Supplementary Material to the paper Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects

Edgar Brunner[1,*], Frank Konietschke[2], Markus Pauly[3] and Madan L. Puri[4]

All authors are given in alphabetic order:
[1] Department of Medical Statistics, University of Göttingen, Germany
[2] Department of Mathematical Sciences, University of Texas, Dallas, USA
[3] Institute of Statistics, University of Ulm, Germany
[4] Department of Mathematics, Indiana University, Bloomington IN, USA

1 Extensions to General Repeated Measures Designs

Some of the results from Section 5 are copied here for the readers convenience.

Let us consider a general nonparametric factorial repeated measures designs given by independent random vectors

$$\boldsymbol{X}_{ik} = (X_{i\ell k})_{\ell=1}^{t} = (X_{i1k}, \ldots, X_{itk})', \quad i = 1, \ldots, d;\ k = 1, \ldots, n_i;\ \ell = 1, \ldots, t \tag{1.1}$$

representing the $t \in \mathbb{N}$ repeated measurements on subject k in group i. As in the paper a factorial structure on the groups (whole-plot / between-subjects factors) and repeated measures (sub-plot / within-subjects factors) can be included by splitting up the indices i and ℓ, respectively. Also in this setting we can define adequate model parameters on the marginals of $X_{i\ell 1} \sim F_{i\ell}$. In particular, these are given by the relative effect of the distribution of group i at time ℓ with respect to the unweighted pooled distribution function $G = \frac{1}{dt}\sum_{i=1}^{d}\sum_{\ell=1}^{t} F_{i\ell}$

$$p_{i\ell} \quad = \quad \int G dF_{i\ell} \quad i = 1, \ldots, d;\ \ell = 1, \ldots, t. \tag{1.2}$$

It can again be written as the mean $p_{i\ell} = \overline{w}_{\cdot\cdot i\ell}$ of the relative marginal effects $w_{rsi\ell} = \int F_{rs} dF_{i\ell}, 1 \leq i, r \leq d, 1 \leq s, \ell \leq t$, Collecting all $p_{i\ell}$ in a vector $\boldsymbol{p} = (p_{11}, p_{12}, \ldots, p_{dt})'$ our linear hypotheses of interest can be written as $H_0^p : \boldsymbol{C}\boldsymbol{p} = \boldsymbol{0}$ for an adequate hypothesis matrices $\boldsymbol{C}$.

For testing H_0^p estimates for the effects $p_{i\ell}$ are obtained by substituting the distribution functions $F_{i\ell}(x)$ in (1.2) with their empirical counterparts $\widehat{F}_{i\ell}(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c(x - X_{i\ell k})$ yielding

$$\widehat{p}_{i\ell} \quad = \quad \int \widehat{G}_t d\widehat{F}_{i\ell}. \tag{1.3}$$

Thus, an estimator for the vector $\boldsymbol{p}$ is given by $\widehat{\boldsymbol{p}} = (\widehat{p}_{11}, \widehat{p}_{12}, \ldots, \widehat{p}_{dt})'$ and its asymptotic behaviour can be studied similar to the univariate case. In particular, defining the vector

$$\boldsymbol{w}_{rs} = (w_{rs11}, w_{rs12}, \ldots, w_{rsdt})'$$

and the matrix

$$\boldsymbol{W} = (\boldsymbol{w}_{11} \vdots \boldsymbol{w}_{12} \vdots \ldots \vdots \boldsymbol{w}_{dt}) \in \mathbb{R}^{dt \times dt}$$

and denoting their empirical counterparts as $\widehat{\boldsymbol{w}}_{rs}$ and $\widehat{\boldsymbol{W}}$, respectively, it holds that

$$\boldsymbol{p} = \boldsymbol{E}_{dt}\text{vec}(\boldsymbol{W}) \quad \text{and} \quad \widehat{\boldsymbol{p}} = \boldsymbol{E}_{dt}\text{vec}(\widehat{\boldsymbol{W}}). \tag{1.4}$$

Here vec denotes the usual matrix operator which stacks the columns of a matrix on top of each other and the matrix $\boldsymbol{E}_{dt}$ is given by

$$\boldsymbol{E}_{dt} = \frac{1}{dt}\boldsymbol{1}_{dt}' \otimes \boldsymbol{I}_{dt}.$$

Thus, by the asymptotic equivalence theorem, the random vector $\sqrt{N}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ has the same asymptotic distribution as

$$\sqrt{N}\boldsymbol{E}_{dt}\boldsymbol{Z}.$$

Here $\boldsymbol{Z} = (\boldsymbol{Z}_{11}', \boldsymbol{Z}_{12}', \ldots, \boldsymbol{Z}_{dt}')'$ with $\boldsymbol{Z}_{i\ell} = (Z_{11i\ell}, Z_{12i\ell}, \ldots, Z_{dti\ell})'$ and

$$Z_{rsi\ell} = \frac{1}{n_i}\sum_{k=1}^{n_i}[F_{rs}(X_{i\ell k}) - w_{rsi\ell}] - \frac{1}{n_r}\sum_{k=1}^{n_r}[F_{i\ell}(X_{rsk}) - w_{i\ell rs}]$$

denote sums of independent random variables. From this expression the following central limit theorem follows.

THEOREM 1.1 *Let* $\boldsymbol{V}_N = \boldsymbol{E}_{dt}\, Cov(\sqrt{N}\boldsymbol{Z})\boldsymbol{E}_{dt}$. *Then* $\sqrt{N}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ *is asymptotically multivariate normally distributed with expectaion* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{V}_N = \boldsymbol{E}_{dt}\, Cov(\sqrt{N}\boldsymbol{Z})\boldsymbol{E}'_{dt}$.

Proof.

