H_0^P formulated in terms of distribution functions or relative effects, respectively).
- `contrast`
  is specified to perform multiple contrast tests. The argument must be given as a `list()` specifying the factor level and the kind of contrast (optional). The user can chose from a pre-implemented list of possible contrasts or commit a user-specific contrast matrix.
- `sci.method`
  defines the computational method of the simultaneous confidence intervals.
- `Factor.Information`
  is a logical argument whether descriptive information (effect estimators, standard error and confidence intervals) for each factor and interaction effect is of interest and shall be displayed.
- `info` and `rounds`
  list optional arguments specifying the numbers of output explanation and decimals.

Plot options: In order to visualize the results of the analysis, the confidence intervals can be plotted by using the generic `plot()` function (being applied to a rankFD object). In two- and higher way layouts, the user is asked to type the name of the main or interaction effect the confidence intervals of which should be drawn. All standard font, width and color arguments apply (lwd, pch, cex, etc.). Furthermore, the argument `cex.ci` sets the "cex" (number indicating the amount by which plotting text and symbols should be scaled relative to the default) of the confidence interval limits.

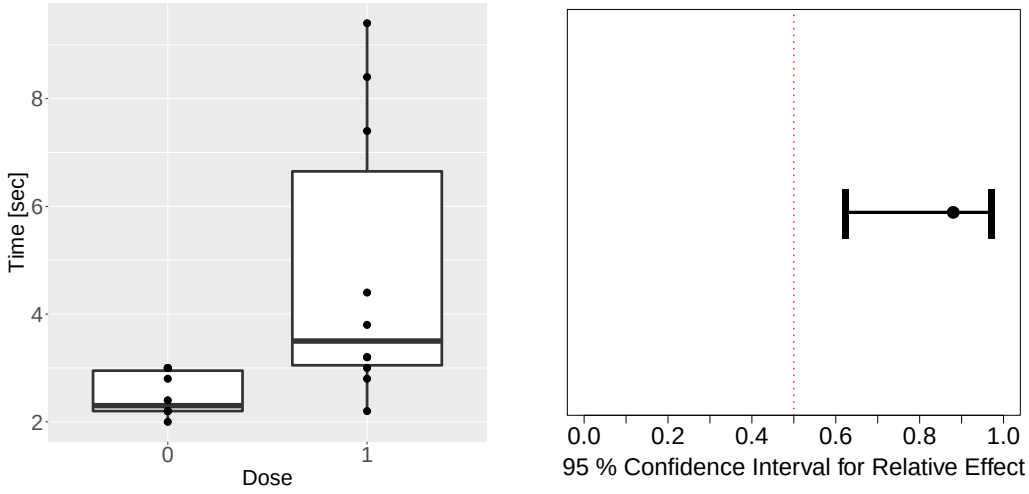

Figure 1: Boxplots (left) and 95%-confidence interval (right) for the relative effect of the reaction time data.

Two independent samples

As an illustrating example, we use a part of the reaction time data provided by Shirley (1977). In this animal experiment, $N = 40$ mice were randomized to $a = 4$ dose groups ($n = 10$ animals per group). The observations are the reaction times [in seconds] of mice to stimuli applied to their tails. Here, we only use the data from dose group 0 (negative control) and dose group 1 and thus reduce the data set to two independent samples. The boxplots of the reaction times as displayed in Figure 1 confirm our initial conjecture of quite skewed distributions. In this case, the *Wilcoxon-Mann-Whitney* effect:

$$\theta = P(X_{01} < X_{11}) + \tfrac{1}{2}P(X_{01} = X_{11}),$$

may have a better interpretation for the researcher than the difference of the two means. Recall that for $\theta < \frac{1}{2}$ the observations coming from the control group tend to be larger than those from group 1. If $\theta = \frac{1}{2}$, then none of the observations tend to be smaller or larger. *No treatment effect* is therefore indicated by $\theta = \frac{1}{2}$. The reaction time data set is analyzed with the `rank.two.samples()` function. As approximate method, we use the `logit` approach, compute the `exact` Wilcoxon-Mann-Whitney test but omit estimation of shift effects:

```
library("rankFD")
data("reaction")

A <- rank.two.samples(Time ~ Group, data = reaction, method = "logit",
+                       wilcoxon = "exact", shift.int = FALSE)


              Nonparametric Methods for 2 Independent Samples

 #Alternative: Relative Effect is unequal to 1/2
 #Method: Logit Transformation
 #Interpretation: If p(0,1) >1/2, then data in group 1 tend to be
                  larger than those in group 0
 #Confidence Level: 95 %
 #Number of permutations: 10000
 #Wilcoxon-Mann-Whitney Test: exact
 #Shift-Effect: NA
------------------------------------------------------------------------------
Call:
Time ~ Group

Descriptive:
  Sample Size
0      0   10
```

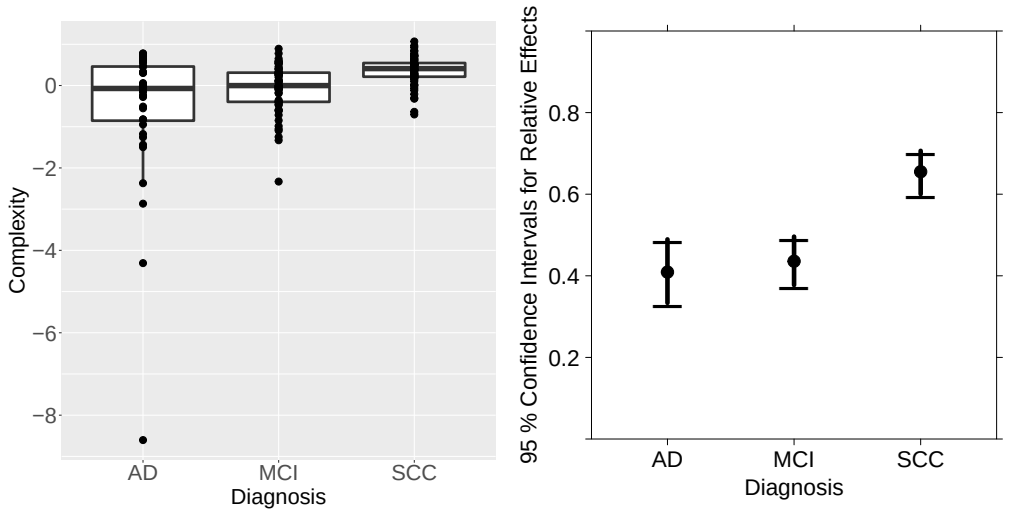

Figure 2: Boxplots (left) and (local) 95%-confidence intervals for the relative effects of the EEG values (right) .

```
1      1   10
----------------------Analysis of Relative Effects-------------------------
Test Results:
 Effect Estimator Std.Error      T  Lower  Upper p.Value
 p(0,1)      0.88    0.0801 2.6277 0.6239 0.9701  0.0086

Studentized Permutation Test:
 Effect Estimator Std.Error      T  Lower  Upper p.Value
 p(0,1)      0.88    0.0801 2.6277 0.6551 0.9673  0.0016

-------------------Analysis of Distribution Functions-----------------------

Wilcoxon-Mann-Whitney Test:
 Effect Estimator Statistic p.Value
 p(0,1)      0.88       143  0.0029

plot(A)
```

The estimated relative effect $\widehat{\theta} = 0.88$ and thus, the estimated probability that untreated mice react faster than treated ones is 88%. Furthermore, the data provide the evidence to reject $H_0^\theta : \theta = \frac{1}{2}$ at 5% level of significance ($p < 5\%$) which is also evident in the compatible confidence interval (not containing 1/2).

