SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2021, Vol. 30(3) 687–701

DOI: 10.1177/0962280220970228
journals.sagepub.com/home/smm
SAGE

Small sample sizes: A big data problem in high-dimensional data analysis

Frank Konietschke[1,2], Karima Schwab[3] and Markus Pauly[4]

Abstract

In many experiments and especially in translational and preclinical research, sample sizes are (very) small. In addition, data designs are often high dimensional, i.e. more dependent than independent replications of the trial are observed. The present paper discusses the applicability of *max* t-test-type statistics (multiple contrast tests) in high-dimensional designs (repeated measures or multivariate) with small sample sizes. A randomization-based approach is developed to approximate the distribution of the maximum statistic. Extensive simulation studies confirm that the new method is particularly suitable for analyzing data sets with small sample sizes. A real data set illustrates the application of the methods.

Keywords

Multiple contrast tests, max t-test, repeated measures, resampling, simultaneous confidence intervals

1 Introduction

Small sample sizes occur in various research experiments and especially in preclinical (animal) studies due to ethical, financial, and general feasibility reasons. Such studies are essential and an important part of translational medicine and other areas (e.g. rare diseases). Often, less than 20 animals per group are involved, and thus making valid inferences in these studies becomes a challenging part. In addition to the small sample sizes, repeated measurements as well as multiple endpoints are often observed on the experimental units (animals), naturally leading to a “large p, small n” situation and thus to a high-dimensional data design. Note that high-dimensional data do not only occur in animal studies, medical imaging and genomics are other well-known application areas. The first statistical problem at hand is neither the high dimensionality of the data nor the relatively low statistical power of the tests when sample sizes are very small—it is the accurate type-1 error rate control of the methods. Many of the existing statistical methods require moderate or large sample sizes and therefore tend to not control the type-1 error rate properly when sample sizes are very small; they either behave liberal and over-reject the null hypothesis or are conservative. Exact techniques (i.e. procedures that rely on the exact distribution of a test statistic for any finite sample size n) would be a great choice, but they typically rely on strict model assumptions that can hardly be verified—at least in more complex models. Indeed, making any assumptions about the underlying distributions (e.g. based on boxplots), verifying equality of variances, specific covariance structures, etc. is quasi impossible when sample sizes are so small and thus, methods, which do not rely on strict model assumptions, are the methods of choice. All in all, besides the often discussed problem of high dimensionality, small sample sizes increase the challenge of a robust and especially accurate data analysis in such situations.

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

Corresponding author:
Frank Konietschke, Charite Universitatsmedizin Berlin, Chariteplatz 1, Berlin 10117, Germany.
Email: Frank.Konietschke@charite.de

Beyond these challenges and even though the statistical designs themselves are usually complex, the research questions and study aims are often very specific. These may be tackled by applying global testing procedures, which have been developed for different high-dimensional repeated measures and multivariate ANOVA models by several authors.[1–8] Furthermore, multivariate tests based on interpoint distances have been proposed.[9–12] Testing global null hypotheses and herewith answering the question whether any difference among the repeated measurements per or across endpoints exists, however, does usually not answer the main question of the practitioners—that is the specific localization of the responsible experimental conditions that lead to the overall significance conclusion. A modern data analysis requires the use of multiple comparison procedures that control the family-wise error rate in the strong sense and that are flexible in the way that they can be used to test arbitrary global and local null hypotheses and lead to *compatible* simultaneous confidence intervals (SCIs) for the underlying treatment effects. Furthermore, due to the often complex dependency structure across the repeated measurements and contrasts, the multiple comparison method should take the correlations of the different tests statistics for powerful data analysis into account. Such methods are also known as *multiple contrast tests* (MCTP) and are based on the maximum value of a vector of possibly correlated t-test type statistics (max t-test statistic). Computing its exact distribution without making strict distributional assumptions is impossible in general designs.[13] Therefore, approximations of its asymptotic distribution are needed for making inferences. Recently, it has been suggested[14] to estimate the distribution of the maximum value within a bootstrap simulation-based framework using the empirical correlation matrix. Simulation studies indicate, however, that sample sizes $n_i \geq 50$ are necessary for an accurate type-1 error rate control. When sample sizes are smaller, the methods tend to be liberal (see Section 5). In the present paper, a modification of the proposed method will be introduced that does not require the estimation of the correlation matrix. Extensive numerical studies show that the new approach controls the type-1 error very accurately even when sample sizes are very small and data do not follow multivariate normal distributions with equal covariance matrices.

The paper is organized as follows. In Section 2, a high-dimensional preclinical study on Azheimer's disease with small sample sizes is described. Existing methodology for the statistical evaluation of such designs is explained in Section 3. Here, its behavior in small sample size situations is also investigated, which motivates the development of a different approximation of the distribution of the max t-test in Section 4. The qualities of the competing approximations are compared in extensive simulation studies in Section 5. The paper closes with the evaluation of the example and a discussion about the results in Sections 6 and 7, respectively. Theoretical properties of the new approximation and proofs are provided in the supplementary material file. Throughout the paper, $\boldsymbol{I}_d$ denotes the d-dimensional unit matrix, $\boldsymbol{J}_d = 1_d 1_d'$ the $d \times d$ matrix of 1 s, where $1_d = (1, \ldots, 1)'_{d\times 1}$.

2 A motivating example

As a motivating example, we consider a part of a preclinical study on Alzheimer's disease conducted in the Institute of Pharmacology at the Charité university medical center in Berlin, Germany. The study involves $n_1 = 10$ wild-type mice (group 1) and $n_2 = 9$ L1 tau-transgenic type (group 2) mice. As usual, the sample sizes of this preclinical research trial are pretty small. The abundance of each of the six different proteins Syntaxin, SNAP25, VAMP2, Synaptophysin, Synapsin-1 and Alpha-synuclein were measured in six different regions of the brain of every mouse. The regions of interest were pre-defined as hippocampal CA1 region (CA1), visual cortex (VC), medial septum (MS), vertical limb of the diagonal band of Broca (VDB), primary motor cortex (M1) and nucleus accumbens (ACB), respectively. Thus, 36 observations were made on every mouse, while the number of dependent replicates exceeds the number of independent replications of the trial. Therefore, the statistical design represents a classical "large p, small n" situation with small sample sizes. We note that generating such data requires very advanced technology. For graphical representation and easy display of the data, the protein abundances were log-transformed. The results are displayed in confidence interval plots (with chosen local level 95%) in Figure 1. For illustration, additional dotplots and boxplots of the data are displayed in the supplementary material file.

The dotplots give the impression that the protein abundances are roughly symmetrically distributed. However, since sample sizes are so small, making any assumption about the underlying distribution would be questionable. It can also be seen that few "outliers" are present. These values have been kept in the data set, because the range of protein abundance measurements is usually very wide. Therefore, these values are not outliers in the classical sense and provide useful information about the protein levels in the respective brain regions. Furthermore, the confidence intervals displayed in Figure 1 show a fairly amount of variance heteroscedasticity. Therefore, the data of this trial can be modeled by independent and identically distributed random vectors

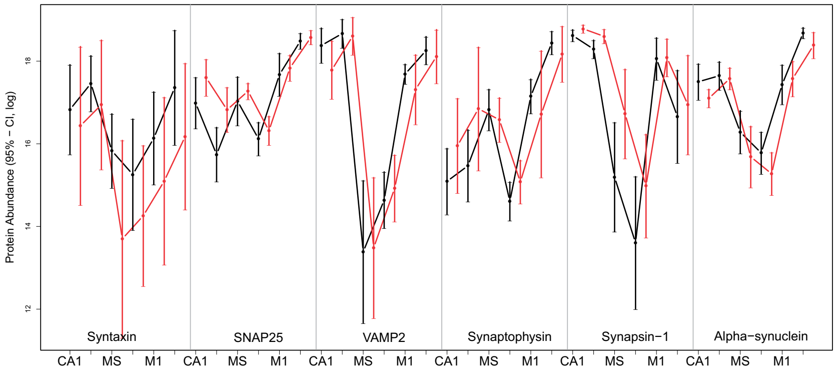

Figure 1. Confidence interval plot (95%) for each protein× region× group combination in the protein abundance trial. Each confidence interval has been computed by inverting the corresponding one-sample *t*-test statistic using 97.5%-t-quantiles from a *t*-distribution with $n_i - 1$ degrees of freedom.