To apply the Cramer-Wold device let $\boldsymbol{k} = (k_{11}, \ldots, k_{dt})'$ denote an arbitrary vector of constants with $||\boldsymbol{k}|| = 1$. It follows from the asymptotic equivalence result stated above that $\sqrt{N}\boldsymbol{k}'(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ is asymptotically equivalent to

$$\begin{aligned}
&\sqrt{N}\sum_{i=1}^{d}\sum_{\ell=1}^{t}\left(\frac{1}{n_i}\sum_{k=1}^{n_i}k_{i\ell}(G(X_{i\ell k}) - F_{i\ell}(X_{i\ell k})) - \frac{1}{dt}\sum_{\substack{r=1\\(r,s)\neq(i,\ell)}}^{d}\sum_{s=1}^{t}\frac{1}{n_r}\sum_{j=1}^{n_r}k_{i\ell}F_{i\ell}(X_{rsj}) + k_{i\ell}(1-2p_{i\ell})\right)\\
= \;&\sqrt{N}\sum_{i=1}^{d}\sum_{\ell=1}^{t}\left(\frac{1}{n_i}\sum_{k=1}^{n_i}k_{i\ell}(G(X_{i\ell k}) - F_{i\ell}(X_{i\ell k})) - \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{1}{dt}\sum_{\substack{r=1\\(r,s)\neq(i,\ell)}}^{d}\sum_{s=1}^{t}k_{rs}F_{rs}(X_{i\ell k}) + k_{i\ell}(1-2p_{i\ell})\right)\\
= \;&\sum_{i=1}^{d}\frac{\sqrt{N}}{n_i}\sum_{k=1}^{n_i}\tilde{Z}_{ik},
\end{aligned}$$

where

$$\tilde{Z}_{ik} = \sum_{\ell=1}^{t}\left(k_{i\ell}(G(X_{i\ell k}) - F_{i\ell}(X_{i\ell k})) - \frac{1}{dt}\sum_{\substack{r=1\\(r,s)\neq(i,\ell)}}^{d}\sum_{s=1}^{t}k_{rs}F_{rs}(X_{i\ell k}) + k_{i\ell}(1-2p_{i\ell})\right)$$

are independent random variables with expectation zero. Since these random variables are uniformly bounded it follows from the Lindeberg-Feller central limit theorem and the Cramer-Wold device that $\sqrt{N}(\widehat{\boldsymbol{p}}-\boldsymbol{p})$ is asymptotically multivariate normally distributed with expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{V}_N = \boldsymbol{E}_{dt}\, Cov(\sqrt{N}\boldsymbol{Z})\boldsymbol{E}'_{dt}$.

Since the involved covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{rsi\ell}) \equiv Cov(\sqrt{N}\boldsymbol{Z})$ is unknown we have to estimate it. Therefore, we first analyze its explicit form and proceed as in Placzek (2013). First, consider the case $i = r$ and $s = \ell$ and set

$$\Sigma_{rsrs} = \left(N\, Cov(Z_{pqrs}, Z_{p'q'rs})\right)_{1\leq p,p'\leq d, 1\leq q,q'\leq t} \equiv \left(\sigma_{rs}(p,q,p',q')\right)_{1\leq p,p'\leq d, 1\leq q,q'\leq t}.$$

Since $Z_{rsrs} = 0$, $Z_{rsi\ell} = -Z_{i\ell rs}$ and $X_{i\ell k}$ is independent from $X_{i'\ell'k'}$ for all $i \neq i'$ or $k \neq k'$ it follows that $\sigma_{rs}(p,q,p',q') =$

$$\begin{cases}
\tau_r^{(s,s)}(p,q,p',q') & r \neq p, p' \wedge p \neq p'\\
\tau_r^{(s,s)}(p,q,p,q') + \tau_p^{(q,q')}(r,s,r,s) & r \neq p, p' \wedge p = p'\\
\tau_r^{(s,s)}(r,q,p',q') - \tau_r^{(q,s)}(r,s,p',q') & if \quad r = p \wedge p' \neq p' \wedge q \neq s\\
\tau_r^{(s,s)}(p,q,r,q') - \tau_r^{(s,q')}(p,q,r,s) & r = p \wedge p' \neq p \wedge q' \neq s\\
\tau_r^{(s,s)}(p,q,r,q') - \tau_r^{(s,q')}(r,q,r,s) - \tau_r^{(q,s)}(r,s,r,q') + \tau_r^{(q,q')}(r,s,r,s) & r = p = p' \wedge q \neq s \wedge q' \neq s\\
0 & \text{else.}
\end{cases}$$

Here

$$\tau_r^{(s,\ell)}(p,q,p',q') \;=\; \frac{N}{n_r} E\left[\left(F_{pq}(X_{rs1}) - w_{pqrs}\right)\left(F_{p'q'}(X_{r\ell 1}) - w_{p'q'r\ell}\right)\right]. \tag{1.5}$$

Now, consider the case $(r,s) \neq (i,\ell)$ and set

$$\Sigma_{rsi\ell} = \left(N\ Cov(Z_{pqrs}, Z_{p'q'i\ell})\right)_{1\leq p,p'\leq d, 1\leq q,q'\leq t} \equiv \left(\sigma_{rsi\ell}(p,q,p',q')\right)_{1\leq p,p'\leq d, 1\leq q,q'\leq t}.$$

