Mathematical Statistics
Stockholm University

Multivariate multiple test procedures

Thorsten Dickhaus

André Neumann

Taras Bodnar

Research Report 2019:17

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2019:17**,
http://www.math.su.se

Multivariate multiple test procedures

Thorsten Dickhaus[a], André Neumann[a], and Taras Bodnar[b]

[a]Department of Mathematics, University of Bremen, Bremen, Germany

[b]Department of Mathematics, Stockholm University, Stockholm, Sweden

December 2019

Keywords: multiple test; copula; effective numbers of tests; FWER; FDR

1 Introduction and preliminaries

Dependencies among data points are present in virtually all modern statistical applications. This holds especially true for studies with multiple endpoints which are all measured for the same observational units. For example, consider the case of a gene expression study. In that context, expression levels of m genes are measured for n individuals. The goal of the study typically is to detect statistically significant expression differences, either in a two-groups model or in a one-group model under different experimental conditions. Due to biological and technological reasons, the expression levels will typically exhibit strong dependencies, at least for genes which are functionally related; cf. [67]. We will discuss multiple tests which take such dependencies explicitly into account. Such multiple tests are called multivariate multiple tests, because they rely on joint distributions or on approximations thereof.

1.1 Motivation

Example 1. *As a simple motivating example for utilizing a multivariate multiple test, consider the case of $m = 2$ simultaneous tests for Gaussian means. Let $Z = (Z_1, Z_2)^\top$ denote an observable $\mathbb{R}^2$-valued random vector which follows the bivariate normal distribution with an unknown mean vector $\mu = (\mu_1, \mu_2)^\top$, but a known covariance (and correlation) matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where $|\rho| < 1$. Consider the two (one-sided) null hypotheses $H_j = \{\mu_j \leq \mu_j^*\}$, $j = 1, 2$, for a given vector $\mu^* = (\mu_1^*, \mu_2^*)^\top \in \mathbb{R}^2$. The corresponding alternative hypotheses are given by $K_j = \{\mu_j > \mu_j^*\}$, $j = 1, 2$. This is a typical setup for a*

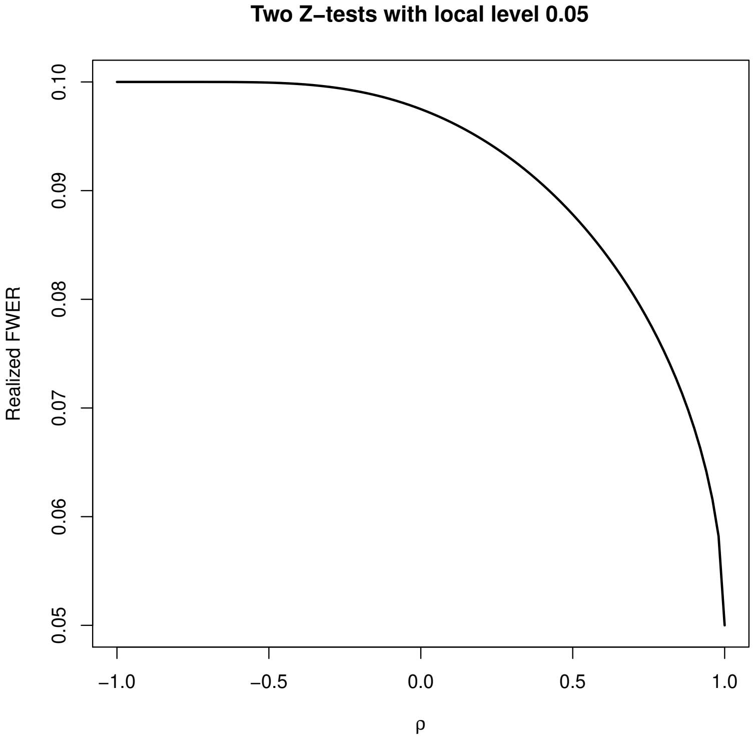

Figure 1: Realized FWER in the case of two simultaneous Z-tests, where the two test statistics are jointly normally distributed with correlation coefficient ρ. Details are provided in Example 1.

multiple test for superiority in a clinical study concerning two endpoints. Let α_{loc} *denote a local significance level, meaning that the two tests* φ_1 *and* φ_2 *for testing* H_1 *and* H_2 *are carried out at level* α_{loc} *each. It is canonical to use* $\varphi_j = \mathbf{1}_{(c,\infty)}(Z_j - \mu_j^*)$, $j = 1, 2$, *with* $c = \Phi^{-1}(1 - \alpha_{loc})$ *denoting the* $(1 - \alpha_{loc})$*-quantile of the univariate standard normal distribution. The Bonferroni correction for family-wise error rate (FWER) control at level* α *would advise us to take* $\alpha_{loc} = \alpha/2$.

Figure 1 displays the effect of choosing $\alpha_{loc} = 5\%$ *for varying values of the correlation coefficient* ρ *in the case that* $\mu = \mu^*$, *implying that both null hypotheses* H_1 *and* H_2 *are true. For strong negative correlations among* Z_1 *and* Z_2 $(\rho \to -1)$, *the realized FWER tends to* $\alpha = 10\% = 2\alpha_{loc}$, *meaning that the Bonferroni inequality becomes an equality for* $\rho \to -1$. *On the other hand, for positive* ρ *the realized FWER is below* α, *and it even monotonically decreases to* $\alpha_{loc} = 5\% = \alpha/2$ *for* $\rho \to +1$, *meaning that in the extreme case of perfect positive correlation among* Z_1 *and* Z_2 *no adjustment for multiplicity would be necessary at all.*

Since ρ *is assumed to be known here, the exhaustion of the FWER level* α *and, hence, the power of the multiple test could be improved by choosing* α_{loc} *by means of a quantile* *of the bivariate joint distribution of* $T = (T_1, T_2)^\top$ *under* $\mu = \mu^*$*, where* $T_j = Z_j - \mu_j^*$*,* $j = 1, 2$*. Denoting the joint cumulative distribution function (cdf) of* T *under* μ^* *by* F_2*, one possibility is to search for the constant* c_α *such that* $F_2(c_\alpha, c_\alpha) = 1 - \alpha$*. The corresponding (exact) local significance level is then given by* $\alpha_{loc} = 1 - \Phi(c_\alpha)$*. If an unequal weighting (for importance) of the two null hypotheses is desired, one can search for a solution of the form* $F_2(c_\alpha^{(1)}, c_\alpha^{(2)}) = 1 - \alpha$ *for* $c_\alpha^{(1)} \neq c_\alpha^{(2)}$*.*

Section 2 will deal with straightforward generalizations of Example 1 to arbitrary dimensions $m \geq 2$, to more general types of null hypotheses, and to more general test statistics which are assumed to be (asymptotically) jointly normally distributed, at least under the global null hypothesis.

1.2 Preliminaries

Throughout the paper, we will assume a finite family $\mathcal{H}_m = (H_1, \ldots, H_m)$ of $m \in \mathbb{N}$ null hypotheses which are to be tested simultaneously under one and the same statistical model $(\mathcal{Y}, \mathcal{F}, (\mathbb{P}_\vartheta : \vartheta \in \Theta))$ with sample space $\mathcal{Y}$ and parameter ϑ taking values in the parameter space Θ. It is convenient to interpret each H_j as a subset of Θ, i. e., we will write $H_j \subset \Theta$, $1 \leq j \leq m$. The corresponding alternative hypotheses will be denoted by $K_j = \Theta \setminus H_j$, $1 \leq j \leq m$. Then we call $(\mathcal{Y}, \mathcal{F}, (\mathbb{P}_\vartheta : \vartheta \in \Theta), \mathcal{H}_m)$ a multiple test problem. A (non-randomized) multiple test $\varphi = (\varphi_1, \ldots, \varphi_m)^\top$ for $\mathcal{H}_m$ is a (measurable) mapping from the sample space $\mathcal{Y}$ to $\{0, 1\}^m$, where $\varphi_j(y) = 1$ means that we reject H_j in favor of K_j on the basis of the observed data y, for $1 \leq j \leq m$ and $y \in \mathcal{Y}$. For given $\vartheta \in \Theta$, let $I_0(\vartheta) \subseteq \{1, \ldots, m\}$ denote the index set of true null hypotheses under ϑ, meaning that exactly those H_j are true for which $j \in I_0(\vartheta)$. For type I error control of φ, we will consider the two random variables $V_m = \sum_{j \in I_0(\vartheta)} \varphi_j$ and $R_m = \sum_{j=1}^m \varphi_j$. The random variable V_m denotes the (random) number of type I errors of φ (under ϑ) and the random variable R_m denotes the total number of rejections of φ (under ϑ). Two important type I error measures, which we will consider in this paper, are the FWER and the false discovery rate (FDR). For a given value of $\vartheta \in \Theta$, they are given by