$$\boldsymbol{X}_{ik} = (X_{i1k}, \ldots, X_{idk})' \sim \boldsymbol{F}_i,\ i = 1, 2;\ k = 1, \ldots, n_i;\ N = n_1 + n_2 \tag{1}$$

with expectation $E(\boldsymbol{X}_{i1}) = \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})'$ and covariance matrix $Cov(\boldsymbol{X}_{i1}) = \boldsymbol{\Sigma}_{\mathrm{i}} > 0$, $i = 1, 2$. For a convenient notation, the index $s = 1, \ldots, d$ represents the repeated measures in the regions under each of the different endpoints. Here, we set

$$d : \left\{ \begin{array}{lll} \text{Dimension} & \text{Protein} & \text{Region} \\ 1 \le d \le 6 & \text{Syntaxin} & (\text{CA1},\ \text{VC},\ \text{MS},\ \text{VDB},\ \text{M1},\ \text{ACB}) \\ 7 \le d \le 12 & \text{SNAP25} & (\text{CA1},\ \text{VC},\ \text{MS},\ \text{VDB},\ \text{M1},\ \text{ACB}) \\ 13 \le d \le 18 & \text{VAMP2} & (\text{CA1},\ \text{VC},\ \text{MS},\ \text{VDB},\ \text{M1},\ \text{ACB}) \\ 19 \le d \le 24 & \text{Synaptophysin} & (\text{CA1},\ \text{VC},\ \text{MS},\ \text{VDB},\ \text{M1},\ \text{ACB}) \\ 25 \le d \le 30 & \text{Synapsin} - 1 & (\text{CA1},\ \text{VC},\ \text{MS},\ \text{VDB},\ \text{M1},\ \text{ACB}) \\ 31 \le d \le 36 & \text{Alpha} - \text{synuclein} & (\text{CA1},\ \text{VC},\ \text{MS},\ \text{VDB},\ \text{M1},\ \text{ACB}) \end{array} \right.$$

and arrange all analyses according to this order. In general, data modeled by Equation (1) can either be repeated measures (measurements on the same scale), multivariate data (measurements on different scales) or combinations thereof. For a convenient notation of the hypotheses, let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1{}', \boldsymbol{\mu}_2{}')'$ denote the combined vector of the expectations in both groups. Besides the questions whether abundances of the proteins in the different regions of the brain differ, the study specifically aims to locate specific group × region interactions for each of the proteins. From a medical point of view, these would expose the proteins and especially the regions of the brain as biomedical biomarkers for Alzheimer's disease. To be more specific, let $\mu_{ij}^{(P)}$ denote the expected protein abundance in group i under region j of protein P, where $i = 1,\ 2;\ j \in \{\text{CA1, VC, MS, VDB, M1, ACB}\}$ and $P \in \mathcal{P} = \{$Syntaxin, SNAP25, ..., Alpha-synuclein$\}$, respectively. For each single protein, the major aim is to (i) decide whether there is a group × region interaction and if so (ii) where. This can be achieved by simultaneously testing whether the group-wise differences $\mu_{1j}^{(P)} - \mu_{2j}^{(P)}$ are identical for all regions $j = 1, \ldots, 6$ for each protein $P = 1, \ldots, 6$. This leads to testing the family of 72 multiple null hypotheses

$$\boldsymbol{\Omega} = \left\{ H_0^{(j,P)} : \mu_{ij}^{(P)} = \bar{\mu}_{i\cdot}^{(P)} + \bar{\mu}_{\cdot j}^{(P)} - \bar{\mu}_{\cdot\cdot}^{(P)}, i = 1, 2, j = 1, \ldots, 6, P \in \mathcal{P} \right\}$$

at multiple level α. Here,

$$\bar{\mu}_{i\cdot}^{(P)} = \frac{1}{6}\sum_{j=1}^{6} \mu_{ij}^{(P)},\ \bar{\mu}_{\cdot j}^{(P)} = \frac{1}{2}\sum_{i=1}^{2} \mu_{ij}^{(P)} \quad \text{and} \quad \bar{\mu}_{\cdot\cdot}^{(P)} = \frac{1}{12}\sum_{i=1}^{2}\sum_{j=1}^{6} \mu_{ij}^{(P)}$$

denote the corresponding means of expectations as known from linear model theory, where $i = 1, 2$; $j = 1, \dots, 6$; $P \in \mathcal{P}$. Thus, the hypotheses are nothing but testing whether the differences of the expectations $\mu_{1j}^{(P)} - \mu_{2j}^{(P)}, j = 1, \dots, 6$ are identical for each protein. For simplicity, we rewrite the above using matrix notation and equivalently obtain

$$\boldsymbol{\Omega} = \{H_0^{(\ell)} : \boldsymbol{c}_\ell{}' \boldsymbol{\mu} = 0, \ell = 1, \dots, q\}$$

where $\boldsymbol{c}_\ell{}'$ denotes the ℓth row vector of the contrast matrix as used in

$$H_0^\mu : \boldsymbol{C}\boldsymbol{\mu} = 0, \quad \text{with} \quad \boldsymbol{C} = \Big(\bigoplus_{P \in \mathcal{P}} \boldsymbol{P}_6\Big)\Big(\boldsymbol{I}_{36} \vdots - \boldsymbol{I}_{36}\Big) \tag{2}$$

Here, in Equation (2), we denote with $\boldsymbol{P}_m = \boldsymbol{I}_m - \frac{1}{m} 1_m 1_m{}'$ the m-dimensional centering matrix, while $\oplus$ describes the direct sum to build a block diagonal matrix.

Note that this summarizing matrix notation enables us to describe the 72 null hypotheses equivalently by only q = 36 which shall be tested using $n_1 = 10$ and $n_2 = 9$ independent replications. Therefore, the above is a high-dimensional multiple testing problem. An existing statistical method to analyze the data will be discussed in the next section.

3 Existing methodology

The high dimensionality of the testing problem considered here makes the data analysis complex in the sense that the computation of the critical values for making statistical inference becomes an issue. Recently, Chang et al.[14] propose a simulation-based inference method for high-dimensional data. The procedure is valid for large dimensions and sample sizes. The case of small sample sizes has not been considered and therefore its applicability in such situations intrigues a detailed investigation. In their original paper, both the cases of studentized and non-studentized statistics have been considered. For the ease of read, we will concentrate on the studentized statistics in the following only. By doing so, we follow the guidelines of resampling studentized statistics.[15–18] First, we will rewrite the null hypothesis and introduce the statistics in the same way as they were described by Chang et al.[14] who propose to test the equality of expectations of the q-variate random vectors $\mathbf{Y}_{ik} = (Y_{i1k}, \dots, Y_{iqk})'$ by $H_0^\theta : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. These considerations show that the two statistical testing problems are identical. However, we will propose a different way of estimating the critical values for making reliable inference later in Section 4.

Note that the null hypothesis $H_0 : \boldsymbol{C}\boldsymbol{\mu} = 0$ as given in Equation (2) can be equivalently written as the "standard" multivariate null hypothesis

$$H_0^\theta : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iq})' = E(\boldsymbol{Y}_{ik})$ denotes the expectation of the transformed vectors $\boldsymbol{Y}_{ik}{}' = \boldsymbol{X}_{ik}{}'\Big(\oplus_{s=1}^6 \boldsymbol{P}_6\Big)$. In order to test H_0^θ against $H_1^\theta : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, consider the maximum of the q component-wise t-test type statistics

$$T_0 = \max\big\{|T_1|, \dots, |T_q|\big\}, \quad \text{where } T_\ell = \sqrt{N}\frac{(\bar{Y}_{1\ell\cdot} - \bar{Y}_{2\ell\cdot}) - (\theta_{1\ell} - \theta_{2\ell})}{\sqrt{\hat{v}_{1,\ell\ell}/n_1 + \hat{v}_{2,\ell\ell}/n_2}} \tag{3}$$

denotes the studentized difference of the means $\bar{Y}_{i\ell\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{i\ell k}$ with the empirical variances $\hat{v}_{i,\ell\ell} = N(n_i - 1)^{-1}\sum_{k=1}^{n_i}(Y_{i\ell k} - \bar{Y}_{i\ell\cdot})^2$, $i = 1, 2$; $\ell = 1, \dots, q$. The use of maximum t-statistics plays an important role in preclinical research, because local test decisions can be made using adjusted p-values for the comparisons $H_0^{(\ell)} : \theta_{1\ell} = \theta_{2\ell}$, $\ell = 1, \dots, q$. Second, each t-statistic describes the distance of the observed mean difference to its respective null hypothesis in units of standard deviations. However, for the computation of the local p-values, the distribution of T_0 must be known, at least approximately. Suppose for a moment that it is known, then the individual null hypothesis $H_0^{(\ell)} : \theta_{1\ell} = \theta_{2\ell}$ will be rejected at multiple level α, if

$$|T_\ell| \geq z_{1-\alpha}(\max) \tag{4}$$

where $z_{1-\alpha}(\max)$ denotes the $(1-\alpha)$-quantile from the distribution of T_0. Compatible SCIs for the effects $\delta_\ell = \theta_{1\ell} - \theta_{2\ell}$ are given by

$$CI_\ell = \left[\bar{Y}_{1\ell\cdot} - \bar{Y}_{2\ell\cdot} \mp \frac{z_{1-\alpha}(\max)}{\sqrt{N}} \sqrt{\hat{v}_{1,\ell\ell}/n_1 + \hat{v}_{2,\ell\ell}/n_2}\right] \tag{5}$$