A one-way factorial design

As an example of a one-way factorial design we use the data set `EEG` that is included in the package **MANOVA.RM** (Friedrich et al., 2019a, 2021). The data set contains EEG measurements of 160 patients who were diagnosed with either Alzheimer's Disease (AD), mild cognitive impairments (MCI), or subjective cognitive complaints without clinically significant deficits (SCC), based on neuropsychological diagnostics (Bathke et al., 2018). For demonstration purposes, we restrict our analysis to the measurement of Hjorth complexity (represents change in frequency) obtained at central electrode positions. The question of interest is whether this EEG value tends to be larger or smaller than the mean Mann-Whitney effect across the different diseases and therefore, the relative effects defined in (3) are used for the analysis.

The EEG data is analyzed using the function `rankFD()`. Here, we calculate confidence intervals with the logit approach and estimate the unweighted relative treatment effects to test the null hypothesis H_0^p. Moreover, we specify a multiple contrast test based on Tukey-type contrasts for the pairwise comparisons of the three diagnosis groups.

```
library("MANOVA.RM")
data("EEGwide")
B <- rankFD(complexity_central ~ diagnosis, data = EEGwide,
```

```
+        CI.method = "logit", effect = "unweighted", hypothesis = "H0p",
+        contrast = list("diagnosis", "Tukey"))


Nonparametric Methods for General Factorial Designs

------------------------------------------------------------------------------
#Hypotheses: Tested in Relative Effects
#Ranking Method: Pseudo-Ranks
#Confidence Intervals: 95 % with Logit-Transformation

#MCTP: Fisher Transformation and multivariate T-Approximation
------------------------------------------------------------------------------

Call:
complexity_central ~ diagnosis

Descriptive:
   diagnosis Size Rel.Effect Std.Error  Lower  Upper
1         AD   36     0.4091    0.0400 0.3335 0.4893
2        MCI   57     0.4357    0.0304 0.3773 0.4960
3        SCC   67     0.6551    0.0272 0.6002 0.7063

Wald.Type.Statistic:
          Statistic df p-Value
diagnosis   36.2624  2       0

ANOVA.Type.Statistic:
          Statistic    df1     df2 p-Value
diagnosis   11.1605 1.6222 80.9562   2e-04

MCTP:
$Contrast.Matrix
1  2 3
C1 -1  1 0
C2 -1  0 1
C3  0 -1 1

$Local.Results
  Effect Std.Error      T   Lower  Upper p.value
C1 0.0266    0.0657 0.4044 -0.1308 0.1827  0.9119
C2 0.2460    0.0613 3.8510  0.0941 0.3868  0.0010
C3 0.2194    0.0415 5.1125  0.1176 0.3167  0.0000

$Global.Result
   T0     p.value
1 5.1125       0

$DF
[1] 46

$Quantile
[1] 2.4042

plot(B)
```

The output consists of several parts: First, a brief description of the methods is given. `B$Descriptive` returns the sample sizes, the estimated relative effects as well as their standard errors and confidence intervals for the factor levels. `B$ Wald.Type.Statistic` and `B$ANOVA.Type.Statistic` return the results of the Wald-type and ANOVA-type test as described in Section 2.3, respectively. Since we specified our null hypothesis in terms of H_0^P, Kruskal-Wallis test is not performed. The part `B$MCTP` finally contains the results of the multiple contrast test: the contrast matrix (Tukey-type), the local test results T_ℓ as well as the global results T_0 along with the t-quantile and the corresponding degrees of freedom (Konietschke et al., 2012) are reported, see Section 2.3.2 for details. The significant difference between the diagnosis groups and the results of the post-hoc tests reveal that SCC patients differ