$$\mathrm{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(V_m > 0) = \mathbb{P}_\vartheta\left(\bigcup_{j \in I_0(\vartheta)} \{\varphi_j = 1\}\right), \quad (1)$$

$$\mathrm{FDR}_\vartheta(\varphi) = \mathbb{E}_\vartheta[\mathrm{FDP}_\vartheta(\varphi)] = \mathbb{E}_\vartheta\left[\frac{V_m}{\max(R_m, 1)}\right]. \quad (2)$$

The random variable $\mathrm{FDP}_\vartheta(\varphi) = V_m / \max(R_m, 1)$ appearing in (2) is called the false discovery proportion (FDP) of φ (under ϑ), where the max in the denominator is to avoid an expression of the form 0/0.

Let $H_0 = \bigcap_{j=1}^m H_j$ denote the global (null) hypothesis in $\mathcal{H}_m$, meaning that all m null hypotheses $H_1, \ldots, H_m$ are true for any $\vartheta \in H_0$. The multiple test φ is said to control the FWER at level $\alpha \in (0, 1)$ in the weak sense, if $\text{FWER}_\vartheta(\varphi) \leq \alpha$ for all $\vartheta \in H_0$. It is said to control the FWER at level α in the strong sense, if $\text{FWER}_\vartheta(\varphi) \leq \alpha$ for all $\vartheta \in \Theta$. Clearly, strong FWER control implies weak FWER control. The reverse implication (i. e., equivalence of weak FWER control and strong FWER control) holds true, if the statistical model, the system $\mathcal{H}_m$ of null hypotheses, and the multiple test φ are such, that the FWER of φ becomes largest for parameter values in H_0. In Sections 2 - 5 below, we will mainly study the FWER behavior of certain multiple tests under H_0, meaning that we design procedures for weak FWER control (in the first place). However, the application of the closed test principle allows one to utilize multiple tests with weak FWER control also for the purpose of strong FWER control. Namely, such tests have to be applied to every intersection hypothesis appearing in the closure of $\mathcal{H}_m$. We will not elaborate further on this strategy in this paper. It has been explained in detail, among many others, in [57], in [9], and in [53], where the latter article also contains applications to real data in a medical context. We note in passing that the concept of weak FDR control is not of independent interest, because FDR and FWER coincide if H_0 holds true. Hence, FDR control always has to be considered in the strict sense.

We assume, that every marginal test φ_j is carried out in terms of a real-valued test statistic T_j or a (random) p-value P_j taking values in $[0, 1]$, respectively, where $1 \leq j \leq m$. Our goal is to utilize the joint null distribution (or an approximation thereof) of the random vector $T = (T_1, \ldots, T_m)^\top$ or $P = (P_1, \ldots, P_m)^\top$, respectively, in the calibration of φ for type I error control.

The rest of the material in this paper is structured as follows. In Section 2, we will consider linear hypotheses testing in the presence of (asymptotically) jointly normally distributed test statistics $T_1, \ldots, T_m$. Here, a quantile of the full m-variate (asymptotic) distribution of T under H_0 is taken as the critical value for FWER control of the multiple test. Sections 3 and 4 deal with the case, that only k-th order marginal distributions of T under H_0 are available, for some $k < m$. This leads to the utilization of probability bounds / approximations (see Section 3), which can be expressed in terms of the "effective number of tests" (see Section 4). The methods in these two sections can be used under (asymptotic) joint normality of T under H_0, but they are not restricted to this distributional assumption. In Section 5, non-Gaussian dependencies are considered, which can conveniently be expressed in terms of so-called copula functions. Here, a point on the contour line of the copula of the test statistics (or their distributional transforms, respectively) determines the local significance level(s) for FWER control. In Section 6, we review some multivariate approaches to FDR control, mainly in terms of an adjustment of the nominal FDR level in the very popular Benjamini-Hochberg procedure. Finally, Section 7 provides some conclusions and practical recommendations.

2 Methods based on multivariate normal distributions

In this section, we will discuss simultaneous test procedures in the sense of [29] and [35].

Let ϑ with values in $\mathbb{R}^k$ for $k \in \mathbb{N}$ denote the statistical parameter of interest. In many applications, multiple test problems concerning ϑ can be formalized as systems of so-called linear hypotheses. To this end, let $C \in \mathbb{R}^{m \times k}$ denote a given matrix (the so-called contrast matrix) and $d \in \mathbb{R}^m$ a given vector, where $m \in \mathbb{N}$ denotes the number of null hypotheses to be tested simultaneously, as outlined in Section 1.2. Then, the system $\mathcal{H}_m = (H_1, \ldots, H_m)$ of (two-sided) linear hypotheses regarding ϑ, which is defined by C and d, can be written as

$$C\vartheta = d, \tag{3}$$

where we interpret (3) as a system of m null hypotheses. This means, that we define H_i by line i of the system of equations in (3), where $1 \leq i \leq m$. Each H_i encodes one linear restriction concerning (components of) ϑ. It is also possible to consider inequality relations in (3), leading to one-sided null hypotheses as in Example 1.

Example 2.

(a) Assume that we want to test, which of the components of ϑ are different from zero. We let $m = k$, $C = I_k$ (the identity matrix in $\mathbb{R}^{k \times k}$), and $d = 0 \in \mathbb{R}^k$. Then, line i of (3) encodes the i-th null hypothesis $H_i = \{\vartheta_i = 0\}$, for $1 \leq i \leq k$.

(b) Assume that we want to compare all components ϑ_i for $1 \leq i \leq k-1$ with the component ϑ_k. This has the interpretation, that the component ϑ_k corresponds to a "control group / treatment" against which all other groups / treatments shall be compared. We let $m = k-1$, $d = 0 \in \mathbb{R}^{k-1}$, and $C = C_{\text{Dunnett}} \in \mathbb{R}^{k-1 \times k}$. The contrast matrix C_{Dunnett} is Dunnett's contrast matrix with $k-1$ rows and k columns, where in each row j the j-th entry equals $+1$, the k-th entry equals -1 and all other entries are equal to zero.

Now, assume for the moment that an (at least asymptotically for large sample sizes) unbiased and normally distributed estimator $\hat{\vartheta}$ of ϑ is at hand. Then, it is near at hand to employ the vector $T = C\hat{\vartheta} - d$ of test statistics for testing $\mathcal{H}_m$ defined by (3). If $\hat{\vartheta}$ (approximately) follows a normal distribution with mean ϑ and covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$ (where Σ is functionally independent of ϑ), then (under standard regularity assumptions) T (approximately) follows a centered normal distribution with covariance matrix $C\Sigma C^\top \in \mathbb{R}^{m \times m}$ under the global hypothesis $H_0 = \bigcap_{i=1}^m H_i$ of $\mathcal{H}_m$. A simultaneous test procedure (STP) based on the vector T chooses a suitable quantile of the (approximate) distribution $\mathcal{N}_m(0, C\Sigma C^\top)$ of T under H_0 as the rejection threshold c_α for each individual test statistic T_i (which is the i-th component of T or its absolute value, respectively). The i-th null hypothesis H_i gets rejected at FWER level α iff T_i exceeds c_α.

Remark 1.