Finally, the global null hypothesis $H_0 : \boldsymbol{C\mu} = 0$ will be rejected, if

$$T_0 \geq z_{1-\alpha}(\max) \tag{6}$$

In such general models (even under the assumption of multivariate normality), however, the exact distribution of T_0 remains unknown[19] and approximate methods are needed for estimating the distribution of T_0. In low-dimensional designs (fixed dimension d and contrasts q), the vector of t-statistics

$$\boldsymbol{T} = (T_1, \ldots, T_q)' \tag{7}$$

follows, asymptotically, as $N \to \infty$, a multivariate normal distribution with expectation 0 and correlation matrix

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{V}\boldsymbol{D}^{-1/2}, \text{ where} \tag{8}$$

$$\boldsymbol{V} = Cov(\sqrt{\mathrm{N}}(\bar{\boldsymbol{Y}}_{1\cdot} - \bar{\boldsymbol{Y}}_{2\cdot})) = \mathrm{N}\left(\ddot{\boldsymbol{C}}\boldsymbol{\Sigma}_1\ddot{\boldsymbol{C}}'/\mathrm{n}_1 + \ddot{\boldsymbol{C}}\boldsymbol{\Sigma}_2\ddot{\boldsymbol{C}}'/\mathrm{n}_2\right) \tag{9}$$

with

$$\ddot{\boldsymbol{C}} = \bigoplus_{s=1}^{6} \boldsymbol{P}_6$$

denotes the covariance matrix of the differences in means and $\boldsymbol{D}$ denotes the diagonal matrix obtained from the diagonal elements of $\boldsymbol{V}$. These considerations show that the (asymptotic) joint distribution of the vector of t-statistics $\boldsymbol{T}$ (and therefore of T_0) depends on the unknown correlation matrix and is non-pivotal. This is intuitively clear, since the higher the statistics are correlated, the smaller should be the critical value $z_{1-\alpha}(\max)$. Indeed, in case of a perfect correlation, the above reduces to a univariate testing problem. Anyway, the correlation matrix is unknown and the above cannot be used for making inferences in its present form. Chang et al.[14] propose to first estimate the correlation matrix by its empirical counterpart

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{D}}^{-1/2}\hat{\boldsymbol{V}}\hat{\boldsymbol{D}}^{-1/2}, \text{ where}$$

$$\hat{\boldsymbol{V}} = N\left(\hat{\boldsymbol{V}}_1/n_1 + \hat{\boldsymbol{V}}_2/n_2\right), \text{and}$$

$$\hat{\boldsymbol{V}}_i = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(\boldsymbol{Y}_{ik} - \bar{\boldsymbol{Y}}_{i\cdot})(\boldsymbol{Y}_{ik} - \bar{\boldsymbol{Y}}_{i\cdot})', \; i = 1, 2$$

denote the empirical covariance matrix of the random vectors $\boldsymbol{Y}_{ik}$. Analogously, $\hat{\boldsymbol{D}}$ denotes the diagonal matrix obtained from the diagonal elements of $\hat{\boldsymbol{V}}$. Next, they propose to generate M random vectors

$$\boldsymbol{Y}_b^* \sim N(0, \hat{\boldsymbol{R}}), \quad b = 1, \ldots, M \tag{10}$$

from a multivariate normal distribution with expectation 0 and correlation matrix $\hat{\boldsymbol{R}}$ and to estimate the critical value $z_1\text{−}\alpha(\max)$ by computing the $(1-\alpha)$-quantile $y^*_{1-\alpha}(\max)$ of the values $Y^*_{0,1}, \ldots, Y^*_{0,M}$, where $Y^*_{0,b} = \max\{|Y^*_{b1}|, \ldots, |Y^*_{bq}|\}$ denotes the maximum value of each of the random vectors $\boldsymbol{Y}^*_b$ (in absolute value). Finally, the unknown quantile $z_{1-\alpha}(\max)$ is replaced with the observable estimator $y^*_{1-\alpha}(\max)$ in Equations (4) to (6), respectively. Note that the quantile $y^*_{1-\alpha}$ can also be computed directly using the R-function *qmvnorm* implemented in the R-package *mvtnorm*[20] (if the dimension is not "too large").

However, the present small sample sizes arise the question whether the method accurately controls the type-1 error rate and thus leads to reliable conclusions. Note that, in the data example, a 36×36 dimensional correlation matrix is estimated upon $n_1 = 10$ and $n_2 = 9$ independent vectors per group. Roughly speaking, the estimator might be too inaccurate when sample sizes are so small. In order to answer this question, a motivating simulation study has been conducted. Data has been generated from (i) multivariate normal and (ii) multivariate T_3-distributions with $df = 3$ degrees of freedom each with group specific covariance matrices $\hat{\boldsymbol{V}}_i$, i.e.

$$(i)\, \boldsymbol{X}_{ik} \sim N(\boldsymbol{0}, \hat{\boldsymbol{V}}_i) \quad \text{and} \quad (ii)\, \boldsymbol{X}_{ik} \sim T_3(\boldsymbol{0}, \hat{\boldsymbol{V}}_i), \quad i = 1, 2; k = 1, \ldots, n_i \tag{11}$$

Here, $\hat{\boldsymbol{V}}_i$ denotes the empirical covariance matrix of group i in the Alzheimer's disease study. Data have been transformed to $\boldsymbol{Y}_{ik}{}' = \boldsymbol{X}_{ik}{}'\left(\oplus_{s=1}^{6} \boldsymbol{P}_6\right)$ as described above. The T_3-distribution is heavy tailed and might be a reasonable candidate to mimic the distributional shape of the protein abundance data. Note that $\hat{\boldsymbol{V}}_i$ is singular and therefore data have been generated using singular value decomposition of $\hat{\boldsymbol{V}}_i$ using the *rmvnorm* function implemented in the *mvtnorm* R-package.[21] The simulation results for varying sample sizes $n_1 = n_2 = 8, 9, \ldots, 50$ at nominal significance level $\alpha = 5\%$ are displayed in Table 1.

It follows that the procedure does not control the type-1 error rate appropriately when sample sizes are very small. With sample sizes $n_i = 9$, the empirical type-1 error rate is about 20% under normality and about 15% under heavy tailed $T_3(\boldsymbol{0}, \hat{\boldsymbol{V}}_i)$-distribution and hence highly inflated. Only with larger sample sizes ($n_i \geq 50$), the method controls the type-1 error rate quite appropriately under normality, while it tends to be slightly conservative under $T_3(\boldsymbol{0}, \hat{\boldsymbol{V}}_i)$, respectively. Digging for the reasons of this behavior, we first find that the procedure does not take the variations of the variance estimators used in the t-statistics in Equation (3) into account and second, the resampling algorithm is based on estimating the full correlation matrix of the vector of t-statistics. If a different resampling algorithm could be defined that overcomes both of these characteristics, a major improvement of the approximation might be available. In the next section, such a solution will be proposed.

4 Approximating the distribution of T_0

The arising challenge is finding a *good* approximation of the joint distribution of $\boldsymbol{T}$ for estimating critical- and p-values. Resampling methods as above are an innovative way to do so. Roughly speaking, the corresponding test will work, if both the limiting and the resampling distributions of the statistic coincide—at least asymptotically—under the null hypothesis of no treatment effect. As explained above, the vector of t-test type statistics follows, asymptotically, a multivariate normal distribution with expectation $\boldsymbol{0}$ and correlation matrix $\boldsymbol{R}$ in low-dimensional settings (d fixed). This means that a proper resampling algorithm must be designed in such a way that the resampling distribution of $\boldsymbol{T}$, say $\boldsymbol{T}^*$, converges to the $N(\boldsymbol{0}, \boldsymbol{R})$ distribution, respectively, where the correlation matrix must be identical to the one defined in Equation (8). Moreover, in high-dimensional settings (with $d \to \infty$) similar observations apply, see the supplementary material, where it is, for example, shown that the distribution of $\boldsymbol{T}$ converges to a discrete Gaussian process. Detailed assumptions, especially on the covariance matrices, are listed in the supplementary material document as well. Having these thoughts in mind, not every resampling method is applicable in high-dimensional designs with an emphasis on small sample sizes. For example, the nonparametric bootstrap (drawing with replacement) shows poor finite sample performances in a similar setting under stronger conditions.[22] Moreover, the therein proposed permutation method for exchangeable designs is in general not applicable in our unbalanced heteroscedastic setting.[23,24] For more details on permutation tests, we refer to existing overviews and monographs.[25–28] Therefore, the generation of the resampling variables and especially the algorithmic build-up play an important role for achieving a adequate bootstrap test in high-dimensional designs. Furthermore, even in low-dimensional settings (d fixed), estimating the approximate null $N(0, \boldsymbol{R})$