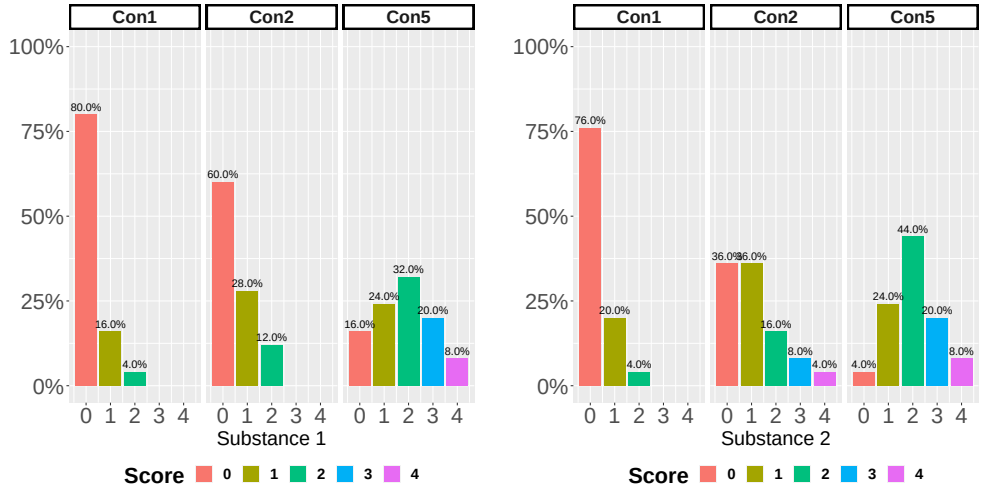

Figure 3: Barplots (percent) of the nasal mucosa scores.

significantly from the other two groups, see also Figure 2.

A two-way factorial design

As an illustrative example of a two-way factorial design, we chose the *Irritation of the Nasal Mucosa* trial provided by Brunner et al. (2019, Chapter B.3.2) and included in the package. In this trial, the researchers investigated the damage of two gaseous substances (factor A) on the nasal mucous membrane of mice. Hereby, both substances were given in three different concentrations (1[ppm], 2[ppm] and 5[ppm]) (factor B) to 25 mice each. The degree of irritation and damage was histopathologically assessed using an ordinal score ranging from 0 to 4 with 0 = "no irritation", 1 = "mild irritation", 2 = "strong irritation", 3 = "severe irritation" and 4 = "irreversible damage", respectively. The outcome is displayed in Figure 3. The code to analyze this data is similar to that provided above, but we additionally include an interaction term in the formula. In this example, we formulate the null hypothesis in terms of the distribution functions to show the R-code for testing this hypothesis. Note that due to the balanced design, both weighted and unweighted estimators give the same results.

```
data(nms)
rankFD(score ~ conc * subst, data = nms,
+          hypothesis = "H0F")


Nonparametric Methods for General Factorial Designs

-------------------------------------------------------------------------
#Hypotheses: Tested in Distribution Functions
#Ranking Method: Pseudo-Ranks
#Confidence Intervals: 95 % with Logit-Transformation

-------------------------------------------------------------------------

Call:
score ~ conc * subst

Descriptive:
  conc subst Size Rel.Effect Std.Error  Lower  Upper
1    1     1   25     0.3053    0.0310 0.2481 0.3693
2    1     2   25     0.3193    0.0320 0.2601 0.3851
3    2     1   25     0.3927    0.0386 0.3200 0.4704
4    2     2   25     0.5296    0.0459 0.4397 0.6176
5    5     1   25     0.6925    0.0429 0.6027 0.7698
6    5     2   25     0.7605    0.0310 0.6947 0.8159

Wald.Type.Statistic:
```