(a) Exemplary model classes under which asymptotically unbiased and normally distributed estimators are available (under certain regularity conditions) have been discussed, among others, in [35], [8], [34], [42], and Section 4.2 of [12]. They comprise, for example, analysis of variance models, multiple linear regression models, generalized linear models, survival models, and various time series models.

(b) Strong FWER control at (asymptotic) level α of the multiple contrast test defined by T and c_α can be established under the (asymptotic) "subset pivotality" condition (see Section 2.2.3 in [64]). Detailed calculations can be found, for example, in Section III of [16] and in Lemma 3.1 of [20].

(c) The multivariate normal distribution can be replaced by a suitable multivariate Student's t distribution, if there is uncertainty about the marginal variances of $T_1, \ldots, T_m$ under H_0 and Studentization techniques are applied; see Example 3 for an application in the context of the analysis of variance.

Example 3 (ANOVA1). *Under the balanced, homoscedastic one-factorial analysis of variance (ANOVA) model with $k \geq 3$ groups, two important multiple test problems are "all pairwise multiple comparisons" (MCA) and "all multiple comparisons with a control group" (MCC). Here, $\vartheta \in \mathbb{R}^k$ is the vector of the group-specific population means, which is estimated by the vector $\hat{\vartheta}$ of the group-specific sample means. The error variance is assumed to be unknown. In the context of multiple contrast tests, MCA leads to the so-called Tukey contrast matrix and the Tukey test, respectively (see [62]), while MCC leads to the so-called Dunnett contrast matrix and the Dunnett test, respectively (see [22] and [23]). These are classical multiple tests which have been treated, for instance, in [34]. The following source code in* `R` *demonstrates how the critical values for the Tukey test can be derived (i) based on the general theory of multiple contrast tests, and (ii) based on the built-in* `R` *routine* `qtukey()`*. One has to notice, that the* `R` *routine* `qtukey()` *actually computes quantiles of the closely related Studentized range distribution. To obtain these quantiles, the critical values of the Tukey test have to be multiplied by $\sqrt{2}$.*

```
library("mvtnorm");
library("multcomp");

###############################################
# ANOVA1 with k groups, "all pairs" contrasts #
###############################################
```

```
n <- c(11,11,11,11);      #group-specific sample sizes
k <- length(n);           #number of groups
C <- contrMat(n, type = "Tukey");
M <- diag(1/n);
combis <- combn(1:k, 2);
D <- diag(sqrt((n[combis[1, ]] * n[combis[2, ]]) /
          ( n[combis[1, ]] + n[combis[2, ]])));

#correlation matrix of the test statistics
R <- D %*% C %*% M %*% t(C) %*% D;
alpha <- 0.1;
my_df <- sum(n) - k;

my_Tukey_quantile <- qmvt(p=1-alpha, tail="both.tails",
                          df=my_df, corr=R)$quantile;
my_StudRangeQuantile <- my_Tukey_quantile*sqrt(2);
my_Tukey_quantile;
my_StudRangeQuantile; my_df; k; alpha;
# In agreement with Table 8 on page 408 of
# Hochberg and Tamhane (1987)!


R_Tukey_quantile <- qtukey(p=1-alpha, nmeans=k, df=my_df);
R_Tukey_quantile;
```

Table 1 tabulates some numerical values of quantiles of the Studentized range distribution. All values in Table 1 are in very good agreement with Table 8 on pages 408 - 409 in [34].

Remark 2.

(a) *Explicit formulas and* R *code for multiple contrast tests under many other model classes can be found in [8]. Applications in nonparametric multiple comparisons, together with* R *code, have been worked out in [37] and [38].*

(b) *While in the specific case of the Tukey test the utilization of the* R *routine* `qtukey()` *appears more convenient than the more involved code referring to the general multiple contrast test methodology, one has to keep in mind that the setup of the latter has a much broader scope. As a very simple example, consider the case of unequal group-specific sample sizes under Example 3 (i. e., an unbalanced design). In that case, the (scaled) Studentized range distribution is not the correct null distribution*

Table 1: Some quantiles of the Studentized range distribution. The FWER level is denoted by α, k denotes the number of groups, n denotes the sample size per group, and $\nu = k(n-1)$ denotes the resulting degrees of freedom. The column "Contrast" contains the solution based on the general methodology of multiple contrast tests, and the column "qtukey" the one obtained by the built-in R routine qtukey().

α	k	n	ν	Contrast	qtukey
0.05	3	10	27	3.506	3.506
0.05	3	20	57	3.403	3.403
0.05	3	50	147	3.348	3.348
0.05	5	10	45	4.018	4.018
0.05	5	20	95	3.932	3.933
0.05	5	50	245	3.887	3.887
0.05	10	10	90	4.592	4.588
0.05	10	20	190	4.528	4.528
0.05	10	50	490	4.493	4.495
0.1	3	10	27	3.030	3.030
0.1	3	20	57	2.962	2.962
0.1	3	50	147	2.925	2.925
0.1	5	10	45	3.591	3.590
0.1	5	20	95	3.531	3.531
0.1	5	50	245	3.498	3.499
0.1	10	10	90	4.215	4.212
0.1	10	20	190	4.170	4.168
0.1	10	50	490	4.141	4.145

for the maximum of the test statistics anymore. However, the code referring to the general multiple contrast test methodology still delivers the correct quantile, if the group-specific sample sizes at hand are entered in its first line.

(c) *One further important generalization of the presented methodology is to extend the scope of multiple contrast test to the case of flexible study designs with several stages; see, e. g., [39] and the references therein.*

3 Methods based on higher-order probability bounds

Let $m \in \mathbb{N}$ denote the number of null hypotheses to be tested simultaneously, and assume that real-valued test statistics $T_1, \ldots, T_m$ are at hand, which tend to larger values under alternatives. For calibrating a multivariate STP φ based on $T = (T_1, \ldots, T_m)^\top$ or for calculating corresponding multiplicity-adjusted p-values, respectively, we have to evaluate expressions of the following form:

$$F_m(x) = \mathbb{P}_0 \left(\bigcap_{j=1}^{m} \{T_j \leq x\} \right), \tag{4}$$

or equivalently

$$\bar{F}_m(x) = 1 - F_m(x) = \mathbb{P}_0 \left(\bigcup_{j=1}^{m} \{T_j > x\} \right), \; x \in \mathbb{R}, \tag{5}$$

where $\mathbb{P}_0$ denotes some probability measure under the global null hypothesis $H_0 = \bigcap_{j=1}^m H_j$. The quantities in (4) or (5), respectively, can often not be evaluated exactly. Reasons for this can be (i) lacking information about the full m-variate null distribution of T, or (ii) computational infeasibility. For example, the `R` package `mvtnorm` (computation of multivariate normal / Student's t probabilities and quantiles) which is based on [31] gives an error message whenever m exceeds 1000. Therefore, two basic ideas for approximating $F_m(x)$, which only require the computation of lower-dimensional marginal distributions, i. e., $F_k(x)$ for some $k < m$, are given by sum-type and product-type probability bounds / approximations.

