# Point By Point Response to: 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*

First, we would like to thank the Editor, Associate Editor and the two anonymuous referees for your careful reading of our manuscript. We addressed all of your comments and highlighted all changes made in the manuscript in blue. In the following, we the list each comment with and our response.

## Reviewer 1

- Overview:

  The article deals with a very current topic, for which there have been some new developments in the recent years. This includes methods and also implementations in the form of R packages. In this sense, the development of rankFD provides a valuable contribution for R users as it closes a gap between existing packages (such as npmv, nparLD, nparMD, MANOVA.RM) that deal with similar problems about multivariate or longitudinal data in one-factor or two-factor designs where classical assumptions like normality cannot be made. The approach, the methods and the technology used is very consistent with the previous publications on this topic which enables a consistent follow-up of the current development of this topic.

  Nonbinding suggestion to the authors: other packages that address related problems could be mentioned in the introduction.

  – Many thanks for the nice overview and the suggestion. We mentioned other R packages (**nparLD, npmv,nparMD**) in the introduction, as suggested.

- Article:
  The topic and its motivation are pointed out very clearly in the introduction of the paper, moreover it explains the necessity why users should know the background of the introduced software package in a very understandable way. The article includes detailed reference/citation to previous pioneering publications and also R packages in the later sections which form the theoretical foundation of rankFD. The central concept of relative effects and their estimators is explained in a detailed manner, such that R users are enabled to grasp the essential difference between rankFD and conventional ANOVA. Throughout, the notation is very clear and - as already mentioned - very consistent with existing literature on this topic. The functions of the introduced package are explained extensively and in a user-friendly manner using data examples and graphical representations of the (example-related) results. As stated in the Summary of the article, no other software package covering the same issues and the same constraints has been published so far.

  – Many thanks for the summary of the paper.

- Package:
  the package has been built according to all well-known best practices. Peculiarities of the code are due to elaborate methods.

  - Many thanks for your careful reading of our manuscript and checking the code. We implemented the package in such a way that it can be used as **nparLD, multcomp** and other related R packages.

## Reviewer 2

- It would be beneficial to list related R packages that use nonparametric methods in one- and multi-way design and in the multivariate setting so that whoever is reading this paper can also learn that there are other carefully crafted (not "ad hoc") packages out there. In other words, this will also answer the central question: "Does anybody know whether there is a nonparametric analog of ANOVA?". Also, mentioning such packages is important for the R community.

  - Many thanks for the valuable comment. We added mentioned related R packages in the Introduction.

- It seems to be the case that unweighted relative effects are preferred to the weighted relative effects based on the discussion on page 3 (and also on page 4). However, that can be more explicitly stated, rather than saying that the weighted relative effects may lead to surprising results in the unbalanced design. Similarly, it appears that the same kind of discussion with regards to the weighted vs. unweighted relative effects was repeated for the multi-way layout on page 5. I am under the impression that this is an important topic to consider in nonparametric inference, perhaps from the theoretical point of view. From the practical point of view, would you also recommend always using the unweighted relative effects? Please clarify.

  - Many thanks for the important question. Indeed, it is not that easy to answer. First, balanced samples (equal) sample sizes are certainly the most convenient situation, because then the unweighted and weighted relative effects and their empirical counterparts are identical. That is not the case when the sample sizes are different. The weighted relative effects depend on the sample sizes and their allocations, and, since classical rank methods (Kruskal-Wallis test etc) base on them, their power and actual test result depend on the allocation of interest. For practical situations, this implies that one may find a sample size configuration to gather a significant result, while the test would not be significant under a different sample size allocation.
    Throughout the manuscript we aimed to be neutral and to mention potential drawbacks of the methods without giving clear guidance and recommendation. Otherwise, one would argue why we implemented all the methods if we are not convinced from using them. We implemented all methods for the sake of completeness.

- On page 4, it states "...equal distribution functions imply equal variances..." However, this is true only when the second moment exists.

  - Thank you, we added the remark.

- On page 6 (Confidence intervals), please define $z_{1-\alpha/2}$. Moreover, is it correct that the confidence intervals presented in the articles need to have a $H_0^p$-type hypothesis for the corresponding test? Or, is it possible to have a $H_0^F$-type hypothesis? Please clarify.