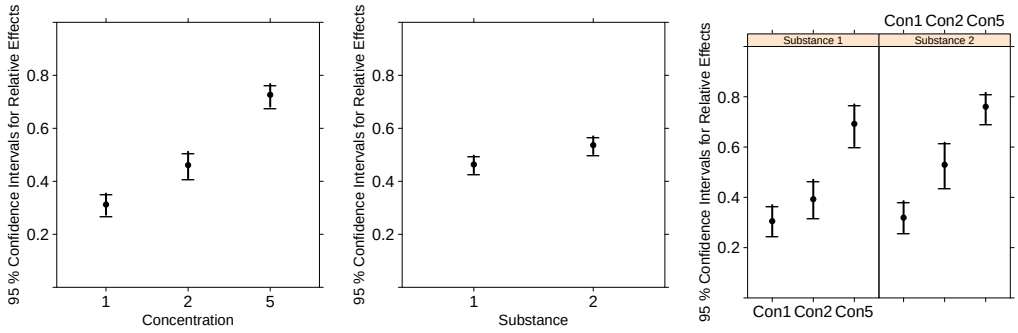

Figure 4: Local 95%-confidence interval for the (relative) main and interaction effects of the reaction time data.

```
           Statistic df p-Value
conc        114.9046  2  0.0000
subst         4.5200  1  0.0335
conc:subst    2.2174  2  0.3300

ANOVA.Type.Statistic:
           Statistic    df1   df2   p-Value
conc         49.8167 1.9289 127.0195  0.0000
subst         4.5200 1.0000 127.0195  0.0354
conc:subst    1.0741 1.9289 127.0195  0.3428
```

The right plot in Figure 4 shows that the relative effects increase at a similar rate in both levels of the main effect suggesting no qualitative interaction between the factor substance and the concentration.

5 Summary

The **rankFD**-package implements current state of the art rank methods for nonparametric inference in general factorial designs with independent observations. It comprises of functions for computing various test statistics for testing null hypotheses formulated either in distribution functions or in relative effects using ranks or pseudo-ranks, respectively. Up until now, no other software package for testing null hypotheses in relative effects in general factorial designs have existed. Besides global procedures (Wald-type and ANOVA-type statistics) using quadratic forms, **rankFD** implements multiple contrast tests and simultaneous confidence intervals for relative effects. The possibility of testing contrasts between the main and interaction effects makes **rankFD** a powerful tool for the application of nonparametric methods in data analysis and a useful addition to **nparcomp** (Konietschke et al., 2015). Besides the inference methods discussed above, **rankFD** furthermore implements formulas for computing sample sizes using the functions `WMWSSP()` and `noether()` (Happ et al., 2019). Since these methods apply for two independent samples only, we did not discuss them in the present manuscript.

We designed the package and its functions to be similar to the well known *R*-functions `lm()`, `aov()` for the analysis of linear models and the `glht()` function of the **multcomp** package for the computation of multiple contrast tests in means. Both **rankFD** and **multcomp** use the **mvtnorm** package (Genz et al., 2021) for the computation of critical values. However, as explained in detail in the Introduction, the effect measures used in multcomp and mvtnorm are different from those used in **rankFD**. In general parlance, this means that the parametric and nonparametric methods are not comparable at hand.

We plan to update **rankFD** frequently with novel procedures. For instance, various international research groups are currently investigating rank-based methods for the analysis of clustered data, see also the package **clusrank** Jiang et al. (2020) for the analysis of two samples, sample size planning, as well as analysis of covariance methods. We plan to add these methods in the future. The package **rankFD** is online available on *CRAN*.

Acknowledgement: The authors are grateful to the Editor and two anonymous referees for helpful comments which considerably improved the paper. This work was supported by the German Research Foundation project DFG KO 4680/4-1.

Bibliography

L. Acion, J. J. Peterson, S. Temple, and S. Arndt. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in medicine*, 25(4):591–602, 2006. [p142]

M. G. Akritas and S. F. Arnold. Fully nonparametric hypotheses for factorial designs i: Multivariate repeated measures designs. *Journal of the American Statistical Association*, 89(425):336–343, 1994. [p143, 145]

M. G. Akritas, S. F. Arnold, and E. Brunner. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437):258–265, 1997. [p143, 145, 148]

A. C. Bathke, S. Friedrich, M. Pauly, F. Konietschke, W. Staffen, N. Strobl, and Y. Höller. Testing mean differences among groups: Multivariate and repeated measures analysis with minimal assumptions. *Multivariate Behavioral Research*, 53(3):348–359, 2018. [p152]

Z. W. Birnbaum and O. M. Klose. Bounds for the variance of the mann-whitney statistic. *The Annals of Mathematical Statistics*, 28(4):933–945, 1957. [p142]

F. Bretz, A. Genz, and L. Hothorn. On the numerical availability of multiple comparison procedures. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 43(5):645–656, 2001. [p147, 149]