Lemma 1.

a) *(Bonferroni inequalities, sum-type probability bounds (STPBs).)*
Let $A_1, \ldots, A_m$ be arbitrary events, and let $\mathbb{P}$ denote any probability measure. Then

$$\forall p \geq 1: \; \sum_{k=1}^{2p} (-1)^{k-1} S_k \leq \mathbb{P} \left(\bigcup_{j=1}^{m} A_j \right) \leq b_{2p-1} := \sum_{k=1}^{2p-1} (-1)^{k-1} S_k, \tag{6}$$

where

$$S_k = \sum_{1 \leq j_1 < j_2 < \ldots < j_k \leq m} \mathbb{P}(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_k}) \tag{7}$$

for $1 \leq k \leq m$, and $S_k = 0$ for $k > m$. See Section 4.7 of [10] and [30] for proofs, related results, and further references. A bivariate variant of the aforementioned upper Bonferroni bounds is due to [65] and is given by

$$\mathbb{P}\left(\bigcup_{j=1}^{m} A_j\right) \leq b_2 := \sum_{j=1}^{m} \mathbb{P}(A_j) - \sum_{j=1}^{m-1} \mathbb{P}(A_j \cap A_{j+1}). \tag{8}$$

For our purposes we have to consider the events $A_j = \{T_j > x\}$ and the probability measure $\mathbb{P} = \mathbb{P}_0$, so that the probability expression in (6) and on the left-hand side of (8) equals $\bar{F}_m(x)$.

b) *(Product-type probability bounds (PTPBs).)*
Define the events $O_j := \{T_j \leq x\} = A_j^c$ for $1 \leq j \leq m$. Due to chain factorization, it holds for any $1 \leq k \leq m-1$ that

$$F_m(x) = \mathbb{P}_0(O_1, \ldots, O_m) = \mathbb{P}_0(O_1, \ldots, O_k) \prod_{j=k+1}^{m} \mathbb{P}_0(O_j | O_{j-1}, \ldots, O_1).$$

Now assume that T is sub-Markovian of order $k \geq 2$ (SM_k) in the sense of Definition 2.2 in [20] under H_0. Then it holds for all $k \leq j \leq m$ that

$$\mathbb{P}_0(O_j | O_{j-1}, \ldots, O_1) \geq \mathbb{P}_0(O_j | O_{j-1}, \ldots, O_{j-k+1}) \tag{9}$$

and, consequently,

$$F_m(x) \geq \beta_k := \mathbb{P}_0(O_1, \ldots, O_k) \prod_{j=k+1}^{m} \mathbb{P}_0(O_j | O_{j-1}, \ldots, O_{j-k+1}). \tag{10}$$

Occasionally, we will write $b_\ell(x)$ or $\beta_k(x)$, respectively, instead of b_ℓ or β_k, respectively, to indicate the argument x at which the approximations are evaluated. Furthermore, we refer to ℓ and k, respectively, as the order of these (sum- or product-type) approximations.

Remark 3.

a) *We note that the complexity of computing the sums S_k in (7) is high, because $\binom{m}{k}$ k-dimensional marginal probabilities have to be evaluated. On the other hand, (6) always holds true, regardless of the dependency structure among $T_1, \ldots, T_m$. A computationally inexpensive alternative is the utilization of b_2 from (8). Under certain structural assumptions, sum-type bounds of higher order can be improved. For example, the derivations in [43, 44] are based on geometric or topological arguments.*

b) *In the general case the inequality relation in (9) is not fulfilled; cf., among others, [4] and [32]. However, β_k often yields a good approximation of $F_m(x)$ already for $k \in \{2, 3\}$; see, for example, Section 4 of [59] for numerical results pertaining to multivariate chi-square probabilities. In the remainder, we refer to β_k as the* *product-type probability approximation (PTPA) of order k to $F_m(x)$. The word "approximation" instead of "bound" indicates, that the inequality in* (10) *may fail if T is not SM_k under H_0.*

c) The vector T is called positive lower orthant dependent (PLOD) under H_0, if for all $t = (t_1, \ldots, t_m)^\top \in \mathbb{R}^m$, it holds that

$$\mathbb{P}_0(T_1 \leq t_1, \ldots, T_m \leq t_m) \geq \prod_{j=1}^{m} \mathbb{P}_0(T_j \leq t_j).$$

This entails, in particular, that

$$F_m(x) \geq \prod_{j=1}^{m} \mathbb{P}_0(T_j \leq x) =: \beta_1. \tag{11}$$

Calibrating a multivariate multiple test by means of (11) *leads to a so-called Šidák test, see [54]. If $T_1, \ldots, T_m$ are jointly independent under H_0, we obtain equality in* (11).

Based on the aforementioned considerations, the following multiplicity- and dependency-adjusted p-values have been proposed in [57].

Definition 1. *For a given order ℓ or k, respectively, the procedure MADAM from [57] transforms the observed values $t_1, \ldots, t_m$ of the test statistics $T_1, \ldots, T_m$ into one of the following multiplicity- and dependency-adjusted p-values:*

$$p_{\Sigma,j} = b_\ell(t_j), \tag{12}$$

$$p_{\Pi,j} = 1 - \beta_k(t_j), \tag{13}$$

for all $1 \leq j \leq m$. The subscript Σ in (12) *indicates, that a STPB is utilized, and the subscript Π in* (13) *indicates, that a PTPA is utilized.*

For (approximate) FWER control at level α, the p-values from (12) or (13), respectively, may simply be thresholded at α.

4 Effective numbers of tests

The STPBs b_ℓ for $\ell \geq 2$ as well as the PTPAs β_k for $k \geq 2$ utilize information about the dependency structure among $T_1, \ldots, T_m$ under H_0 by incorporating ℓ-variate or k-variate marginal distributions, respectively, of $T = (T_1, \ldots, T_m)^\top$ under H_0. The bounds b_1 and β_1 only utilize univariate marginal distributions. One question of practical interest is, how much gain in FWER exhaustion and, consequently, power can be achieved by the exploitation of higher-order marginal distributions. This may also aid in selecting the appropriate order ℓ or k, respectively. On the one hand, the order should be chosen as large as possible in order to exhaust α as tightly as possible. On the other hand, as mentioned in Remark 3, the computational complexity of computing the bounds / approximations increases with increasing order.

One way to quantify the aforementioned gain is to compute the effective number of tests (see [20] and the references therein) corresponding to b_ℓ or β_k, respectively. To this end, assume that FWER control at a given level α is targeted and that the α_{loc}-quantile of T_j under H_0 is chosen as the critical value for the marginal test φ_j, where $1 \leq j \leq m$. Due to multiplicity, it is immediately clear that $\alpha_{loc} \leq \alpha$. For b_1 or β_1, respectively, the value of α_{loc} can be calculated straightforwardly. Namely, we obtain the (first-order) Bonferroni correction $\alpha_{loc} \equiv \alpha_{loc}(b_1) = \alpha/m$ in the case of b_1 and the Šidák correction $\alpha_{loc} \equiv \alpha_{loc}(\beta_1) = 1 - (1-\alpha)^{1/m}$ in the case of β_1.

By equating b_ℓ or β_k, respectively, for ℓ or k larger than one with α (where $x = x_j$ is chosen as the α_{loc}-quantile of the distribution of T_j under H_0), we can (numerically) determine $\alpha_{loc}(b_\ell)$ or $\alpha_{loc}(\beta_k)$, respectively. The effective number of tests of order ℓ or k, respectively, which we will denote by $M_{\text{eff}}^{(\ell)}$ or $M_{\text{eff}}^{(k)}$, is now found by (numerically) solving

$$\begin{aligned} M_{\text{eff}}^{(\ell)} \alpha_{loc}(b_\ell) &= \alpha \quad \text{or} & (14)\\ (1 - \alpha_{loc}(\beta_k))^{M_{\text{eff}}^{(k)}} &= 1 - \alpha, & (15) \end{aligned}$$

respectively. This means, that we write $\alpha_{loc}(b_\ell)$ in the form of a univariate Bonferroni correction with m replaced by $M_{\text{eff}}^{(\ell)}$ or that we write $\alpha_{loc}(\beta_k)$ in the form of a Šidák correction with m replaced by $M_{\text{eff}}^{(k)}$, respectively. Now, if the effective number of tests is smaller than m, we interpret this as "effectively" having to correct for only $M_{\text{eff}}^{(\ell)}$ or $M_{\text{eff}}^{(k)}$ comparisons instead of m ones. This reduction / relaxation of the "effective" multiplicity correction is due to the fact that we exploit dependencies among the tests (or test statistics), such that not every marginal test "fully counts" in $M_{\text{eff}}^{(\ell)}$ or $M_{\text{eff}}^{(k)}$, because of certain similarities between them.