Table 1. Type-1 error ($\alpha = 5\%$) simulation results of the simulation-based test.[14]

Dist\\n_i	8	9	10	15	20	30	40	50
Nor	0.2134	**0.1908**	0.1723	0.1216	0.0975	0.0793	0.0720	0.0651
T_3	0.1681	**0.1483**	0.1133	0.0873	0.0781	0.0647	0.0430	0.0420

significance is provided with alpha = 5%

distribution of $\boldsymbol{T}$ using a plug-in estimator $\hat{\boldsymbol{R}}$ usually requires large sample sizes for an appropriate approximation. We therefore propose to approximate the limiting distribution of $\boldsymbol{T}$ without estimating the parameter of the distribution using a Wild-bootstrap randomization approach that is applicable in low- as well as high-dimensional situations. The method follows the same ideas proposed for matched pairs[18,29] and in high-dimensional linear models,[30] and is as follows. Let

$$\boldsymbol{Z}_{ik} = \boldsymbol{Y}_{ik} - \bar{\boldsymbol{Y}}_{i\cdot}, \quad i = 1, 2;\ k = 1, \ldots, n_i$$

denote the centered random vectors and let W_{ik} denote N independent and identically distributed random signs with $P(W_{ik} = \pm 1) = \frac{1}{2}$. Now, let

$$\boldsymbol{Z}_{ik}^* = W_{ik}\boldsymbol{Z}_{ik}$$

denote the resampling variables, $\bar{\boldsymbol{Z}}_{i\cdot}^* = \frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{Z}_{ik}^* = (\bar{Z}_{i1\cdot}^*, \ldots, \bar{Z}_{iq\cdot}^*)'$ their empirical means and let

$$\hat{v}_{i,\ell\ell}^* = N\frac{1}{n_i - 1}\sum_{k=1}^{n_i}(Z_{i\ell k}^* - \bar{Z}_{i\ell\cdot}^*)^2, \quad i = 1, 2 \text{ and } \ell = 1, \ldots, q \tag{12}$$

denote the empirical variance of the variables obtained under the ℓth condition. Now, the resampling version of the original test statistic T_0 is given by

$$T_0^* = \max\Big\{|T_1^*|, \ldots, |T_q^*|\Big\}, \quad \text{where } T_\ell^* = \sqrt{N}\frac{(\bar{Z}_{1\ell\cdot}^* - \bar{Z}_{2\ell\cdot}^*)}{\sqrt{\hat{v}_{1,\ell\ell}^*/n_1 + \hat{v}_{2,\ell\ell}^*/n_2}} \tag{13}$$

In comparison to the existing methodology discussed in Section 3, the statistic T_0^* mimics the computational process that lead to the original statistic T_0. Moreover, it is shown in the supplementary material that for both low- and high-dimensional settings, the conditional distribution of the vector of statistics (given the data $\boldsymbol{X}$)

$$\boldsymbol{T}^* = (T_1^*, \ldots, T_q^*)' \tag{14}$$

mimics the null distribution of $\boldsymbol{T}$. For making statistical inference, the critical value $z_{1-\alpha}(\max)$ is now estimated by the following steps:

1. Fix the data $\boldsymbol{X}$ (or $\boldsymbol{Y}$) and compute the centered variables $\boldsymbol{Z}_{ik}$.
2. Generate random weights W_{ik}, compute the resampling variables $\boldsymbol{Z}_{ik}^*$, the test statistics $\boldsymbol{T}^*$ and safe the value of T_0^* in $T_{0,b}^*$.
3. Repeat the previous step a large number of times (e.g. $M = 10,\ 000$) and compute the values $T_{0,1}^*, \ldots, T_{0,M}^*$.
4. Estimate $z_{1-\alpha}(\max)$ by the empirical $(1-\alpha)$-quantile $z_{1-\alpha}^*(\max)$ of $T_{0,1}^*, \ldots, T_{0,M}^*$.

Finally, the unknown quantile $z_{1-\alpha}(\max)$ is replaced with the observable value $z_{1-\alpha}^*(\max)$ in Equations (4) to (6), respectively. One-sided tests and p-values are estimated analogously. The estimation of $z_{1-\alpha}(\boldsymbol{R})$ thus gets by without estimating the full correlation matrix $\boldsymbol{R}$ and additionally takes the variability of the variance estimators into account. Note that the set $\{H_0^{(\ell)}, T_\ell, \ell = 1, \ldots, q\}$ consisting of the null hypotheses and corresponding test statistics constitutes a joint testing family in the sense of Gabriel.[31] Therefore, the simultaneous test procedure controls the family-wise error rate in the strong sense asymptotically in case of fixed q. Its accuracy in terms of controlling the type-1 error rate and power to detect alternatives when sample sizes are small will be investigated in the next section.

Remark: Throughout the manuscript, we consider the maximum statistic T_0 as a combination of the q possibly correlated test statistics only. It originates from finding appropriate real valued $c_{1-\alpha}$ such that

$$P\left(\bigcap_{\ell=1}^{q}\{-c_{1-\alpha} \le T_\ell \le c_{1-\alpha}\}\right) = 1 - \alpha \Longleftrightarrow P\left(\max_{\ell=1}^{q}|T_\ell| \le c_{1-\alpha}\right) = 1 - \alpha$$

The right-hand side holds with $z_{1-\alpha}(\max) = c_{1-\alpha}$. The resulting test further allows inversion of the test statistics into simultaneous confidence intervals. However, we note that other combining functions than computing the maximum statistic would be possible, for instance Fisher's weighted combining function. A general overview of nonparametric combination terminologies are provided in the monographs of Pesarin and Salmaso[32] and Salmaso et al.[27] and the references therein.

5 Simulations

In this section, we investigate the small sample properties of the proposed randomization technique within extensive simulation studies. The study aims to compare the two different approximations of the distribution of T_0 presented in the paper. As the true distribution of T_0 remains unknown, the type-1 error control of the competing methods will be used as a quality criterion. Later, the all-pairs and the any-pairs powers of the two methods will be compared. We conducted the extensive simulation studies in *R* (version 3.6.1). Marozzi[12] discusses different methods to compute the numbers *nsim* and *nboot* of simulation and resampling runs in detail. Using his result and under some assumptions, $nsim = 10{,}000$ simulations lead to $nboot = 8\sqrt{nsim} = 800$ resampling runs and a maximal simulation error of $0.006 \approx 1\%$. Since the methods proposed in the manuscript are computationally feasible, we chose $nsim = 10{,}000$ and $nboot = 1000$ runs for each setting. Furthermore, setting $\alpha = 5\%$, we can compute the 95% precision interval $[5\% \mp \frac{1.96}{\sqrt{1{,}000}}\sqrt{5\% * 95\%}] = [4.6\%; 5.4\%]$. If the empirical type-1 error of a test is within this interval, the method can be seen as accurate. The simulation code is displayed in the supplementary material file for reproducibility.

5.1 Type-1 error simulation results

Due to the abundance of possible factorial designs and hypotheses, two-way designs with varying dimension $d \in \{2, 4, \ldots, 150\}$ will be simulated and the hypothesis $H_0 : \boldsymbol{P}_d(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$ of *no interaction effect* will be tested at 5% level of significance. Data were generated from model

$$\boldsymbol{X}_{ik} \sim \boldsymbol{F}_i(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_i) + \boldsymbol{\mu}_i,\ i = 1, 2,\ k = 1, \ldots, n_i, \tag{15}$$

where $\boldsymbol{F}_i(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_i)$ represents a multivariate distribution with expectation vector $\boldsymbol{\mu}_0$, correlation matrix $\boldsymbol{\Sigma}_i$ and location shifts $\boldsymbol{\mu}_i$. As representative marginal data distributions, we selected three differently tailed symmetric distributions (normal, logistic, T_3) and three skewed distributions (ranging from mildly to very skewed) (χ^2_7, χ^2_{15}, exponential) each with sample sizes $n_i \in \{10, 20\}$. A major assessment criteria of the quality of the proposed approximations is the impact of both the chosen contrast as well as the dependency structures of the data—especially when data has different covariance matrices and thus covering a typical Behrens-Fisher situation. Here, we used normal copulas in order to generate rather complex dependency structures of the repeated measurements using the *R*-package *copula*.[33] The different allocations of the correlation matrices used in the simulation studies are summarized in Table 2.