From similar considerations as above it follows for each entry that $\sigma_{rsi\ell}(p,q,p',q') =$

$$\begin{cases}
\tau_r^{(s,\ell)}(p,q,p',q') & r=i \wedge p\neq i, p'\wedge r\neq p' \\
-\tau_r^{(s,q')}(p,q,i,\ell) & r=p' \wedge p\neq i, p'\wedge r\neq i \\
-\tau_p^{(q,q')}(r,s.i,\ell) & p=i \wedge r\neq i, p'\wedge p\neq p' \\
\tau_p^{(q,q')}(r,s,i,\ell) & p=p' \wedge r\neq i, p'\wedge p\neq i \\
\tau_r^{(s,\ell)}(p,q,r,q') - \tau_r^{(s,q')}(p,q,r,j) & r=i=p' \wedge p\neq i, p'\wedge q'\neq \ell \\
-\tau_p^{(q,j)}(r,s,p,q') + \tau_p^{(q,q')}(r,s,i,\ell) & p=i=p' \wedge r\neq i, p'\wedge q'\neq \ell \\
\tau_r^{(s,\ell)}(r,q,p',q') - \tau_r^{(q,\ell)}(r,s,p',q') & r=i=p \wedge p'\neq i, p\wedge q\neq s \\
-\tau_r^{(s,q')}(r,q,i,\ell) + \tau_r^{(q,q')}(r,s,i,\ell) & p=r=p' \wedge i\neq r, p\wedge q\neq s \\
\tau_r^{(s,\ell)}(p,q,p,q') + \tau_p^{(q,q')}(r,s,r,j) & r=i\wedge p=p' \wedge r\neq p' \wedge p\neq i \\
-\tau_r^{(s,q')}(p,q,p,\ell) - \tau_p^{(q,\ell)}(r,s,r,q') & r=p' \wedge p=i\wedge r\neq i\wedge p\neq p' \\
\tau_r^{(s,\ell)}(r,q,r,q') - \tau_r^{(s,q')}(r,q,r,\ell) - \tau_r^{(q,\ell)}(r,s,r,q') + \tau_r^{(q,q')}(r,s,r,\ell) & r=p=p'=i\wedge q\neq s\wedge q'\neq \\
 & \ell\wedge q\neq \ell\wedge s\neq q'\wedge q\neq q' \\
0 & else
\end{cases}$$

Thus, for estimating the unknown covariance $\boldsymbol{V}_N$ we only have to estimate the unknown quantities given in (1.5). Similar to the paper consistent estimaors $\widehat{\tau}_r^{(s,\ell)}(p,q,p',q')$ are obtained by calculating the arithmetic means of the empirical counterparts of (1.5). This yields a consistent estimator $\widehat{\boldsymbol{V}}_N$ of $\boldsymbol{V}$ and an ANOVA-type-statistic for H_0^p is given by

$$Q_N(\boldsymbol{C}) \;=\; Q_N(\boldsymbol{T}) \;=\; \frac{N}{\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{V}}_N)} \widehat{\boldsymbol{p}}'\boldsymbol{T}\widehat{\boldsymbol{p}}, \tag{1.6}$$

where again $\boldsymbol{T} = \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^+\boldsymbol{C}$ is the unique projection matrix on the column space of $\boldsymbol{C}$, see e.g. Brunner et al. (1997) or Brunner and Puri (2001). As in Theorem 4.1 of the paper $Q_N(\boldsymbol{C})$ has, asymptotically under the null $H_0^p : \boldsymbol{T}\boldsymbol{p} = \boldsymbol{0}$, the same distribution as a weighted sum of independent χ_1^2-distributed random variables. An *ANOVA-eigen-type-p-test* can be obtained by estimating the unknown weights using the consistent matrix estimator $\widehat{\boldsymbol{V}}_N$. The investigation of this approach will be part of future work together with a simultaneous inference procedure.

2 Interpretation of the Nonparametric Effects

Here we outline an interpretation of the nonparametric effects by using a decomposition of the distribution functions as in Akritas and Arnold (1994). It is similar to the interpretation for the relative effects considered in the supporting information in de Neve and Thas (2015). For ease of convenience we only consider the situation of a crossed two-way layout. To this end, write

$$F_{ij} = G + A_i + B_j + (AB)_{ij} \qquad (i=1,\ldots,a; j=1,\ldots,b)$$

for functions $G, A_i, B_j, (AB)_{ij}$ satisfying $\sum_{i=1}^{a} A_i = \sum_{j=1}^{b} B_j = 0$, $\sum_{i=1}^{a}(AB)_{ij} = 0$ for all $j = 1, \ldots, b$ and $\sum_{j=1}^{b}(AB)_{ij} = 0$ for all $i = 1, \ldots, a$. This expression is related to the classical mean decomposition in linear models. In particular, we can write $G = \overline{F}_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} F_{ij}$, $A_i = \overline{F}_{i\cdot} - G = \frac{1}{b}\sum_{j=1}^{b} F_{ij} - G$, $B_j = \overline{F}_{\cdot j} - G = \frac{1}{a}\sum_{i=1}^{a} F_{ij} - G$ and $(AB)_{ij} = F_{ij} - \overline{F}_{i\cdot} - \overline{F}_{\cdot j} + G$, for $i = 1, \ldots, a, j = 1, \ldots, b$. Inserting the above decomposition into the nonparametric effects p_{ij} now results in

$$p_{ij} \;=\; \int G dF_{ij} \;=\; \frac{1}{2} + \underbrace{\int G dA_i}_{\alpha_i} + \underbrace{\int G dB_j}_{\beta_j} + \underbrace{\int G d(AB)_{ij}}_{(\alpha\beta)_{ij}}.$$

Here the additive effects all fulfill the side conditions $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a}(\alpha\beta)_{ij} = 0$ for all $j = 1, \ldots, b$ and $\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$ for all $i = 1, \ldots, a$ which they inherit from the corresponding functions. Thus, as in the supporting information in de Neve and Thas (2015), we can interpret the additive effect α_i as

$$\alpha_i = \int G dF_{i\cdot} - \frac{1}{2} = P(Z_G < Z_{ai}) + \frac{1}{2}P(Z_G = Z_{ai}) - \frac{1}{2},$$

where $Z_G \sim G$ and $Z_{ai} \sim \overline{F}_{i\cdot}$. Similar interpretations hold for β_j (see the paper for details) and $(\alpha\beta)_{ij}$, respectively.

3 More Details on the Analysis of the Data Example

Some parts from Section 8 are copied here for the readers convenience.