Example 4. *Let $m = 50$ and assume that the vector $T = (T_1, \ldots, T_{50})^\top$ follows under the global hypothesis H_0 a centered m-variate normal distribution. The correlation matrix of T is assumed to be an equi-correlation matrix, such that all non-diagonal elements of it are identical and equal to ρ, where ρ ranges from 0 to 0.9 in steps of 0.1. In Table 2, we display the local significance levels and the effective numbers of tests resulting from three different calibration methods for a target FWER level of $\alpha = 5\%$: (i) Exact calibration (up to numerical inaccuracies) by means of the full 50-variate null distribution of T. We have used the* R*-routine* `qmvnorm()` *for this purpose. (ii) Approximate calibration by means of the STPB b_2 from* (8)*. (ii) Approximate calibration by means of the PTPA β_3 from* (10)*. For the marginal variances of each T_j, $1 \leq j \leq m$, we have chosen the value $1/n$ for $n = 30$. This mimics the case of a multiple test for Gaussian means (in the case of a unit error variance), where the sample size for each marginal test problem equals $n = 30$.*

Table 2: Local significance levels and effective numbers of tests corresponding to `qmvnorm()`, b_2 from (8), and β_3 from (10), respectively, assuming jointly normally distributed test statistics under the global null hypothesis. Their covariance matrix equals $\Sigma = n^{-1}\left(\rho\mathbf{1} + (1-\rho)\,I\right)$, where $\mathbf{1}$ denotes the matrix with every entry equal to one and I is the identity matrix. The parameter ρ is the equi-correlation coefficient. The values of $\alpha_{loc}^{\text{qmvnorm}}$ and $M_{\text{eff}}^{\text{qmvnorm}}$ are based on the full m-variate joint distribution of the test statistics. The global FWER level was chosen as $\alpha = 0.05$, the number of hypotheses equals $m = 50$, and the marginal variances are all equal to $1/n$, for $n = 30$. The values of $M_{\text{eff}}^{\text{qmvnorm}}$ have been computed according to (14).

ρ	$\alpha_{loc}^{\text{qmvnorm}}$	$M_{\text{eff}}^{\text{qmvnorm}}$	$\alpha_{loc}^{b_2}$	$M_{\text{eff}}^{b_2}$	$\alpha_{loc}^{\beta_3}$	$M_{\text{eff}}^{\beta_3}$
0	0.001025	48.78	0.001001	49.95	0.001026	49.97
0.1	0.001070	46.73	0.001003	49.85	0.001027	49.92
0.2	0.001161	43.07	0.001007	49.65	0.001038	49.39
0.3	0.001321	37.85	0.001015	49.26	0.001067	48.05
0.4	0.001573	31.79	0.001030	48.54	0.001140	44.97
0.5	0.001991	25.11	0.001057	47.30	0.001288	39.80
0.6	0.002609	19.16	0.001106	45.21	0.001575	32.54
0.7	0.003756	13.31	0.001194	41.88	0.002137	23.98
0.8	0.005900	8.47	0.001371	36.47	0.003259	15.71
0.9	0.011086	4.51	0.001836	27.23	0.006695	7.64

The numerical values displayed in Table 2 are very much in line with the FWER behavior of the multiple test discussed in our simple motivating Example 1: The stronger the positive correlation among the test statistics, the larger α_{loc} and, consequently, the smaller M_{eff}. In the setting studied here, the PTPA β_3 delivers a conservative approximation, which is much closer to the exact value (both in terms of α_{loc} and in terms of M_{eff}) than the STPB b_2.

Remark 4. *The concept of effective numbers of tests is very popular in the context of genetic association studies; see [13], Chapter 9 in [12], Section 4.1 in [20], Section 5 in [59], and Section 5.1 in [14] for details and many references. In that context, m contingency tables (one per considered genetic locus) have to be analyzed simultaneously with respect to association of a (typically binary) phenotype and the genotype at the respective locus. In this, m can be a very large number of an order of magnitude of up to 10^5 or 10^6 in the case of a genome-wide association study. Typically, at least in the presence of large sample sizes, a chi-square test statistic T_j is computed for each contingency table j, $1 \leq j \leq m$. Due to the biological mechanism of inheritance (and due to technical aspects of the measurements), there exist pronounced dependencies among the T_j's, and the vector $T = (T_1, \ldots, T_m)^\top$ follows a multivariate (central) chi-square distribution under the global hypothesis of independence of the phenotype of interest and the genotype at all m loci under investigation. The computation of multivariate chi-square probabilities is rather involved (see [18], [59], and the references therein), and it seems that up to date explicit analytical formulas only exist for two-, three- and four-variate chi-square probabilities. Therefore, the utilization of PTPAs of order $2-4$ has been proposed in this context in [20] and [59]. Another multivariate approach to addressing the multiplicity problem in genetic association analyses is to combine multiple test procedures with (inherently multivariate) statistical learning methods like support vector machines; see [41] and [15] for details.*

5 Copula-based methods

Again, we assume that $m \in \mathbb{N}$ null hypotheses are to be tested simultaneously, and that real-valued test statistics $T_1, \ldots, T_m$ are at hand, each tending to larger values under the alternative. As discussed around (4), the joint cdf F_m under a fixed parameter value $\vartheta^* \in H_0$ is needed in order to calibrate a multivariate STP based on $T = (T_1, \ldots, T_m)^\top$ for FWER control (under ϑ^*). The idea in this section is, to decompose F_m into the marginal cdfs of the T_j's and the dependency structure among the T_j's. To this end, denote by G_j the marginal cdf of T_j under ϑ^*, for $1 \leq j \leq m$. Then, we have the following result.

Theorem 1 (Sklar's Theorem, see [55, 56].)**.** *There exists a function $C_T : [0,1]^m \to [0,1]$,* *called the copula (function) of T, such that for all $t = (t_1, \ldots, t_m)^\top \in \bar{\mathbb{R}}^m$, it holds*

$$F_m(t) = C_T(G_1(t_1), \ldots, G_m(t_m)). \tag{16}$$

If G_j is a continuous function for all $1 \leq j \leq m$, then the copula C_T is unique.

Equation (16) formalizes the decomposition of F_m into the marginal cdfs $G_1, \ldots, G_m$ and the dependency structure among $T_1, \ldots, T_m$, which is mathematically described by the copula (or: dependence function) C_T. The advantages of working with (16) are threefold: (i) The transformation with the marginal cdfs leads to a distributional standardization under H_0, meaning that by the principle of quantile / probability integral transformation, we have that $G_j(T_j)$ is uniformly distributed on $[0, 1]$ under ϑ^* in the case of a continuous G_j, $1 \leq j \leq m$. Of course, the same holds true for the corresponding (random) p-value $P_j = 1 - G_j(T_j)$. The random variable $G_j(T_j)$ is often referred to as the distributional transform of T_j, cf. [50]. If all G_j's are strictly increasing on their supports, then the copula of the distributional transforms coincides with C_T. (ii) By means of (16), we obtain a very high degree of modeling flexibility. Namely, the marginal models referring to $G_1, \ldots, G_m$ can be coupled with any copula C_T. For example, in the case of marginal tests for means, univariate normal cdfs $G_1, \ldots, G_m$ can be coupled with a non-Gaussian copula, to describe situations where univariate normality of marginal arithmetic means can (approximately) be assumed, but multivariate normality of the vector T (under ϑ^*) may be questionable. (iii) Modeling dependencies by means of copula functions has meanwhile become a standard tool in applied multivariate statistics and quantitative risk management; see, e. g., [45], [36], [33], [24], and Chapter 5 of [40]. There exists a rich and ever growing body of literature on appropriate copula models for many applications. By means of (16), these models are available for multiple testing.

The following result connects Sklar's Theorem with FWER control under $\vartheta^* \in H_0$.