In Setting 1, both correlation matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are identical and represent an autoregressive structure. In Settings 2 and 3, the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ have different off-diagonal elements models, whereas an autoregressive structure depending on the dimension d is modeled by $\boldsymbol{\Sigma}_1$ in Setting 2, and a linearly decreasing (symmetric) Toeplitz structure is covered by $\boldsymbol{\Sigma}_1$ in Setting 3 (see Table 2), see Pauly et al.[7] for similar choices. Note that Setting 2 models a pretty extreme scenario. For a detailed overview of copulas, we refer to Nelsen[34] or Marozzi.[35]

All these four settings will be simulated for all four sample sizes ($n_i \in \{10, 20\}$), dimensions ($d \in \{2, 4, \ldots, 150\}$) and distributional configurations as described above. The type-1 error simulation results obtained under Setting 1 are displayed in Figure 2. All others are displayed in the supplementary material file.

Table 2. Different correlation matrices used in the simulation study.

Setting 1:	$\boldsymbol{\Sigma}_1 = (\sigma_{1,ij}) = 0.6^{\lvert i-j \rvert}$	$\boldsymbol{\Sigma}_2 = (\sigma_{2,ij}) = 0.6^{\lvert i-j \rvert}$
Setting 2:	$\boldsymbol{\Sigma}_1 = (\sigma_{1,ij}) = 0.6^{\lvert i-j \rvert/(d-1)}$	$\boldsymbol{\Sigma}_2 = (\sigma_{2,ij}) = 0.6^{\lvert i-j \rvert}$
Setting 3:	$\boldsymbol{\Sigma}_1 = (\sigma_{1,ij}) = 1 - \lvert i-j \rvert/d$	$\boldsymbol{\Sigma}_2 = (\sigma_{2,ij}) = 0.6^{\lvert i-j \rvert/(d-1)}$
Setting 4:	$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_d + 0.5 \cdot (\boldsymbol{J}_d - \boldsymbol{I}_d)$	$\boldsymbol{\Sigma}_2 = \boldsymbol{I}_d + 0.25 \cdot (\boldsymbol{J}_d - \boldsymbol{I}_d)$.

First, it can be seen that the underlying covariance matrices significantly impact the accuracy of the simulation-based procedure proposed by Chang et al.[14] in small sample size situations. It can also readily be seen that this test shows an increasing liberal behavior for increasing dimension d. The over-rejection of the hypotheses occurs, because the test decision is based upon quantiles from the $N(0, \hat{\boldsymbol{R}})$ distribution, which neither takes the variability nor the distribution of the variance estimators into account. On the contrary, the randomization-based test T_0 in Equation (13) tends to control the nominal type-1 error rate very well, even in case of very small sample sizes and large dimensions. The underlying covariance structures seem to impact the results only minor (if even). In case of mildly skewed distributions, the simulation results indicate that the resampling test controls the type-1 error accurately. However, in case of skewed data with different covariance matrices, the test might be very liberal when sample sizes are small. This behavior especially depends on the type of contrast of interest and whether it induces positive or negative correlations. The simulation results obtained under Setting 2 (see the supplementary material) indicate that none of the methods should be applied in these (rather extreme) cases in practice. Other methods, e.g. nonparametric methods rather than mean-based procedures, might be more appropriate for the analysis of small skewed data in such cases. However, the liberality vanishes with increasing sample sizes.[36] In the three other settings considered here, the randomization procedure controls the type-1 error rate accurately, even when sample sizes are small and data follow skewed distributions. Pauly et al.[24] report similar conclusions for general linear models with independent observations. Furthermore, in all of the settings considered here, the randomization-based resampling method is accurate when data is heavy-tailed but symmetric. As many factors might impact the behavior of the tests, however, the procedure might be sensitive to such distributional shapes in different scenarios than the ones considered here. For the generation of other copula models, e.g. Elliptical and Archimedean copulas, we refer to Marozzi.[35]

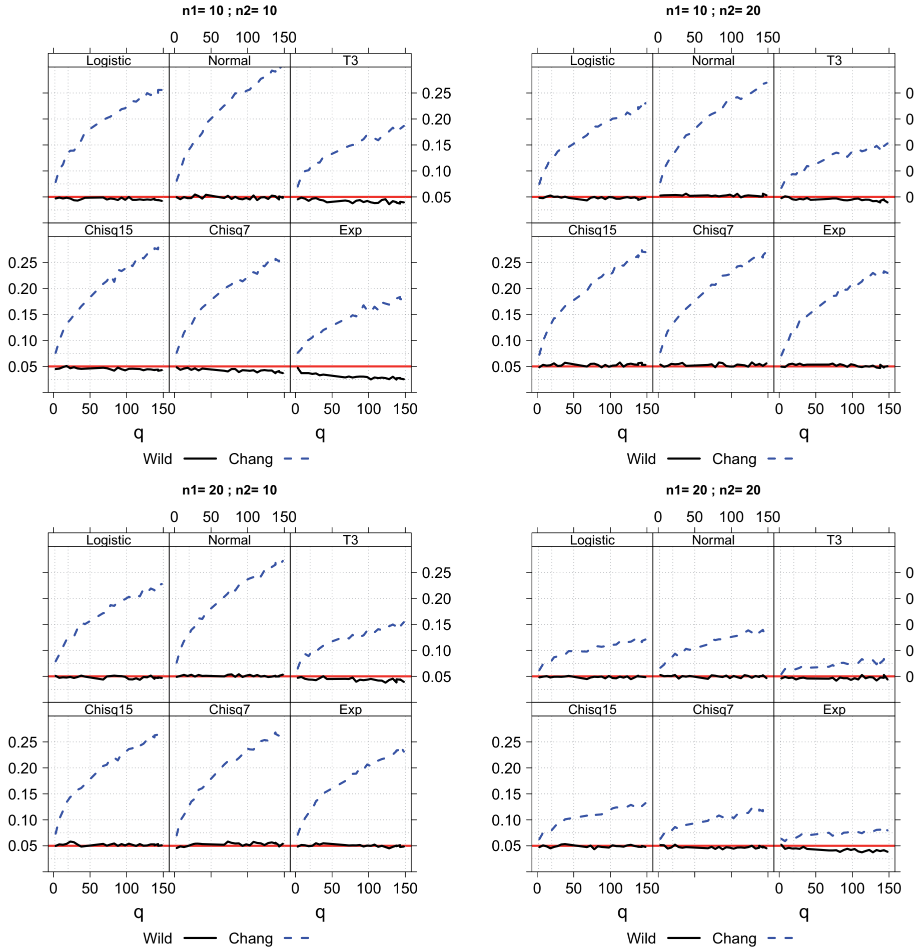

Figure 2. Type-1 error ($\alpha = 5\%$) simulation results of the Wild-bootstrap randomization test $\boldsymbol{T}$ in Equation (14) (*Wild*) and simulation-based test $\boldsymbol{T}$ in Equation (10) (*Chang*). Data have covariance matrices as described in Setting 1 in Table 2.

Remark: Instead of using copulas for generation of the multivariate distributions, an alternative method is generating data from model

$$\boldsymbol{X}_{ik} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{-1/2}\epsilon_{ik},\ i = 1, 2,\ k = 1, \ldots, n_i$$

where the error terms is generated from standardized distributions, respectively. Additional s”imulation studies indicate that the empirical behavior of the test procedures is very similar.

5.2 Power simulation results

Next the *all-pairs* power P (“reject all false null hypotheses”) as well as the *any-pairs* powers P (“reject any true or false null hypothesis”) of the competing methods to reject the null hypothesis $H_0 : \boldsymbol{P}_d(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$ ($\alpha = 5\%$) will be simulated for selected alternatives. The aim of the simulation study is investigating the impact of the underlying distributions, dependency structures of the data, sample size allocations and dimensions on the powers of the tests. Data have been generated (under the alternative) from model Equation (15) with expectations

$$\boldsymbol{\mu}_1 = (\delta, 2\delta, \delta, 0, \ldots, 0)' \text{ and } \boldsymbol{\mu}_2 = (2\delta, \delta, 2\delta, 0, \ldots, 0)'$$

and varying $\delta \in \{0, 0.1, \ldots, 2\}$ from the same six distributions as above (normal, logistic, T_3, χ^2_7, χ^2_{15} and exponential) having all of the four different dependency structures displayed in Table 2, respectively. The dimension of the random vectors was set to $d = 30$. Due to the liberality of Chang et al.’s method for small sample sizes, large sample sizes were simulated ($n_i = 100$) in order to be able to compare the powers of the methods on a fair basis, i.e. in a situation where both of them control the type-1 error rate accurately. For illustration, an additional power simulation with small sample sizes ($n_i = 10$) has been conducted. First, it turns out that the types of covariance structures affect the powers of the tests. This is not surprising, because the higher the correlation the smaller are the variances of the effect size estimators. Overall, the simulation results indicate that the competing methods have comparable powers when sample sizes are large. Chang et al.’s method has slightly larger any-pairs and all-pairs powers than the randomization test (about 1% higher). However, when sample sizes are small, the randomization method controls the size and has a reasonable power. The simulations of the all-pairs power furthermore indicate the strong control of the FWER of both methods. Under the situations considered here, the shapes of the underlying distributions impact the results. As expected, the power of the methods under χ^2-distributions appears to be rather low. The reason is the pretty large variance of the χ^2-distribution compared with the other distributions. It should be noted that the above findings only hold for the settings considered here and might be different under other scenarios. The all-pairs and the any-pairs power curves are displayed in the supplementary material file.