In order to provide freely available software for data analysis and educational purposes we implemented an *R* software package called **rankFD** for rank based analysis of factorial designs with independent observations. For a user-friendly implementation it is equipped with a graphical user interface. The package contains the ANOVA-type-p-test (who turned out to be the best in our simulation study) for making inference in one-, two- or arbitrary higher-way layouts as well specific nested designs. Furthermore, all test procedures for testing the hypothesis H_0^F formulated in terms of the distribution functions are implemented. Besides of a descriptive overview it also provides p-values and confidence intervals for the main treatment effects along with plotting options. The *R* package will be updated regularly. The *R*-package is freely available at CRAN. Here it has been exemplified for analysing the motivating data example described in the Introduction of the paper and in more detail below.

In a placebo-controlled trial, the effect of a drug on the immune system was examined under consideration of stress (food deprivation) using 40 mice. A main response variable was the number of leucocytes migrating into the peritoneum. Half of the mice received a diet low in protein, the other half received normal food. One day before opening the peritoneum, 20 mice in each group received an injection with the drug, while the other 20 received an equal amount placebo. Eight hours later, migration of leucocytes was stimulated by injecting glycogen into every mouse. Then, for the resulting four groups, the number of leucocytes (among other attributes) was determined for each mouse. Because of copy right and confidentiality reasons only a part of the data from the complete trial is given in Table 1. We are grateful to Fa. Schaper & Brümmer (Salzgitter) for making available these data from a common research project.

Table 1: Number of Leukocytes $[10^6/ml]$ for 40 mice. All combinations of the following two treatments were examined: normal diet vs. low protein diet and drug vs. placebo.

Number of Leukocytes $[10^6/ml]$			
Normal Food		Reduced Food	
Placebo	Drug	Placebo	Drug
7.5	15.9	7.5	5.7
8.1	12.0	5.7	8.1
5.4	12.3	3.3	6.0
6.0	44.4	3.9	6.0
16.2	13.5	3.9	11.4
7.8	19.8	6.6	5.1
8.1	15.3	6.3	11.1
5.7	32.7	3.3	12.9
6.9	18.0	4.5	5.4
5.1	15.0	4.2	8.4

Applying the *R*-package **rankFD** to the above data set yields the following statistics and p-values for testing the main effects A (*food condition*) and B (*treatment*) as well as the interaction AB between the food condition and the treatment shown in Table 2.

Table 2: Analysis of the data example with the ANOVA-type$-p$-test φ_N given in Theorem 4.2b)(3). The value of the test statistic $Q_N(\boldsymbol{T})$ is compared with the quantile of an F-distribution with estimated degrees of freedom $\widehat{f}_1$ and $\widehat{f}_2$.

Factor	Statistic	$\widehat{f}_1$	$\widehat{f}_2$	p-value
Food Condition	42.450	1	26.492	< 0.0001
Treatment	33.191	1	26.492	< 0.0001
Interaction	1.868	1	26.492	0.1832

It appears from Table 2 that both the factors *Food* as well as *Treatment* have a significant impact on the numbers of leucocytes at 5% level. The data do not provide any evidence for an interaction between the treatment and the food condition.

So far the state of the art nonparametric approach using ranks to test the null hypotheses of no treatment effects or no interaction would have been using the procedures based on the distribution functions $F_{ij}(x)$, i.e. $H_0^F : \boldsymbol{TF} = \boldsymbol{0}$, where $\boldsymbol{T}$ denotes an appropriate contrast matrix (for details see Akritas et al., 1997). the hypothesis of no food effect would be written as $H_0^F(A) : F_{11} + F_{12} - F_{21} - F_{22} \equiv 0$. Here the index i in F_{ij} refers to the factor A (food condition: $i = 1$, normal food; $i = 2$, reduced food) while the second index j refers to the factor B (treatment: $j = 1$, placebo; $j = 2$, drug). A rejection or acception of these hypotheses would help for a first intuition about underlying effects, however, the testing procedures would not help for deducing the same elaborated interpretations and conclusions as done with the unweighted relative effects p_{ij} in Section 8 of the paper. The only possibility for more intuition would be to plot the empirical versions of the sums and differences of distribution functions defining the hypotheses. To demonstrate this we plot the so-called empirical interaction function $x \mapsto (\widehat{AB}_{11})(x) = \frac{1}{4}[\widehat{F}_{11}(x) - \widehat{F}_{12}(x) - \widehat{F}_{21}(x) + \widehat{F}_{22}(x)]$ and the empirical

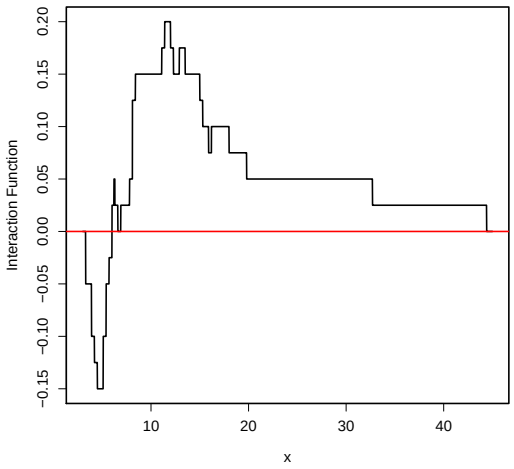

Figure 1: Plot of the empirical interaction function $(\widehat{AB}_{11}) : x \mapsto \frac{1}{4}[\widehat{F}_{11}(x) - \widehat{F}_{12}(x) - \widehat{F}_{21}(x) + \widehat{F}_{22}(x)]$

main effect functions $x \mapsto (\widehat{A}_1)(x) = \frac{1}{4}[\widehat{F}_{11}(x) + \widehat{F}_{12}(x) - \widehat{F}_{21}(x) - \widehat{F}_{22}(x)]$ and $x \mapsto (\widehat{B}_1)(x) = \frac{1}{4}[\widehat{F}_{11}(x) - \widehat{F}_{12}(x) + \widehat{F}_{21}(x) - \widehat{F}_{22}(x)]$ in Figures 1-3 below.