Lemma 2 (Theorem 2 in [16], Lemma 3.5 in [46].)**.** *Under the assumptions of Theorem 1, assume that $G_1, \ldots, G_m$ are known and fixed, at least asymptotically for large sample sizes. Furthermore, assume that C_T does not depend on ϑ. Let critical values $c_j(\alpha)$ for $1 \leq j \leq m$ be given, such that the j-th null hypothesis H_j is rejected at FWER level α by the STP φ iff T_j exceeds $c_j(\alpha)$. Finally, let $\alpha_{loc}^{(j)} = 1 - G_j(c_j(\alpha))$ denote a local significance level for the j-th marginal test problem, $1 \leq j \leq m$. Then we have that*

$$FWER_{\vartheta^*, C_T}(\varphi) = 1 - C_T\left(1 - \alpha_{loc}^{(1)}, \ldots, 1 - \alpha_{loc}^{(m)}\right).$$

Lemma 2 shows, that the multiplicity-adjusted local significance levels $\alpha_{loc}^{(1)}, \ldots, \alpha_{loc}^{(m)}$ or, equivalently, the multiplicity-adjusted critical values $c_1(\alpha), \ldots, c_m(\alpha)$ can be determined by means of the contour line of the copula C_T at contour level $1 - \alpha$. In practice, it is convenient to carry out the resulting multiple test φ in terms of the (realized) p-values $p_j = 1 - G_j(t_j)$, where t_j denotes the observed value of T_j, $1 \leq j \leq m$. We reject H_j at

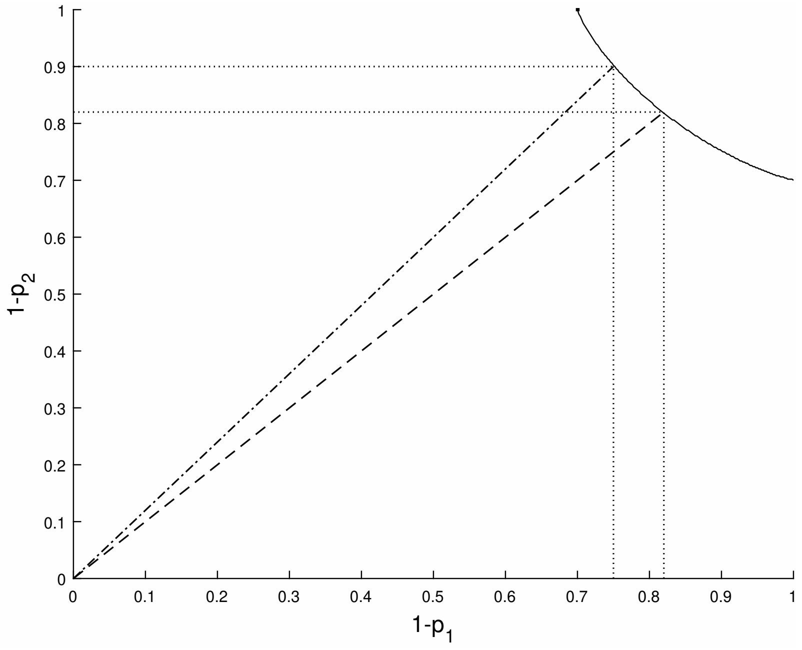

Figure 2: Copula-based calibration of a local significance level. The solid line displays the contour line of a specific bivariate copula at contour level 0.7, corresponding to an FWER level of $\alpha = 0.3$. The dashed line corresponds to an equal weighting of the two null hypotheses to which the p-values p_1 and p_2 refer. By projecting the point of intersection of the dashed line and the contour line of the copula onto the coordinate axes, we obtain an equi-coordinate local significance level of $1 - 0.82 = 0.18 > 0.15 = \alpha/2$. The dash-dotted line corresponds to a case in which the first null hypothesis gets a higher weight than the second one. We obtain a local significance level of $1 - 0.75 = 0.25$ for the first null hypothesis and a local significance level of $1 - 0.9 = 0.1$ for the second null hypothesis.

FWER level α, iff p_j is smaller than $\alpha_{loc}^{(j)}$. Figure 2 is an adapted and extended version of Figure 2 in [16] and depicts this construction in the bivariate case ($m = 2$). Figure 2 also shows how a possible weighting for importance of the null hypotheses is automatically incorporated in the methodology.

Remark 5. *Lemma 2 refers to a given, fixed copula C_T. In practice, the dependency structure among the test statistics $T_1, \ldots, T_m$ will however often not be known exactly. In such cases, C_T has to be (pre-)estimated. In [58], parametric estimation methods for copula functions and their impact on the FWER behavior of multivariate STPs have been studied. In particular, the authors provide a method to derive confidence regions for the realized FWER in the case that there is estimation uncertainty about C_T . In [46], analogous results have been obtained for certain nonparametric copula estimators, in particular so-called Bernstein copula estimators.*

Example 5. *Let us assume that the test statistics or their distributional transforms, re-*

Table 3: Comparison of the local significance levels $\alpha_{loc}^{\text{Sidak}} = 1 - (1-\alpha)^{1/m}$ and $\alpha_{loc}^{\text{Frank}(\eta)}$ calibrated under the Frank copula with parameter η. The global FWER level equals $\alpha = 0.05$ and the number of null hypotheses equals $m = 50$. The line $\eta = 0$ corresponds to joint independence.

η	$\alpha_{loc}^{\text{Sidak}}$	$\alpha_{loc}^{\text{Frank}(\eta)}$
0	0.00103	0.00101
2	0.00103	0.00107
4	0.00103	0.00113
6	0.00103	0.00116
8	0.00103	0.00122
10	0.00103	0.00129
12	0.00103	0.00135
14	0.00103	0.00142

spectively, follow under the global hypothesis H_0 an m-variate Frank copula with parameter η. Denoting this copula by C_η, it holds that

$$C_\eta(u_1, \ldots, u_m) = \psi_\eta \left(\sum_{j=1}^{m} \psi_\eta^{-1}(u_j) \right),$$

where $u_j \in [0,1]$ for all $1 \leq j \leq m$. The function ψ_η is called the generator function of C_η, and it is given by

$$\psi_\eta(t) = -\frac{1}{\eta} \log \left(1 - \frac{1 - \exp(-\eta)}{\exp(t)} \right), \ t \in [0,1].$$

For $\eta \to 0$, the model tends to the case of joint independence, while the degree of positive dependency increases with increasing $\eta > 0$.

Table 3 compares the equi-coordinate local significance levels resulting from Lemma 2 (applied to $C_T = C_\eta$ with varying values of η) with the local significance level of the Šidák test (which is exact under H_0 when assuming joint independence of the test statistics or p-values, respectively). In the case of $\eta = 0$ and $\eta = 2$, the two local significance levels are very similar. For larger values of η, however, the copula-based (multivariate) calibration method exploits the positive dependencies among the test statistics, and it leads to a markedly larger local significance level than the Šidák method. All values in Table 3 refer to $m = 50$ and an FWER level of $\alpha = 0.05$.

6 Multivariate multiple tests for control of the false discovery rate

Up to date, the still by far most popular multiple test procedure for control of the FDR is the linear step-up test φ^{LSU}, which had been proposed in the seminal paper [1] by Benjamini and Hochberg. The multiple test φ^{LSU} is also often referred to as the Benjamini-Hochberg procedure or simply *the* FDR procedure. In our notation, the decision rule of φ^{LSU} may be written as follows.

(i) Let $p_{1:m}, \ldots, p_{m:m}$ denote the ordered values of the m (random) p-values $P_1, \ldots, P_m$.

(ii) Determine
$$k = \max\{1 \leq j \leq m : p_{j:m} \leq j\alpha/m\}, \tag{17}$$
where α denotes the target FDR level.

(iii) If the maximum in (17) does not exist, retain all m null hypotheses $H_1, \ldots, H_m$. Otherwise, reject exactly $H_{1:m}, \ldots, H_{k:m}$, where $H_{1:m}, \ldots, H_{m:m}$ denote the re-ordered null hypotheses in $\mathcal{H}_m$, according to the ordering of the corresponding p-values.

Early results on FDR control of φ^{LSU} dealt with the case of jointly independent p-values (or test statistics); see [1] and [28]. However, since the beginning of the 21st century several authors have analyzed the FDR behavior of φ^{LSU} and related stepwise rejective multiple tests under (positive) dependency; cf., e. g., [2], [51], [25], [52], [5], [6], [7], [26], and references therein.