E. Brunner and U. Munzel. The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(1): 17–25, 2000. [p146, 149]

E. Brunner and M. L. Puri. Nonparametric methods in design and analysis of experiments. *Handbook of Statistics*, 13:631–703, 1996. [p143, 145]

E. Brunner, H. Dette, and A. Munk. Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, 92(440):1494–1502, 1997. [p148]

E. Brunner, F. Konietschke, M. Pauly, and M. L. Puri. Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society-Series B*, 79 (5):1463–1485, 2017. [p143, 146, 148]

E. Brunner, A. C. Bathke, and F. Konietschke. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*. Springer, 2019. [p143, 144, 146, 147, 148, 154]

E. Brunner, F. Konietschke, A. C. Bathke, and M. Pauly. Ranks and Pseudo-ranks – Surprising Results of Certain Rank Tests in Unbalanced Designs. *International Statistical Review*, 2020. [p143, 146, 148]

W. W. Burchett, A. R. Ellis, S. W. Harrar, and A. C. Bathke. Nonparametric inference for multivariate data: the r package npmv. *Journal of Statistical Software*, 76:1–18, 2017. [p143]

J. M. Chambers and T. J. Hastie, editors. *Statistical Models in S*. Chapman & Hall, London, 1992. [p149, 150]

W. J. Conover and R. L. Iman. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129, 1981. [p143]

R. B. D'Agostino, M. Campbell, and J. Greenhouse. The mann–whitney statistic: continuous use and discovery: Special papers for the 25th anniversary of statistics in medicine, 2006. [p142]

D. Dobler, S. Friedrich, and M. Pauly. Nonparametric manova in meaningful effects. *Annals of the Institute of Statistical Mathematics*, pages 1–26, 2019. [p142]

C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955. [p147]

M. A. Fligner and G. E. Policello. Robust rank procedures for the behrens-fisher problem. *Journal of the American Statistical Association*, 76(373):162–168, 1981. [p146]

S. Friedrich, F. Konietschke, and M. Pauly. Gfd: an r package for the analysis of general factorial designs. *Journal of Statistical Software*, 79:1–18, 2017. [p143]

S. Friedrich, F. Konietschke, and M. Pauly. Resampling-Based Analysis of Multivariate Data and Repeated Measures Designs with the R Package MANOVA.RM. *The R Journal*, 2(11):380–400, 2019a. doi: 10.32614/RJ-2019-051. [p152]

S. Friedrich, F. Konietschke, and M. Pauly. Resampling-based analysis of multivariate data and repeated measures designs with the r package manova. rm. *R Journal*, 11(2):380, 2019b. [p143]

S. Friedrich, F. Konietschke, and M. Pauly. **MANOVA.RM***: Resampling-Based Analysis of Multivariate Data and Repeated Measures Designs*, 2021. R package version 0.4.3. [p152]

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, B. Bornkamp, M. Maechler, and T. Hothorn. Package 'mvtnorm'. *Journal of Computational and Graphical Statistics*, 11:950–971, 2021. [p155]

M. Happ, A. C. Bathke, and E. Brunner. Optimal sample size planning for the wilcoxon-mann-whitney test. *Statistics in medicine*, 38(3):363–375, 2019. [p155]

M. Happ, G. Zimmermann, E. Brunner, and A. C. Bathke. Pseudo-ranks: How to calculate them efficiently in R. *Journal of Statistical Software, Code Snippets*, 95(1):1–22, 2020. [p144, 148]

T. P. Hettmansperger and R. M. Norton. Tests for patterned alternatives in k-sample problems. *Journal of the American Statistical Association*, 82(397):292–299, 1987. [p145]

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):346–363, 2008. [p147, 149]

A. Janssen. Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & Probability Letters*, 36(1):9–21, 1997. [p142]

A. Janssen. Nonparametric symmetry tests for statistical functionals. *Mathematical Methods of Statistics*, 8(3):320–343, 1999a. [p142]