From Figures 2 and 3 it is obvious that the main effect functions are different from the 0-function. But this is also true for the plot of the empirical interaction function in Figure 1. No intuitive conclusion regarding an interaction can be drawn from this figure. This demonstrates the gap between the hypotheses of the procedures based on H_0^F and the set of alternatives for which they are consistent. One must note that the above main and interaction effects defined by the distribution functions are functional-valued quantities which are difficult to interpret.

This is different, however, for the nonparametric effects $p_{ij} = \int G dF_{ij}$ considered in the main paper. For these real-valued effect measures point estimators for each drug×food combination can be computed. Also two-sided (range preserving) 95%-confidence intervals for the p_{ij} are computed where the logit transformation $g(x) = \log(x/(1-x))$ has been used. The results are listed in Table 3 and displayed in Figure 4.

Table 3: Estimates and 95%-confidence intervals for the nonparametric treatment effects $p_{ij} = \int G dF_{ij}$ in the leucocytes trial. The index i refers to the food condition while the index j refers to the treatment. The range-preserving limit are obtained by the logit-transformation $g(x) = \log(x/(1-x))$.

Factor Level Combination		Sample Size	Effect	95%-Confidence Limits	
Food Condition	Treatment	n_{ij}	$\widehat{p}_{ij}$	Lower	Upper
$i = 1$ - Normal	$j = 1$ - Placebo	10	0.460	0.355	0.568
$i = 1$ - Normal	$j = 2$ - Drug	10	0.855	0.818	0.885
$i = 2$ - Reduced	$j = 1$ - Placebo	10	0.209	0.140	0.301
$i = 2$ - Reduced	$j = 2$ - Drug	10	0.476	0.375	0.579

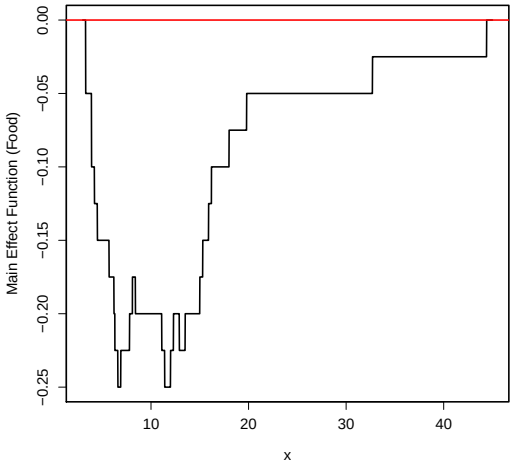

Figure 2: Plot of the empirical main effect function $(\widehat{A}_1) : x \mapsto \frac{1}{4}[\widehat{F}_{11}(x) + \widehat{F}_{12}(x) - \widehat{F}_{21}(x) - \widehat{F}_{22}(x)]$

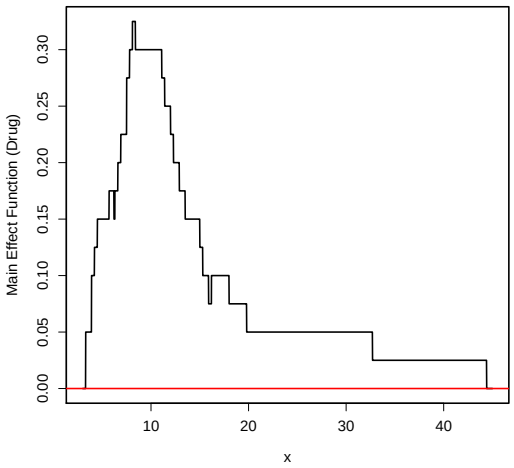

Figure 3: Plot of the empirical main effect function $(\widehat{B}_1) : x \mapsto \frac{1}{4}[\widehat{F}_{11}(x) - \widehat{F}_{12}(x) + \widehat{F}_{21}(x) - \widehat{F}_{22}(x)]$

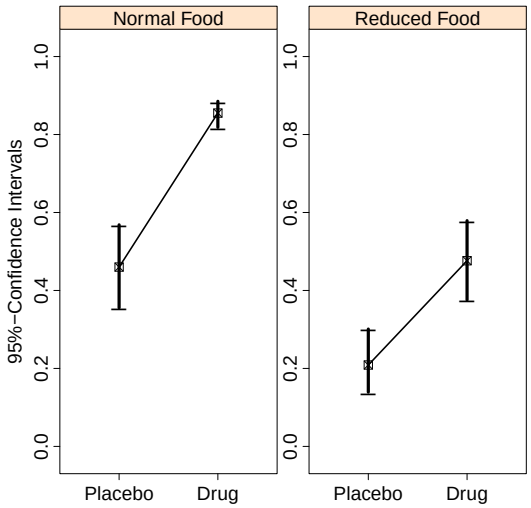

Figure 4: Plot of the 95% confidence intervals for each drug$\times$ food combination.

The effect $\widehat{p}_{21} = 0.209$ for the reduced food under placebo means that the observations from F_{21} tend to be smaller than those from the mean distribution $G = \frac{1}{4}\sum_{i,j=1}^{2} F_{ij}$, or more precisely, the probability that a randomly selected observation Z from the mean distribution G is smaller than a randomly selected observation X_{21} from F_{21} equals 0.209. Similarly, the effect $\widehat{p}_{12} = 0.855$ for the normal food under the drug means that the observations from F_{12} tend to be larger than those from the mean distribution G. We note that the confidence intervals for Placebo and Drug do not overlap within each food condition which may be interpreted that the drug is effective in both cases, as seen from Figure 4. Such an interpretation is difficult to conclude from plots of the empirical effect functions.

4 More Simulation Results

Again most parts from Section 7 are copied here for the readers convenience.