Let ϑ denote the parameter of the statistical model under consideration, denote by $\mathbb{P}_\vartheta$ the distribution of the data sample under ϑ, and let $P_1, \ldots, P_m$ denote the (random) p-values on which φ^{LSU} operates. In [2], Benjamini and Yekutieli proved the following result (see also Section 4 in [27] for slightly more general calculations).

Lemma 3 (see Equation (10) in [2].). *Assume that exactly $m_0 \equiv m_0(\vartheta)$ out of the m null hypotheses that are to be tested simultaneously are true under ϑ. Without loss of generality, assume that H_i is true for $1 \leq i \leq m_0$, and that H_i is false for $m_0 + 1 \leq i \leq m$. Then it holds for the FDR of φ^{LSU} under ϑ, that*
$$FDR_\vartheta\left(\varphi^{LSU}\right) = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \mathbb{P}_\vartheta(P_i \leq q_k \cap C_k^{(i)}), \tag{18}$$
where $q_j = j\alpha/m$, $1 \leq j \leq m$, and $C_k^{(i)}$ denotes the event that exactly $k-1$ null hypotheses additionally to H_i are rejected by φ^{LSU}.

Clearly, the probability expression on the right-hand side of (18) refers to the joint distribution of $P_1, \ldots, P_m$ under ϑ. Hence, it is possible to employ the multivariate approaches from our earlier sections to calculate / bound the FDR of φ^{LSU} under well-defined dependency structures among $P_1, \ldots, P_m$. This can for instance be used to adjust the nominal FDR level α. Namely, if the FDR of φ^{LSU} (or an upper bound for it) under the assumed dependency structure exceeds α, we may replace the nominal value of α by some smaller value such that the FDR is controlled. For example, Benjamini and Yekutieli showed in [2] that replacing α by $\alpha / \left(\sum_{i=1}^{m} i^{-1}\right)$ in the definition of φ^{LSU} always controls the FDR, no matter the dependency structure among $P_1, \ldots, P_m$. On the other hand, in [6] it has been shown that the FDR of φ^{LSU} is typically strictly smaller than α if the copula of $P_1, \ldots, P_m$ is an Archimedean copula with a completely monotone generator function (which does not depend on ϑ), and the authors derived an adjustment factor by which the nominal level α can be increased in order to exhaust the FDR level and, hence, optimize the power of φ^{LSU} by exploiting the (positive) dependencies among $P_1, \ldots, P_m$.

Both of these proposals follow the general construction method (exploitation of FDR bounds) which has been mentioned on page 81 of [12]. However, it has to be mentioned that the calculations referring to the right-hand side of (18) can become rather tedious, due to the complicated structure of the event $\{P_i \leq q_k \cap C_k^{(i)}\}$. Furthermore, as indicated for instance in [25] and in [3], the false discovery proportion (FDP) is typically not well concentrated around its expectation (the FDR) under dependency. Hence, many authors consider it more appropriate to control exceedance probabilities (over some given threshold) of the FDP under dependency rather than its mean; cf. also [11] and the references therein. The distribution of the FDP relies on the joint distribution of all m (random) p-values, too; cf. [3], [63], and the references therein.

7 Conclusions and practical recommendations

We have presented multivariate approaches to the calibration of multiple tests for control of the FWER and the FDR, respectively. Multivariate statistical models and resulting multiple tests have several advantages over generic procedures which only take into account univariate marginal distributions of test statistics or p-values: (i) Multivariate statistical models are often more realistic, because they take into account the dependencies in the data. Such dependencies are ubiquitous in nowadays' (high-throughput) measurements, because of the underlying (neuro-)biological or technological mechanisms. Hence, data from such experiments typically exhibit strong temporal, spatial, or spatio-temporal dependencies. (ii) Especially in the presence of positive dependencies, the power of the multiple test can be enhanced by explicitly modeling and incorporating marginal distributions of higher order into the decision rule of the multiple test. This has been demonstrated by numerical examples in Sections 4 and 5. Disadvantages of the presented methods are a potentially high computational effort, and the need for information about the kind of dependencies among the test statistics or p-values, leading to a higher model complexity than in the case of univariate marginal approaches. Therefore, it is recommendable in practice to apply multivariate techniques whenever computationally feasible, given that the type of dependency structure among the test statistics is known. In the case of a totally unknown dependency structure among $T_1, \ldots, T_m$, nonparametric copula (pre-)estimation methods may be applied, but the methodology of [46] at least requires that the copula of T is a nuisance parameter in the sense that it does not depend on the parameter ϑ to which the null hypotheses $H_1, \ldots, H_m$ refer.

One different class of multivariate multiple tests, which has not been treated in this paper, is constituted by resampling-based procedures. Such procedures implicitly take into account the aformentioned dependencies by employing appropriate resampling schemes which approximate the (joint) null distribution of the entire vector of test statistics or p-values, respectively. Resampling-based multiple tests for FWER control have been worked out for instance in [64], [60], [48], [49], and [9]. Resampling-based methods for FDR control have been derived, among others, in [66], [61], [47], and [21].

Of course, it is also possible to combine explicit multivariate modeling of the data and resampling-based calibration of a multiple test for control of the FWER or FDR, respectively. For example, a model-based bootstrap procedure for multiple specification tests in dynamic factor models has been theoretically derived in [17] and implemented and applied in [19].

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc., Ser. B, 57(1):289–300, 1995.

[2] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. Ann. Stat., 29(4):1165–1188, 2001.

[3] G. Blanchard, T. Dickhaus, E. Roquain, and F. Villers. On least favorable configurations for step-up-down tests. Statistica Sinica, 24(1):1–23, 2014.

[4] Henry W. Block, Tim Costigan, and Allan R. Sampson. Product-type probability bounds of higher order. Probab. Eng. Inf. Sci., 6(3):349–370, 1992.

[5] Henry W. Block, Thomas H. Savits, Jie Wang, and Sanat K. Sarkar. The multivariate-t distribution and the Simes inequality. Stat. Probab. Lett., 83(1):227–232, 2013.

[6] Taras Bodnar and Thorsten Dickhaus. False discovery rate control under Archimedean copula. Electron. J. Stat., 8(2):2207–2241, 2014.

[7] Taras Bodnar and Thorsten Dickhaus. On the Simes inequality in elliptical models. Ann. Inst. Statist. Math., 69(1):215–230, 2017.

[8] Frank Bretz, Torsten Hothorn, and Peter Westfall. Multiple Comparisons Using R. Chapman and Hall/CRC., 2010.

[9] EunYi Chung and Joseph P. Romano. Multivariate and multiple permutation tests. Journal of Econometrics, 193(1):76 – 91, 2016.

[10] Louis Comtet. Advanced combinatorics. The art of finite and infinite expansions. D. Reidel Publishing Co., Dordrecht, enlarged edition, 1974.

[11] Sylvain Delattre and Etienne Roquain. New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. Ann. Statist., 43(3):1141–1177, 2015.

[12] T. Dickhaus. Simultaneous Statistical Inference with Applications in the Life Sciences. Berlin, Heidelberg: Springer, 2014.

[13] T. Dickhaus, K. Strassburger, D. Schunk, C. Morcillo-Suarez, T. Illig, and A. Navarro. How to analyze many contingency tables simultaneously in genetic association studies. Stat. Appl. Genet. Mol. Biol., 11(4):Article 12, 2012.

[14] Thorsten Dickhaus. Simultaneous Bayesian analysis of contingency tables in genetic association studies. Stat. Appl. Genet. Mol. Biol., 14(4):347–360, 2015.

[15] Thorsten Dickhaus. Combining high-dimensional classification and multiple hypotheses testing for the analysis of big data in genetics. In Statistics and its applications, volume 244 of Springer Proc. Math. Stat., pages 47–50. Springer, Singapore, 2018.

[16] Thorsten Dickhaus and Jakob Gierl. Simultaneous test procedures in terms of p-value copulae. In Proceedings on the 2nd Annual International Conference on Computational Mathematics, Computational Geometry & Statistics (CMCGS 2013), pages 75–80. Global Science and Technology Forum (GSTF), 2013.