  - Thank you, we added definition of $z_{1-\alpha/2}$. You are actually correct. For computing a confidence interval for the relative effect $\theta$, one needs a $H_0^p$-type hypothesis. The reason is that the Wilcoxon-Mann-Whitney test uses a variance estimator that is only consistent under the null hypothesis $H_0^F$. Therefore, we cannot invert this statistic to obtain a confidence interval for $\theta$. We added a remark to the respective Section on page 6.

- On page 7 (Global test procedures), it says "the ANOVA-type statistic $A_N(\mathbf{C})$ controls the type-I error much better". Do you have any suggestion for the minimal sample size necessary to make this approximation work? I understand that giving such a recommendation may be dangerous, but it could also be worthwhile for practitioners. In addition, this is perhaps where one could emphasize advantage(s) of the permutation method.

  – We added a recommendation to the manuscript. As you said, since we cover general factorial designs, we need to be cautious in providing a general recommendation. In the global testing scenario (several sample case), there actually is no permutation test available for testing the general nonparametric null hypothesis $H_0^P$. We implement a permutation test for the analysis of two samples only. We therefore did not mention permutation tests in the respective section.

- On page 8 (multiple contrast test procedures), the correlation matrix $\mathbf{R}$ is a theoretical quantity so that it must be estimated in order to compute the two-sided $(1-\alpha)$-equicoordinate quantile in practice. To make that more explicit, it may be better to use $\widehat{\mathbf{R}}$ instead.

  – Many thanks for pointing this out. We have changed it accordingly.

- Illustration: Is there a particular reason for using "H0P" in the one-way factorial design example, and "H0F" in the two-way factorial design example?

  – Both examples are just for illustration and we randomly picked "H0P" for the one-way and "H0F" for the two-way design example. Since the paper is pretty long already, we prefer to not add both versions to each example.

- A two-way factorial design: The authors conclude that there are "significant effects of both factors, but no significant interaction". Even though this conclusion is clear from p-values, could Figure 4 help the readers better understand visually why we have these conclusions? For example, for the main effects (Concentration and Substance), the left and middle plot show different levels of relative effects with relatively narrow confidence intervals. On the other hand, the right plot shows that the relative effects go up in a similar manner (in parallel). Could that be an indication of the presence of main effects and absence of interaction effect? Please clarify.

  – Many thanks for the question. The interpretation of the results is very similar to the one in linear models in means. If there is no interaction, then the main effects can be interpret as "main effects" and just being caused from the actual condition. The software implements all three graphics, one for each main effect plus the interaction. In principle, the right plot shows that the relative effects go up similarly (in parallel, say). This indicates that there is no interaction. We therefore can interpret the main effects in the remaining panels. The intervals do not overlap indicating significant main effects. We added a remark to the corresponding section.

- A one-way factorial design: Please describe in more detail what contrast = list("diagnosis", "Tukey") is doing. This seems somewhat different from what the help file of the package says.

  – We apologize if there is confusion. The contrast argument sets the factor (in the running example its "diagnosis") on the levels of which a Tukey contrast should be computed, i.e. all-pairs comparisons on the levels of "diagnosis". We are somewhat unsure to what extent the help file differs. We just checked the version available on CRAN. It is the one we are workig with. Anyway, we noticed that the contrast statement is not used in any of the examples listed in the help files. We updated them and will update the package on CRAN soon.

- Software: Is it easy to extract $\mathbf{C}\widehat{\boldsymbol{\psi}}$ from the package? Many journals require reporting effect size along with test statistic and confidence interval, so a demonstration of how $\mathbf{C}\widehat{\boldsymbol{\psi}}$ can be obtained would be helpful.
  - The estimators of the effects are contained in the output and are listed within the MCTP, see the table "Local.Results" (column Effect):

```
library("MANOVA.RM")
data("EEGwide")
B <- rankFD(complexity_central ~ diagnosis, data = EEGwide,
+ CI.method = "logit", effect = "unweighted", hypothesis = "H0p"
+ contrast = list("diagnosis", "Tukey"))
B[5]$MCTP$Local.Results

Effect   Std.Error    T   Lower  Upper p.value
C1 0.0266    0.0657 0.4044 -0.1308 0.1826  0.9119
C2 0.2460    0.0613 3.8510  0.0941 0.3867  0.0010
C3 0.2194    0.0415 5.1125  0.1176 0.3166  0.0000
```