Here we investigate the small sample properties of the three statistical tests $\widehat{\varphi}_N$, $\widetilde{\varphi}_N$, and φ_N based on the ANOVA-type statistic $Q_N(\boldsymbol{T})$ and given in Theorem 4.2.(b) within extensive simulation studies with regard to their

(a) maintenance of the preassigned type I error level ($\alpha = 5\%$) under the hypothesis $H_0^p(\boldsymbol{T}) : \boldsymbol{Tp} = \boldsymbol{0}$ and

(b) their powers to detect specific alternatives.

All simulations were performed using *R* (version 2.15.0, R Development Core Team, 2010) with $nsim = 10,000$ simulation runs for each setting. As in the main paper the distribution of $\widehat{Q}(\boldsymbol{T})$ was approximated using $n_{MC} = 10,000$ Monte-Carlo runs, and the critical values were estimated from this distribution. Hereby, the eigenvalues of the matrix $\boldsymbol{T}\widehat{\boldsymbol{V}}_N$ were computed with the base *R*-function *eigen*.

In order to compare the newly developed methods with other procedures we first restrict our considerations to the one-way layout (balanced and unbalanced) with $a = 4$ independent treatment groups,

and by using both symmetric and skewed distributions. In this set-up the above procedures test the null hypothesis $H_0^p : p_1 = p_2 = p_3 = p_4$. As competitors the classical Kruskal-Wallis rank test and two Wald-type tests are considered: The test $\varrho_N = \mathbf{1}\{W_N(\boldsymbol{C}) > \chi^2_{1-\alpha;r(\widehat{\boldsymbol{M}}_N)}\}$ based on the WTS given in Section 4 of the paper and a related test in a Wald-type statistic for a probabilistic index model (PIM, Thas et al., 2012) using a sandwich-type covariance matrix estimator, say $\widehat{\boldsymbol{S}}$, and weighted rank estimators for the PIM effects, say $\widehat{\boldsymbol{\alpha}}$, instead of $\widehat{\boldsymbol{V}}_N$ and $\widehat{\boldsymbol{p}}$, respectively, and a χ^2-quantile with estimated degrees of freedom given by $r(\boldsymbol{C}\widehat{\boldsymbol{S}}\boldsymbol{C}')$. The latter is motivated from the considerations in de Neve and Thas (2015) and denoted as DTS. We note that it is a test for the related null hypothesis $H_0^\alpha : \alpha_1 = \cdots = \alpha_4$ formulated in terms of the weighted PIM effects α_i (see Equation (4) in de Neve and Thas, 2015, for its explicit definition) which is equal to H_0^p in the balanced case. The ingredients of the test statistic were calculated as described in the supplementary material of de Neve and Thas (2015) with the *R* package PIM (Version 1.1.5.6). Moreover, note that the Kruskal-Wallis test has been developed for testing the more restrictive null hypothesis $H_0^F : F_1 = F_2 = \cdots = F_a$ formulated in terms of the distribution functions.

Symmetrically distributed data was generated from the model

$$X_{ik} = \mu_i + \sigma_i \epsilon_{ik}, \quad i = 1, \ldots, a; \quad k = 1, \ldots, n_i,$$

where the random error terms

$$\epsilon_{ik} = \frac{\widetilde{\epsilon}_{ik} - E(\widetilde{\epsilon}_{i1})}{\sqrt{Var(\widetilde{\epsilon}_{i1})}}$$

were generated from different standardized symmetric distributions, i.e., the random variables $\widetilde{\epsilon}_{ik}$ were generated from standard normal or the double exponential distribution, respectively. Skewed data was generated from log-normal-distributions by $X_{ik} = \exp(\eta_{ik})$, where $\eta_{ik} \sim N(0, \sigma_i^2)$ and possibly different variances σ_i^2. Note that the null hypothesis $H_0^p : \boldsymbol{P}_a\boldsymbol{p} = \mathbf{0}$ holds in both cases, because of the symmetry and the monotonicity of the exponential function.

A major assessment criterion for the accuracy of the methods is their behavior when different sample sizes and variances are combined, i.e. when increasing sample sizes are combined with increasing variances (*positive pairing*) or with decreasing variances (*negative pairing*) (see Pauly et al., 2015a). We consider balanced situations with sample size vector $\boldsymbol{n}_1 = (n_1, n_2, n_3, n_4) = (5, 5, 5, 5)$ and unbalanced situations with sample size vector $\boldsymbol{n}_2 = (n_1, n_2, n_3, n_4) = (10, 20, 30, 40)$, respectively. The scaling vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ was chosen from $(1, 1, 1, 1), (1, \sqrt{2}, 2, \sqrt{5})$ or $(\sqrt{5}, 2, \sqrt{2}, 1)$, respectively. In order to investigate the behavior of the tests when the sample sizes increase, a constant $m \in \{5, 10, 20, 25\}$ was added to each component of the vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$, i.e. $\boldsymbol{n}_i + m\mathbf{1}_4' = (n_1 + m, n_2 + m, n_3 + m, n_4 + m), i = 1, 2$. The different simulation settings are summarized in Table 4.

Table 4: Simulated one-way layout with $a = 4$ samples, where $m \in \{0, 5, 10, 20, 25\}$ and $\boldsymbol{n}_1 = (5, 5, 5, 5)$, and $\boldsymbol{n}_2 = (10, 20, 30, 40)$.