[17] Thorsten Dickhaus and Markus Pauly. Simultaneous statistical inference in dynamic factor models. In Ignacio Rojas and Héctor Pomares, editors, Time Series Analysis and Forecasting, pages 27–45. Springer, 2016.

[18] Thorsten Dickhaus and Thomas Royen. A survey on multivariate chi-square distributions and their applications in testing multiple hypotheses. Statistics, 49(2):427–454, 2015.

[19] Thorsten Dickhaus and Natalia Sirotko-Sibirskaya. Simultaneous statistical inference in dynamic factor models: chi-square approximation and model-based bootstrap. Comput. Statist. Data Anal., 129:30–46, 2019.

[20] Thorsten Dickhaus and Jens Stange. Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. Calcutta Statistical Association Bulletin, 65(257-260):123–144, 2013.

[21] Sandrine Dudoit and Mark J. van der Laan. Multiple testing procedures with applications to genomics. Springer Series in Statistics. Springer, New York, 2008.

[22] Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc., 50:1096–1121, 1955.

[23] Charles W. Dunnett. New tables for multiple comparisons with a control. Biometrics, 20:482–491, 1964.

[24] Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas and applications to risk management. In S.T. Rachev, editor, Handbook of Heavy Tailed Distributions in Finance, pages 329–384. Elsevier Science B.V., 2003.

[25] H. Finner, T. Dickhaus, and M. Roters. Dependency and false discovery rate: Asymptotics. Ann. Stat., 35(4):1432–1455, 2007.

[26] H. Finner, M. Roters, and K. Strassburger. On the Simes test under dependence. Statist. Papers, 58(3):775–789, 2017.

[27] Helmut Finner, Thorsten Dickhaus, and Markus Roters. On the false discovery rate and an asymptotically optimal rejection curve. Ann. Stat., 37(2):596–618, 2009.

[28] Helmut Finner and M. Roters. On the false discovery rate and expected type I errors. Biom. J., 43(8):985–1005, 2001.

[29] K. R. Gabriel. Simultaneous test procedures - some theory of multiple comparisons. Ann. Math. Stat., 40:224–250, 1969.

[30] Janos Galambos and Italo Simonelli. Bonferroni-type inequalities with applications. Probability and its Applications (New York). Springer-Verlag, New York, 1996.

[31] Alan Genz and Frank Bretz. Computation of multivariate normal and t probabilities. Lecture Notes in Statistics 195. Berlin: Springer, 2009.

[32] Joseph Glaz and Bruce McK. Johnson. Probability inequalities for multivariate distributions with dependence structures. J. Am. Stat. Assoc., 79:436–440, 1984.

[33] Wolfgang Karl Härdle and Ostap Okhrin. De copulis non est disputandum - Copulae: an overview. AStA Adv. Stat. Anal., 94(1):1–31, 2010.

[34] Yosef Hochberg and Ajit C. Tamhane. Multiple comparison procedures. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York etc.: John Wiley & Sons, Inc., 1987.

[35] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. Biom. J., 50(3):346–363, Jun 2008.

[36] Harry Joe. Dependence modeling with copulas. Boca Raton, FL: CRC Press, 2014.

[37] Frank Konietschke, Ludwig A. Hothorn, and Edgar Brunner. Rank-based multiple test procedures and simultaneous confidence intervals. Electron. J. Stat., 6:738–759, 2012.

[38] Frank Konietschke, Marius Placzek, Frank Schaarschmidt, and Ludwig Hothorn. nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. Journal of Statistical Software, 64(9):1–17, 2015.

[39] D. Magirr, T. Jaki, and J. Whitehead. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika, 99(2):494–501, 2012.

[40] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. Quantitative risk management. Concepts, techniques, and tools. Princeton, NJ: Princeton University Press, 2005.

[41] B. Mieth, M. Kloft, J. A. Rodriguez, S. Sonnenburg, R. Vobruba, C. Morcillo-Suarez, X. Farre, U. M. Marigorta, E. Fehr, T. Dickhaus, G. Blanchard, D. Schunk, A. Navarro, and K. R. Müller. Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. Scientific Reports, 6:Article 36671, Nov 2016.

[42] Rupert G. Miller, Jr. Simultaneous statistical inference. Springer-Verlag, New York-Berlin, second edition, 1981. Springer Series in Statistics.

[43] Daniel Q. Naiman and Henry P. Wynn. Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. Ann. Stat., 20(1):43–76, 1992.

[44] D.Q. Naiman and H.P. Wynn. The algebra of Bonferroni bounds: discrete tubes and extensions. Metrika, 62(2-3):139–147, 2005.

[45] Roger B. Nelsen. An introduction to copulas. 2nd ed. Springer Series in Statistics. New York, NY: Springer., 2006.

[46] André Neumann, Taras Bodnar, Dietmar Pfeifer, and Thorsten Dickhaus. Multivariate multiple test procedures based on nonparametric copula estimation. Biom. J., 61(1):40–61, 2019.

[47] Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. Control of the false discovery rate under dependence using the bootstrap and subsampling. Test, 17(3):417–442, 2008.

[48] Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. J. Am. Stat. Assoc., 100(469):94–108, 2005.

[49] Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. Econometrica, 73(4):1237–1282, 2005.

[50] Ludger Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. J. Stat. Plann. Inference, 139(11):3921–3927, 2009.

[51] Sanat K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. Ann. Stat., 30(1):239–257, 2002.

[52] Sanat K. Sarkar. On the Simes inequality and its generalization. IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen, 1:231–242, 2008.

[53] K. Schildknecht, S. Olek, and T. Dickhaus. Simultaneous Statistical Inference for Epigenetic Data. PLOS ONE, 10(5):Article e0125587, 2015.

[54] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc., 62:626–633, 1967.

[55] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, 8:229–231, 1959.

[56] A. Sklar. Random variables, distribution functions, and copulas - a personal look backward and forward. In Distributions with Fixed Marginals and Related Topics., pages 1–14. Institute of Mathematical Statistics, Hayward, CA, 1996.

[57] J. Stange, T. Dickhaus, A. Navarro, and D. Schunk. Multiplicity- and dependency-adjusted p-values for control of the family-wise error rate. Stat. Probab. Lett., 111:32–40, 2016.

[58] Jens Stange, Taras Bodnar, and Thorsten Dickhaus. Uncertainty quantification for the family-wise error rate in multivariate copula models. AStA Adv. Stat. Anal., 99(3):281–310, 2015.

[59] Jens Stange, Nina Loginova, and Thorsten Dickhaus. Computing and approximating multivariate chi-square probabilities. Journal of Statistical Computation and Simulation, 86(6):1233–1247, 2016.

[60] James F. Troendle. A stepwise resampling method of multiple hypothesis testing. J. Am. Stat. Assoc., 90(429):370–378, 1995.

[61] James F. Troendle. Stepwise normal theory multiple test procedures controlling the false discovery rate. J. Stat. Plann. Inference, 84(1-2):139–158, 2000.

[62] J. W. Tukey. The problem of multiple comparisons. In The collected works of John W. Tukey. Volume VIII: Multiple comparisons: 1948-1983. Ed. by Henry I. Braun, with the assistance of Bruce Kaplan, Kathleen M. Sheehan, Min-Hwei Wang., pages 1–300. New York, NY: Chapman & Hall, 1953.

[63] Jonathan von Schroeder and Thorsten Dickhaus. Efficient Calculation of the Joint Distribution of Order Statistics. Preprint, available at https://arxiv.org/abs/1812.09063.

[64] Peter H. Westfall and S. Stanley Young. Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, New York, 1993.

[65] K.J. Worsley. An improved Bonferroni inequality and applications. Biometrika, 69:297–302, 1982.

[66] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plann. Inference, 82(1-2):171–196, 1999.

[67] G. Yona, W. Dirks, S. Rahman, and D. M. Lin. Effective similarity measures for expression profiles. Bioinformatics, 22(13):1616–1622, Jul 2006.