A. Janssen. Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference*, 81(1):71–93, 1999b. [p150]

Y. Jiang, X. He, M.-L. T. Lee, B. Rosner, and J. Yan. Wilcoxon rank-based tests for clustered data with r package clusrank. *Journal of Statistical Software*, 96:1 – 26, 2020. [p155]

E. Kahler, A. Rogausch, E. Brunner, and W. Himmel. A parametric analysis of ordinal quality-of-life data can lead to erroneous results. *Journal of Clinical Epidemiology*, 61(5):475–480, 2008. [p142]

M. Kiefel, A. C. Bathke, and M. M. Kiefel. Package 'nparmd'. 2022. [p143]

F. Konietschke, L. A. Hothorn, and E. Brunner. Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, 6:738–759, 2012. [p143, 146, 148, 149, 153]

F. Konietschke, M. Placzek, F. Schaarschmidt, and L. A. Hothorn. nparcomp: an r software package for nonparametric multiple comparisons and simultaneous confidence intervals. *Journal of statistical software*, 64(1):1–17, 2015. [p155]

F. Konietschke, K. Schwab, and M. Pauly. Small sample sizes: A big data problem in high-dimensional data analysis. *Statistical Methods in Medical Research*, 30(3):687–701, 2021. [p149]

W. H. Kruskal. A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, pages 525–540, 1952. [p145]

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952. [p145]

H. B. Mann and D. R. Whitney. 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, 1947. [p142]

C. R. Mehta, N. R. Patel, and P. Senchaudhuri. Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association*, 83(404):999–1005, 1988. [p143]

K. Neubert and E. Brunner. A studentized permutation test for the non-parametric Behrens–Fisher problem. *Computational Statistics & Data Analysis*, 51(10):5192–5204, 2007. [p150]

K. Noguchi, Y. R. Gel, E. Brunner, and F. Konietschke. nparld: an r software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical software*, 50:1–23, 2012. [p143]

M. Pauly, E. Brunner, and F. Konietschke. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 77(2):461–473, 2015. [p142]

M. Pauly, T. Asendorf, and F. Konietschke. Permutation-based inference for the auc: a unified approach for continuous and discontinuous data. *Biometrical Journal*, 58(6):1319–1337, 2016. [p150]

J. Putter. The treatment of ties in some nonparametric tests. *The Annals of Mathematical Statistics*, 26(3): 368–386, 1955. [p142]

F. H. Ruymgaart. A unified approach to the asymptotic distribution theory of certain midrank statistics. In *Statistique non Parametrique Asymptotique*, pages 1–18. Springer, 1980. [p143]

E. Shirley. A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. *Biometrics*, 33(2):386–389, 1977. [p151]

Ł. Smaga. Wald-type statistics using {2}-inverses for hypothesis testing in general factorial designs. *Statistics & Probability Letters*, 107:215–220, 2015. [p142]

B. Streitberg and J. Röhmel. Exact distributions for permutation and rank tests: an introduction to some recently published algorithms. *Statistical Software Newsletter*, 12(1):10–17, 1986. [p143]

O. Thas, J. D. Neve, L. Clement, and J.-P. Ottoy. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):623–671, 2012. [p142]

M. Umlauft, M. Placzek, F. Konietschke, and M. Pauly. Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *Journal of Multivariate Analysis*, 171:176–192, 2019. [p149]

A. Zeileis, M. A. Wiel, K. Hornik, and T. Hothorn. Implementing a class of permutation tests: the **coin** package. *Journal of Statistical Software*, 28(8):1–23, 2008. [p150]

G. Zimmermann, E. Brunner, W. Brannath, M. Happ, and A. C. Bathke. Pseudo-ranks: The better way of ranking? *The American Statistician*, 76(2):124–130, 2022. [p146]

Frank Konietschke
Charité Universitätsmedizin Berlin
Institute of Biometry and Clinical Epidemiology
Reinhardstr. 58
10117 Berlin, Germany
Frank.Konietschke@charite.de

Edgar Brunner
University Medical School Göttingen
Institut für Medizinische Statistik
Humboldtallee 32
37073 Göttingen, Germany
ebrunne1@gwdg.de