6 Evaluation

The extensive simulation studies show that the newly proposed randomization test controls the type-1 error rate very satisfactorily, even when sample sizes are very small and data do not follow a multivariate normal distribution. In a first step, we perform a further type-1 error simulation study and investigate the accuracy of the method for analyzing this specific data set in the same way as presented in Table 1. As before in Equation (11), we mimic the data set using multivariate normal $N(\boldsymbol{0}, \hat{\boldsymbol{V}}_i)$ and multivariate $T_3(\boldsymbol{0}, \hat{\boldsymbol{V}}_i)$ distributions, respectively. The type-1 error simulation results are displayed in Table 3. It appears that the method controls the type-1 error accurately. The high-dimensional preclinical study on Alzheimer’s disease introduced in Section 2 can therefore now be analyzed with this method. For comparisons, the data set will be analyzed using both of the discussed approximations.

Table 3. Type-1 error ($\alpha = 5\%$) simulation results of the Wild-bootstrap randomization test.

Dist$\backslash n_i$	8	9	10	15	20	30	40	50
Nor	0.0530	**0.0512**	0.0494	0.0496	0.0474	0.0502	0.0501	0.0497
T_3	0.0392	**0.0411**	0.0431	0.0464	0.0458	0.0512	0.0527	0.0502

significance is provided with alpha = 5%

In addition to testing for interactions motivated in Equation (2), multiple comparisons inferring the region as well as the group effects are of interest. These will be performed using the contrast matrices

$$H_0^{\mu} : \boldsymbol{C}\boldsymbol{\mu} = 0, \text{ where } \boldsymbol{C} = \left(\oplus_{s=1}^{6} \boldsymbol{P}_6\right)\left(\boldsymbol{I}_{36} \vdots \boldsymbol{I}_{36}\right) \text{ (Region), and}$$

$$H_0^{\mu} : \boldsymbol{C}\boldsymbol{\mu} = 0, \text{ where } \boldsymbol{C} = \left(\boldsymbol{I}_{36} \vdots -\boldsymbol{I}_{36}\right) \text{ (Group)} \tag{16}$$

Note that for testing the impact of the region, the test statistics are given by

$$T_\ell = \sqrt{N} \frac{\bar{Y}_{1\ell\cdot} + \bar{Y}_{2\ell\cdot} - (\theta_{1\ell} + \theta_{2\ell})}{\sqrt{\hat{v}_{1,\ell\ell}/n_1 + \hat{v}_{2,\ell\ell}/n_2}}$$

where $\bar{Y}_{i\ell\cdot}$ denotes the mean of the ℓth component of the vector $\boldsymbol{Y}_{ik}{}' = \boldsymbol{X}_{ik}{}'\left(\oplus_{s=1}^{6} \boldsymbol{P}_6\right)$. Therefore, the correlation matrix of the vector of test statistics $\boldsymbol{T} = (T_1, \ldots, T_q)'$ is identical to the one using interaction contrasts as described in Section 3. The randomization approach for approximating the joint null distribution of these statistics is adapted accordingly. Testing for the group effects is the "standard" multivariate hypothesis. Means and empirical variances of the protein abundance data under each protein × region × group combination are provided in the supplementary material file.

As already indicated by the confidence interval plots in Figure 1, data show a fairly amount of variance heteroscedasticity. Therefore, assuming equal covariance matrices across the groups is doubtful. Next, the multiple hypotheses will be tested using the two different approaches. The point estimators of the contrasts in means $\hat{\delta}_\ell$, values of the test statistics T_ℓ equation (3), the estimated quantile $z_{95\%}(\max)$ as well as 95%-simultaneous confidence intervals equation (5) using both the simulation as well as randomization technique will be displayed for all of the three multiple hypotheses. In total, $M = 100{,}000$ simulation and randomization runs have been performed. The results are displayed in Table 4. Different decisions (at 5% level) between the two competing methods are highlighted in boldface.

First, for all of the three different testing problems, the estimated quantiles of the maximum statistic $z_{95\%}(\max)$ are way larger using the randomization approach than with the simulation-based method. This is not surprising when reflecting the liberal behavior of the test. The simultaneous confidence intervals are therefore wider using the randomization procedure. In the following, results obtained for each of the three multiple null hypotheses will be discussed separately. Neither of the two competing methods detects an interaction between group and region under any of the six investigated proteins. The estimated quantiles differ remarkably, though (3.10 vs. 3.76). But, since the maximum t-statistic is $T_0 = 2.63$ and thus $T_0 < z_{95\%}(\max)$, data do not provide the evidence to reject the null hypothesis at 5% level of significance. Investigating differences in the regions, the approximation methods provide different local conclusions at 5% level of significance. The simulation-based method declares the regions ACB under Syntaxin, M1 under VAMP2 as well as M1 under Alpha-synuclein significantly different from the average of the others, while the randomization method does not. Taking a look at the boxplots give the impression that these values differ only slightly from the mean of the others. Clearly, given the amount of regional deviations, those regions differ significantly on a pairwise level. Also, overall, the protein abundances differ significantly across the regions. Investigating differences between the groups, no significant differences can be detected using any of the competing methods. It should be noted that the estimated quantile using the randomization method increases from 3.75 to a value of 3.88, while the simulation-based estimator is still about 3.10 (3.09).

7 Discussion and outlook

Research experiments in translational medicine and especially in preclinical areas are usually small due to ethical reasons and animal welfare. Clearly, animal studies should be abandoned, but, roughly speaking, medical research has not arrived at the point yet to replace and refine every experiment to avoid animals. Often, sample sizes of such trials are smaller than 20 per experimental group, which might be a reason to argue the quality of the outcome. However, since animals are kept under homogeneous conditions, heteroscedasticity across the animals is usually smaller compared to other scenarios in humans, depending on the outcome measures. Anyway, since

Table 4. Multiple contrast test results of the protein abundance trial.