Setting	Sample Size	Scaling Factors	Meaning
1	$\boldsymbol{n} = \boldsymbol{n}_1 + m\mathbf{1}_4'$	$\boldsymbol{\sigma} = (1, 1, 1, 1)$	Balanced homoscedastic
2	$\boldsymbol{n} = \boldsymbol{n}_2 + m\mathbf{1}_4'$	$\boldsymbol{\sigma} = (1, 1, 1, 1)$	Unbalanced homoscedastic
3	$\boldsymbol{n} = \boldsymbol{n}_1 + m\mathbf{1}_4'$	$\boldsymbol{\sigma} = (1, \sqrt{2}, 2, \sqrt{5})$	Balanced heteroscedastic
4	$\boldsymbol{n} = \boldsymbol{n}_2 + m\mathbf{1}_4'$	$\boldsymbol{\sigma} = (1, \sqrt{2}, 2, \sqrt{5})$	Unbalanced heteroscedastic (Positive Pairing)
5	$\boldsymbol{n} = \boldsymbol{n}_2 + m\mathbf{1}_4'$	$\boldsymbol{\sigma} = (\sqrt{5}, 2, \sqrt{2}, 1)$	Unbalanced heteroscedastic (Negative Pairing)

Since the balanced settings have been discussed in the main paper we here only comment on the unbalanced settings.

In Table 5 the simulation results for the unbalanced homoscedastic designs (Setting 2) for various distributions are displayed. As in Setting 1 the hypothesis H_0^F holds here and it is not surprising that similar observations can be drawn. First, the Kruskal-Wallis test controls the nominal type-1 error level ($\alpha = 5\%$) very satisfactorily for all investigated distributions. Second, both of the Wald-type statistics (DTS and WTS) tend to be considerably liberal, where their liberality again slowly decreases with increasing sample sizes. However, even for the scenarios with larger sample sizes their type-I-error control is not acceptable. Finally, the behaviour of the ANOVA-type tests is again similar to the main paper: For smaller sample sizes ($N \leq 120$) both the tests $\widehat{\varphi}_N$ and $\widetilde{\varphi}_N$ are slightly liberal. For larger sample sizes their type-I error control is acceptable. In contrast, the ANOVA-type test φ_N based on the F-approximation shows a better control of the type-1 error level and is only slightly liberal in case of the smallest simulated sample sizes.

In the two unbalanced heteroscedastic Settings 4 and 5 the null hypothesis H_0^F is violated and only H_0^p is true. The corresponding results are shown in Tables 6–7.

In case of positive pairings (see the results for Setting 4 in Table 6), the Kruskal-Wallis test tends to be conservative in all considered scenarios. In comparison to Settings 1 - 3, both the methods $\widehat{\varphi}_N$ and $\widetilde{\varphi}_N$ tend to be less liberal and fairly control the type-1 error rate α. The two Wald-type tests (DTS and WTS) are still but less liberal. Also in this setup, the ANOVA-type test φ_N controls the type-1 error rate very satisfactorily.

The most severe case from all investigated scenarios is when larger sample sizes are combined with smaller variances (negative pairing – Setting 5). The simulation results are displayed in Table 7 below. It can be readily seen that the Kruskal-Wallis test and both Wald-type tests tend to quite liberal conclusions. Moreover, both the methods $\widehat{\varphi}_N$ and $\widetilde{\varphi}_N$ do not control the error rate $\alpha = 5\%$ in this set-up. The method φ_N tends to be slightly liberal in case of the smallest simulated sample sizes, but controls the type-1 error rate superior to all other methods. Thus, the ANOVA-type test φ_N is recommended for practical applications.

Table 5: Type-I error ($\alpha = 5\%$) simulations of the Kruskal-Wallis test (KW), the two Wald-type tests in the test statistics WTS and the test statistic of De Neve and Thas (DTS) and the three different ANOVA-type tests $\widehat{\varphi}_N$, $\widetilde{\varphi}_N$, and φ_N using the distributional approximations as given in (4.21), (4.23), and (4.24) of the main paper under Setting 2 as described in Table 4.

Distribution	Sample Sizes	KW	DTS	WTS	$\widehat{\varphi}_N$	$\widetilde{\varphi}_N$	φ_N
DExp	10 20 30 40	0.0518	0.1114	0.0908	0.0758	0.0777	0.0659
DExp	15 25 35 45	0.0483	0.0926	0.0727	0.0604	0.0607	0.0538
DExp	20 30 40 50	0.0475	0.0804	0.0740	0.0562	0.0564	0.0515
DExp	30 40 50 60	0.0497	0.0674	0.0629	0.0559	0.0555	0.0520
DExp	35 45 55 65	0.0492	0.0650	0.0604	0.0531	0.0538	0.0500
LogNor	10 20 30 40	0.0480	0.1112	0.0903	0.065	0.0667	0.0580
LogNor	15 25 35 45	0.0482	0.0874	0.0730	0.0588	0.0597	0.0526
LogNor	20 30 40 50	0.0481	0.0824	0.0711	0.0543	0.0555	0.0498
LogNor	30 40 50 60	0.0505	0.0726	0.0654	0.0549	0.0553	0.0510
LogNor	35 45 55 65	0.0468	0.0720	0.0696	0.0506	0.0520	0.0476
Normal	10 20 30 40	0.0477	0.1026	0.0912	0.0650	0.0667	0.0576
Normal	15 25 35 45	0.0483	0.0916	0.0764	0.0578	0.0588	0.0504
Normal	20 30 40 50	0.0478	0.0820	0.0708	0.0517	0.0515	0.0475
Normal	30 40 50 60	0.0518	0.0732	0.0645	0.0559	0.0570	0.0525
Normal	35 45 55 65	0.0531	0.0712	0.0639	0.0579	0.0583	0.0535

Table 6: Type-I error ($\alpha = 5\%$) simulations of the Kruskal-Wallis test (KW), the two Wald-type tests in the test statistics WTS and the test statistic of De Neve and Thas (DTS) and the three different ANOVA-type tests $\widehat{\varphi}_N$, $\widetilde{\varphi}_N$, and φ_N using the distributional approximations as given in (4.21), (4.23), and (4.24) of the main paper under Setting 4 as described in Table 4.