	Interaction						Region						Group					
	Chang ($z_{95\%} = 3.10$)			Randomization ($z_{95\%} = 3.76$)			Chang ($z_{95\%} = 3.10$)			Rand. ($z_{95\%} = 3.75$)			Chang ($z_{95\%} = 3.09$)			Rand. ($z_{95\%} = 3.88$)		
$\ell =$	$\hat{\delta}_\ell$	T_ℓ	95%-L	95%-U	95%-L	95%-U	$\hat{\delta}_\ell$	T_ℓ	95%-L	95%-U	95%-L	95%-U	$\hat{\delta}_\ell$	T_ℓ	95%-L	95%-U	95%-L	95%-U
1	−0.65	−1.20	−2.32	1.03	−2.67	1.37	1.35	2.50	−0.32	3.03	−0.68	3.38	0.39	0.41	−2.57	3.36	−3.32	4.11
2	−0.53	−1.00	−2.17	1.11	−2.50	1.44	2.49	4.71	0.85	4.13	0.51	4.48	0.51	0.69	−1.77	2.80	−2.35	3.38
3	1.09	1.37	−1.37	3.55	−1.88	4.05	−2.38	−3.00	−4.84	0.07	−5.36	0.59	2.13	1.93	−1.28	5.55	−2.15	6.42
4	−0.04	−0.10	−1.41	1.32	−1.69	1.60	−2.40	−5.43	−3.77	−1.03	−4.06	−0.74	1.00	1.06	−1.92	3.91	−2.66	4.66
5	−0.01	−0.02	−1.39	1.37	−1.67	1.66	−0.68	−1.53	−2.06	0.70	−2.36	0.99	1.04	1.03	−2.08	4.15	−2.87	4.94
6	0.14	0.28	−1.41	1.70	−1.74	2.02	1.62	3.23	**0.06**	**3.18**	**−0.26**	**3.51**	1.18	1.20	−1.85	4.21	−2.62	4.99
7	−0.22	−0.72	−1.15	0.71	−1.34	0.91	0.18	0.58	−0.76	1.11	−0.95	1.30	−0.61	−1.84	−1.64	0.41	−1.90	0.68
8	−0.68	−2.50	−1.53	0.16	−1.71	0.34	−1.85	−6.75	−2.70	−1.00	−2.88	−0.82	−1.08	−2.90	−2.23	0.07	−2.53	0.36
9	0.15	0.74	−0.49	0.80	−0.63	0.94	−0.10	−0.50	−0.75	0.54	−0.89	0.68	−0.24	−0.89	−1.09	0.60	−1.30	0.82
10	0.20	0.98	−0.43	0.83	−0.56	0.96	−1.97	−9.69	−2.60	−1.34	−2.74	−1.21	−0.20	−0.85	−0.91	0.52	−1.10	0.70
11	0.24	1.41	−0.28	0.75	−0.39	0.86	1.10	6.58	0.58	1.62	0.47	1.73	−0.16	−0.61	−0.98	0.66	−1.19	0.87
12	0.31	2.42	−0.09	0.71	−0.17	0.79	2.65	20.58	2.25	3.05	2.17	3.13	−0.09	−0.77	−0.43	0.25	−0.51	0.34
13	0.46	1.44	−0.53	1.45	−0.74	1.66	2.62	8.18	1.63	3.62	1.42	3.83	0.59	1.66	−0.51	1.69	−0.79	1.97
14	−0.07	−0.31	−0.79	0.64	−0.93	0.79	3.73	16.17	3.02	4.45	2.87	4.60	0.06	0.24	−0.70	0.82	−0.90	1.02
15	−0.23	−0.29	−2.67	2.21	−3.18	2.72	−6.67	−8.45	−9.12	−4.23	−9.63	−3.71	−0.10	−0.09	−3.37	3.17	−4.20	4.00
16	−0.41	−1.80	−1.13	0.30	−1.27	0.45	−3.98	−17.27	−4.69	−3.26	−4.84	−3.11	−0.28	−0.62	−1.70	1.14	−2.07	1.50
17	0.24	0.55	−1.11	1.59	−1.39	1.87	1.46	3.34	**0.11**	**2.81**	**−0.18**	**3.10**	0.37	0.98	−0.80	1.54	−1.10	1.84
18	0.01	0.04	−0.91	0.94	−1.10	1.13	2.83	9.47	1.90	3.75	1.71	3.95	0.14	0.45	−0.84	1.13	−1.09	1.38
19	−0.57	−1.11	−2.14	1.01	−2.47	1.34	−1.78	−3.49	−3.36	−0.20	−3.69	0.13	−0.86	−1.41	−2.74	1.02	−3.22	1.49
20	−1.08	−1.89	−2.86	0.69	−3.22	1.05	−0.50	−0.88	−2.28	1.27	−2.65	1.64	−1.38	−1.83	−3.70	0.94	−4.29	1.53
21	0.54	1.62	−0.49	1.56	−0.70	1.77	0.58	1.75	−0.44	1.61	−0.66	1.82	0.24	0.76	−0.74	1.23	−0.99	1.48
22	−0.17	−0.56	−1.14	0.79	−1.34	0.99	−3.14	−10.09	−4.10	−2.17	−4.30	−1.97	−0.47	−1.52	−1.42	0.48	−1.66	0.72
23	0.72	1.35	−0.93	2.37	−1.27	2.71	1.05	1.97	−0.60	2.70	−0.95	3.05	0.43	0.62	−1.70	2.55	−2.24	3.09
24	0.57	2.32	−0.19	1.32	−0.35	1.48	3.79	15.52	3.03	4.55	2.87	4.71	0.27	0.86	−0.70	1.25	−0.95	1.49
25	0.45	1.16	−0.75	1.65	−1.00	1.90	3.31	8.52	2.11	4.52	1.85	4.77	−0.16	−2.19	−0.39	0.07	−0.45	0.13
26	0.30	0.72	−1.00	1.61	−1.27	1.88	2.80	6.63	1.49	4.11	1.21	4.38	−0.31	−2.55	−0.69	0.07	−0.78	0.16
27	−0.92	−1.93	−2.39	0.56	−2.70	0.86	−2.17	−4.56	−3.65	−0.70	−3.96	−0.38	−1.53	−2.05	−3.84	0.78	−4.43	1.37
28	−0.76	−1.40	−2.44	0.92	−2.78	1.26	−5.50	−10.16	−7.18	−3.83	−7.54	−3.47	−1.37	−1.54	−4.13	1.38	−4.83	2.08
29	0.59	2.64	−0.10	1.28	−0.25	1.43	2.05	9.18	1.36	2.75	1.21	2.89	−0.02	−0.08	−0.95	0.90	−1.18	1.13
30	0.33	0.86	−0.87	1.53	−1.11	1.78	−0.49	−1.26	−1.68	0.71	−1.94	0.96	−0.28	−0.39	−2.50	1.93	−3.07	2.50
31	0.11	0.52	−0.55	0.78	−0.69	0.92	0.45	2.08	−0.22	1.11	−0.36	1.26	0.40	1.87	−0.26	1.06	−0.43	1.23
32	−0.22	−0.93	−0.94	0.50	−1.09	0.65	1.07	4.59	0.35	1.79	0.20	1.94	0.07	0.37	−0.51	0.65	−0.66	0.80
33	0.31	0.96	−0.69	1.31	−0.89	1.52	−2.18	−6.77	−3.18	−1.18	−3.40	−0.97	0.60	1.52	−0.62	1.82	−0.93	2.12
34	0.22	0.81	−0.61	1.05	−0.79	1.22	−3.10	−11.54	−3.93	−2.27	−4.11	−2.09	0.51	1.60	−0.47	1.48	−0.72	1.73
35	−0.43	−1.81	−1.16	0.31	−1.32	0.46	0.85	3.59	**0.12**	**1.59**	**−0.04**	**1.75**	−0.14	−0.50	−1.00	0.72	−1.22	0.94
36	0.01	0.06	−0.41	0.42	−0.49	0.51	2.91	21.80	2.50	3.33	2.41	3.41	0.30	2.02	−0.16	0.75	−0.27	0.86
	$T_0 = 2.63$						$T_0 = 21.80$						$T_0 = 2.90$					

Here, $\ell = 1, \ldots, 36$ corresponds to each of the 36 contrasts tested by the hypothesis of no interaction, region or group as given in Equation (2) or Equation (16), respectively. In addition, $\hat{\delta}_\ell$ is the estimated contrast, T_ℓ the value of the *t*-test type statistic and 95%-L or 95%-U the 95% lower or upper bound of the 95% simultaneous confidence intervals. The estimated quantiles $z_{95\%}(\max)$ are given in the headers. Different test decisions between the competing methods are marked boldface.

preclinical studies play a significant role in medical sciences in terms of transferring the results towards the next phase, a major concern is the quality of the used statistical methods. Most of them control the type-1 error rate accurately with large sample sizes only and, indeed, they show a very liberal or conservative behavior when sample sizes are small. This observation holds for a variety of statistical procedures designed for different questions and fields, including analysis of variance methods[24,37] as well as multiple contrast test procedures using maximum *t*-test type statistics[38,39] for repeated measures and multivariate data. When the number of comparisons is "small" compared to the sample sizes, approximate and exact methods are available.[19,40–43] These methods are, however, limited to the number of comparisons to be made and are not applicable in high-dimensional situations. Note that the methods are not applicable because of the test statistic itself (maximum *t*-test) or because of any computational difficulty, and information about its distribution is only available for low-dimensional designs. Recently, Chang et al.[14] tackled the problem and proposed a simulation-based algorithm to approximate the distribution of the maximum statistic in high-dimensional designs. Extensive simulations show, however, that large sample sizes are needed for an accurate type-1 error rate control making their method not applicable for trials with small sample sizes. In the present paper, we modified their strategy towards a robust randomization technique to approximate the null distribution of the max *t*-test. Simulation studies indicate that the method approximates the null distribution very satisfactorily and greatly improves the applicability of their method.

Furthermore, the power of the method can be considerably improved by adapting it to a two-step screening procedure. For a given significance level α, define the index set

$$\mathcal{S}_1 = \{1 \leq \ell \leq q : |T_\ell| \leq \sqrt{2\log(q)} + \{2\log(q)\}^{-1/2} + \sqrt{2\log(1/(1-\alpha))}\} \tag{17}$$

which contains the indices of all test statistics T_ℓ that—in absolute value—do not exceed the given bound in $\mathcal{S}_1$. If the cardinality of the set is equal to q, the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ is not rejected. Otherwise, if its cardinality is equal to s (say) and smaller than q ($s < q$), select all corresponding test statistics in the sub-vector $\tilde{\boldsymbol{T}} = \{T_s, s \notin \mathcal{S}_1\}$. The null hypothesis will be rejected, if $\max\{|T_s|, s \notin \mathcal{S}_1\} \geq y^*_{1-\alpha}(\mathcal{S}_1)$. Here, $y^*_{1-\alpha}(\mathcal{S}_1)$ denotes the $(1-\alpha)$-quantile of the values $\max_{s\notin\mathcal{S}_1}|Y^*_{1s}|, \ldots, \max_{s\notin\mathcal{S}_1}|Y^*_{Ms}|$, where the Y^*'s are defined in equation (10). Thus, the dimension of the testing problem is basically reduced to testing the null hypotheses $H_0^{(s)} : \tilde{\boldsymbol{\theta}}_1 = \tilde{\boldsymbol{\theta}}_2$, where the hypothesis matrix $\tilde{\boldsymbol{C}}$ has appropriate dimensions and $\tilde{\boldsymbol{\mu}}$ collects all corresponding μ_i's being excluded from $\mathcal{S}_1$. Moreover, the dimension reduction implies that the critical value of the screening modification does not exceed the original one, i.e. $y_1-\alpha^*(\mathcal{S}_1) \leq y^*_{1-\alpha}$. As the value of the test statistic does not change ($\max\{|T_s|, s \notin \mathcal{S}_1\} = T_0$ by definition), screening indeed improves the power. For the same reason, however, screening might result in even more liberal test decisions than the original version without screening. In addition, the interpretation of the screening results may be challenging in case of arbitrary contrasts or especially interaction effects. For these two reasons, we did not consider an additional screening stage. Detailed power investigations and dimension reductions will be part of future research. All of the methods considered in the paper are based on means of data. Nonparametric methods, for instance based on ranks of the data, or simultaneous methods based on interpoint distances as well as investigating other combination functions than maximum,[12] will be part of further investigations as well.

Acknowledgements

The authors are grateful to the Editor, Associate Editor and three anonymous referees for their helpful suggestions, which greatly improved the manuscript.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research is supported by the German Science Foundation awards number DFG KO 4680/3-2 and PA 2409/3-2.

ORCID iD

Frank Konietschke https://orcid.org/0000-0002-5674-2076

Supplemental material

Supplemental material for this article is available online. It contains theoretical investigations, all proofs, simulation results as well as software code.

References

1. Ahmad MR, Werner C and Brunner E. Analysis of high-dimensional repeated measures designs: the one sample case. *Computat Stat Data Analys* 2008; **53**: 416–427.
2. Chen SX and Qin YL. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann Stat* 2010; **38**: 808–835.
3. Brunner E, Bathke AC and Placzek M. Estimation of box's ε for low-and high-dimensional repeated measures designs with unequal covariance matrices. *Biometric J* 2012; **54**: 301–316.
4. Secchi P, Stamm A, Vantini S, et al. Inference for the mean of large p small n data: A finite-sample high-dimensional generalization of hotelling's theorem. *Electronic J Stat* 2013; **7**: 2005–2031.
5. Cai T, Liu W and Xia Y. Two-sample test of high dimensional means under dependence. *J Royal Stat Soc: Ser B (Stat Methodol)* 2014; **76**: 349–372.
6. Gregory KB, Carroll RJ, Baladandayuthapani V, et al. A two-sample test for equality of means in high dimension. *J Am Stat Assoc* 2015; **110**: 837–849.
7. Pauly M, Ellenberger D and Brunner E. Analysis of high-dimensional one group repeated measures designs. *Statistics* 2015; **49**: 1243–1261.
8. Sattler P, Pauly M, et al. Inference for high-dimensional split-plot-designs: A unified approach for small to large numbers of factor levels. *Electronic J Stat* 2018; **12**: 2743–2805.
9. Baringhaus L and Franz C. On a new multivariate two-sample test. *J Multivariate Analysis* 2004; **88**: 190–206.
10. Rosenbaum PR. An exact distribution-free test comparing two multivariate distributions based on adjacency. *J Royal Stat Soc: Ser B (Stat Methodol)* 2005; **67**: 515–530.
11. Jurečková J and Kalina J. Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli* 2012; **18**: 229–251.
12. Marozzi M. Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Stat Meth Med Res* 2016; **25**: 2593–2610.
13. Hasler M and Hothorn LA. Multiple contrast tests in the presence of heteroscedasticity. *Biometric J* 2008; **50**: 793–780.
14. Chang J, Zheng C, Zhou WX, et al. Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics* 2017; **73**: 1300–1310.
15. Hall P and Wilson SR. Two guidelines for bootstrap hypothesis testing. *Biometrics* 1991; **47**: 757–762.
16. Janssen A. Resampling student's t-type statistics. *Ann Institute Stat Math* 2005; **57**: 507–529.
17. Delaigle A, Hall P and Jin J. Robustness and accuracy of methods for high dimensional data analysis based on student's t-statistic. *J Royal Stat Soc: Ser B (Stat Methodol)* 2011; **73**: 283–301.
18. Konietschke F and Pauly M. Bootstrapping and permuting paired t-test type statistics. *Stat Comput* 2014; **24**: 283–296.
19. Hasler M. Multiple contrasts for repeated measures. *Int J Biostat* 2013; **9**: 49–61.
20. Mi X, Miwa T and Hothorn T. mvtnorm: New numerical algorithm for multivariate normal probabilities. *R J* 2009; **1**: 37–39.
21. Genz A, Bretz F, Miwa T, et al. *Package mvtnorm. J Computat Graph Stat* 2019; **11**: 950–971.
22. Troendle JF, Korn EL and McShane LM. An example of slow convergence of the bootstrap in high dimensions. *Am Stat* 2004; **58**: 25–29.
23. Huang Y, Xu H, Calian V, et al. To permute or not to permute. *Bioinformatics* 2006; **22**: 2244–2248.
24. Pauly M, Brunner E and Konietschke F. Asymptotic permutation tests in general factorial designs. *J Royal Stat Soc: Ser B* 2015; **77**: 461–473.
25. Pesarin F and Salmaso L. Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *J Nonparametric Stat* 2010; **22**: 669–684.
26. Pesarin F and Salmaso L. On the weak consistency of permutation tests. *Communicat Stat-Simulat Computat* 2013; **42**: 1368–1379.
27. Salmaso L. Combination-based permutation tests: Equipower property and power behavior in presence of correlation. *Commun Stat-Theory Meth* 2015; **44**: 5225–5239.
28. Pesarin F and Salmaso L. A review and some new results on permutation testing for multivariate problems. *Stat Comput* 2012; **22**: 639–646.
29. Janssen A. Nonparametric symmetry tests for statistical functionals. *Math Meth Stat* 1999; **8**: 320–343.
30. Mammen E. Bootstrap and wild bootstrap for high dimensional linear models. *Ann Stat* 1993; **21**: 255–285.
31. Gabriel KR. Simultaneous test procedures–some theory of multiple comparisons. *Ann Math Stat* 1969; **4**: 224–250.
32. Pesarin F and Salmaso L. *Permutation tests for complex data: theory, applications and software*. Hoboken, NJ: John Wiley & Sons, 2010.
33. Yan J et al. Enjoy the joy of copulas: with a package copula. *J Stat Software* 2007; **21**: 1–21.

34. Nelsen RB. *An introduction to copulas*. New York, NY: Springer Science & Business Media, 2007.
35. Marozzi M. Multivariate multidistance tests for high-dimensional low sample size case-control studies. *Stat Med* 2015; **34**: 1511–1526.
36. Davidson R and Flachaire E. The wild bootstrap, tamed at last. *J Econometrics* 2008; **146**: 162–169.
37. Dobler D, Friedrich S and Pauly M. Nonparametric MANOVA in meaningful effects. *Ann Inst Stat Math* 2019; **72**: 1–26.
38. Gunawardana A and Konietschke F. Nonparametric multiple contrast tests for general multivariate factorial designs. *J Multivariate Analys* 2019; **173**: 165–180.
39. Umlauft M, Placzek M, Konietschke F, et al. Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *J Multivariate Analys* 2019; **171**: 176–192.
40. Hothorn T, Bretz F and Westfall P. Simultaneous inference in general parametric models. *Biometric J* 2008; **50**: 346 –363.
41. Liu W, Ah-Kine P, Bretz F, et al. Exact simultaneous confidence intervals for a finite set of contrasts of three, four or five generally correlated normal means. *Computat Stat Data Analys* 2013; **57**: 141–148.
42. Hasler M. Multiple contrast tests for multiple endpoints in the presence of heteroscedasticity. *Int J Biostat* 2014; **10**: 17–28.
43. Hasler M and Böhlendorf K. Multiple comparisons for multiple endpoints in agricultural experiments. *J Agri Biol Environ Stat* 2013; **18**: 1–16.