Distribution	Sample Sizes	KW	DTS	WTS	$\widehat{\varphi}_N$	$\widetilde{\varphi}_N$	φ_N
DExp	10 20 30 40	0.0247	0.1008	0.0749	0.0563	0.0566	0.0492
DExp	15 25 35 45	0.0318	0.0822	0.0696	0.0534	0.0533	0.0486
DExp	20 30 40 50	0.0371	0.0828	0.0682	0.0578	0.0573	0.0541
DExp	30 40 50 60	0.0414	0.0730	0.0640	0.0543	0.0549	0.0515
DExp	35 45 55 65	0.0421	0.0696	0.0584	0.0525	0.0533	0.0498
LogNor	10 20 30 40	0.0364	0.1000	0.0873	0.0653	0.0670	0.0586
LogNor	15 25 35 45	0.0392	0.0904	0.0736	0.0567	0.0562	0.0511
LogNor	20 30 40 50	0.0387	0.0752	0.0686	0.0538	0.0546	0.0495
LogNor	30 40 50 60	0.0420	0.0718	0.0665	0.0527	0.0535	0.0494
LogNor	35 45 55 65	0.0407	0.0684	0.0597	0.0507	0.0511	0.0479
Normal	10 20 30 40	0.0248	0.0938	0.0706	0.0545	0.0547	0.0475
Normal	15 25 35 45	0.0287	0.0858	0.0710	0.0525	0.0524	0.0470
Normal	20 30 40 50	0.0340	0.0834	0.0674	0.0546	0.0549	0.0493
Normal	30 40 50 60	0.0411	0.0690	0.0652	0.0523	0.0528	0.0492
Normal	35 45 55 65	0.0432	0.0702	0.0587	0.0512	0.0517	0.0479

Table 7: Type-I error ($\alpha = 5\%$) simulations of the Kruskal-Wallis test (KW), the two Wald-type tests in the test statistics WTS and the test statistic of De Neve and Thas (DTS) and the three different ANOVA-type tests $\widehat{\varphi}_N$, $\widetilde{\varphi}_N$, and φ_N using the distributional approximations as given in (4.21), (4.23), and (4.24) of the main paper under Setting 5 as described in Table 4.

Distribution	Sample Sizes	KW	DTS	WTS	$\widehat{\varphi}_N$	$\widetilde{\varphi}_N$	φ_N
DExp	10 20 30 40	0.1178	0.1108	0.0907	0.0755	0.0760	0.0628
DExp	15 25 35 45	0.1029	0.1008	0.0768	0.0652	0.0665	0.0580
DExp	20 30 40 50	0.1014	0.0916	0.0721	0.0602	0.0609	0.0538
DExp	30 40 50 60	0.0888	0.0722	0.0671	0.0544	0.0554	0.0502
DExp	35 45 55 65	0.0879	0.0746	0.0655	0.0544	0.0557	0.0496
LogNor	10 20 30 40	0.0695	0.1060	0.0912	0.0753	0.0760	0.0644
LogNor	15 25 35 45	0.0667	0.0962	0.0795	0.0654	0.0660	0.0583
LogNor	20 30 40 50	0.0580	0.0796	0.0685	0.0550	0.0557	0.0508
LogNor	30 40 50 60	0.0586	0.0768	0.0629	0.0497	0.0506	0.0465
LogNor	35 45 55 65	0.0623	0.0666	0.0638	0.0563	0.0565	0.0515
Normal	10 20 30 40	0.1287	0.1132	0.0935	0.0719	0.0727	0.0619
Normal	15 25 35 45	0.1198	0.0882	0.0798	0.0656	0.0675	0.0583
Normal	20 30 40 50	0.1167	0.0800	0.0713	0.0624	0.0635	0.0549
Normal	30 40 50 60	0.1044	0.0724	0.0678	0.0598	0.0608	0.0546
Normal	35 45 55 65	0.0992	0.0746	0.0617	0.0562	0.0571	0.0510

REFERENCES

AKRITAS, M. G., AND ARNOLD, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *J. Amer. Statist. Assoc.* **89**, 336–343.

AKRITAS, M. G., ARNOLD, S. F., AND BRUNNER, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Amer. Statist. Assoc.* **92**, 258–265.

AKRITAS, M. G. AND BRUNNER, E. (1997). A unified approach to ranks tests in mixed models. *J. Statist. Plann. Inference* **61**, 249–277.

AKRITAS, M. G. (2011). Nonparametric Models for ANOVA and ANCOVA Designs. In *International Encyclopedia of Statistical Science*, Springer, 964–968.

BRUNNER, E., MUNZEL, U. AND PURI, M. L. (1999). Rank-Score Tests in Factorial Designs with Repeated Measures. *J. Mult. Analysis* **70**, 286–317.

BRUNNER, E., AND PURI, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers* **42**, 1–52.

BRUNNER, E., KONIETSCHKE, F., PAULY, M. AND PURI, M.L. (2016). Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects.

DE NEVE, J., AND THAS, O. (2015). A Regression Framework for Rank Tests Based on the Probabilistic Index Model. *J. Amer. Statist. Assoc.*, DOI: 10.1080/01621459.2015.101622.

DOMHOF, S. (2001). Nichtparametrische relative Effekte. Ph.D. Thesis, University of Göttingen.

GAO, X., AND ALVO, M. (2005). A unified nonparametric approach for unbalanced factorial designs. *J. Amer. Statist. Assoc.* **100**, 926–941.

GAO, X., AND ALVO, M. (2008). Nonparametric multiple comparison procedures for unbalanced two-way layouts. *Journal of Statistical Planning and Inference* **138**, 3674–3686.

GAO, X., ALVO, M., CHEN, J., AND LI, G. (2008). Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *Journal of Statistical Planning and Inference* **138**, 2574–2591.

KONIETSCHKE, F., BATHKE, A. C., HOTHORN, L. A., AND BRUNNER, E. (2010). Testing and estimation of purely nonparametric effects in repeated measures designs. *Computational Statistics and Data Analysis*, **54**, 1895–1905.

PLACZEK, M. (2013). Nichtparametrische simultane Inferenz für faktorielle Repeated Measures Designs. Master Thesis, University of Göttingen.