Mathematical Statistics
Stockholm University

Estimating the Proportion of True Null Hypotheses under Copula Dependency

André Neumann
Taras Bodnar
Thorsten Dickhaus

Research Report 2017:09

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 **2017:09**,
http://www.math.su.se

Estimating the Proportion of True Null Hypotheses under Copula Dependency

June 2017

André Neumann[1], Taras Bodnar[2], and Thorsten Dickhaus[3]

It is a well known result in multiple hypothesis testing that the proportion π_0 of true null hypotheses is not identified under general dependencies. However, it is possible to estimate π_0 if structural information about the dependency structure among the test statistics or p-values, respectively, is available. We demonstrate these points, and propose a marginal parametric bootstrap method. A pseudo-sample of bootstrap p-values is generated, which still carry information about π_0, but behave like realizations of stochastically independent random variables. Theoretical properties of resulting estimation procedures for π_0 are analyzed and their usage is illustrated on synthetic and real data.

MSC 2010 classification numbers: Primary 62J15; secondary 62F03.
Key words: Bootstrap, copula, family-wise error rate, p-values, Schweder-Spjøtvoll estimator.

[1]André Neumann is Ph. D. candidate (E-mail: neumann@uni-bremen.de) at University of Bremen, Institute for Statistics, P. O. Box 330 440, D-28344 Bremen, Germany. His research was supported by the Deutsche Forschungsgemeinschaft through the Research Unit FOR 1735 "Structural Inference in Statistics: Adaptation and Efficiency".

[2]Taras Bodnar is associate professor (E-mail: taras.bodnar@math.su.se) at Stockholm University, Department of Mathematics, Roslagsvgen 101, SE-10691 Stockholm, Sweden.

[3]Thorsten Dickhaus is professor (E-mail: dickhaus@uni-bremen.de) at University of Bremen, Institute for Statistics, P. O. Box 330 440, D-28344 Bremen, Germany.

1. INTRODUCTION AND MOTIVATION

Under the multiple testing framework, estimating the proportion π_0 of true null hypotheses is informative for various reasons. On the one hand, in applications like quality control or anomaly detection, the presence of a certain number of untypical data points already indicates the necessity for an intervention, no matter which of the data points are responsible for that. On the other hand, data-adaptive multiple test procedures (Dickhaus 2014, sec. 3.1.3) incorporate an estimate $\hat{\pi}_0$ into their decision rules in order to optimize power (Langaas et al. 2005; Finner and Gontscharuk 2009; Celisse and Robin 2010; Dickhaus et al. 2012).

Throughout the remainder, we assume that m null hypotheses, which relate to the (main) parameter $\boldsymbol{\vartheta}$ of one and the same statistical model, are simultaneously under consideration. We let $m_0 = m_0(\boldsymbol{\vartheta})$ denote the number of true nulls, hence $\pi_0 = m_0/m$. The number of false null hypotheses is denoted by $m_1 = m_1(\boldsymbol{\vartheta}) = m - m_0$. Furthermore, we assume that test statistics $T_1, \ldots, T_m$ and corresponding p-values $P_1, \ldots, P_m$ are at hand. Without loss of generality, we will assume throughout that the p-values $P_1, \ldots, P_{m_0}$ correspond to true null hypotheses, while $P_{m_0+1}, \ldots, P_m$ correspond to false null hypotheses. Under independence assumptions regarding the joint distribution of the p-values, the very popular Schweder-Spjøtvoll estimator $\hat{\pi}_0^{\text{Schweder}}$ for π_0 has originally been proposed by Schweder and Spjøtvoll (1982). Theoretical properties of $\hat{\pi}_0^{\text{Schweder}}$ and slightly modified versions of it have been investigated by Storey et al. (2004), Langaas et al. (2005), Finner and Gontscharuk (2009), Dickhaus et al. (2012), Dickhaus (2013), and Cheng et al. (2015). Based on the EM algorithm, a novel estimation procedure for π_0 has recently been proposed by Oyeniran (2016), also under independence assumptions. Competing estimators have been compared by Hwang et al. (2014) and Nguyen and Matias (2014).

To our knowledge, the case of dependent test statistics or p-values, respectively, has not been treated yet in depth in the literature. Under the assumption of a linear factor model, Friguet and Causeur (2011) proposed an adjustment procedure prior to the application of $\hat{\pi}_0^{\text{Schweder}}$. Under monotonicity and convexity constraints regarding the mixture density of the p-values, Ostrovnaya and Nicolae (2012) worked out a (maximum likelihood) estimator based on a multinomial model. However, in many applications in modern life sciences, where the involved technical and biological mechanisms of data generation typically lead to involved temporal, spatial, or spatio-temporal dependencies (Stange et al. 2016), it is hard to verify such explicit

model assumptions. Therefore, we express dependency structures in this work in the most general manner by means of copula functions (Sklar 1996). Unfortunately, as we will demonstrate in Example 1 below, π_0 is not identified under general dependencies. This seems to be a well known fact in multiple test theory. Meinshausen and Bühlmann (2005) established an upper bound for π_0 based on a bounding function approach. However, the choice of an appropriate bounding function is only straightforward in the case of a multi-sample problem. Wang et al. (2011) employed a sliding linear model (SLIM) approach which is based on the empirical cumulative distribution function (ecdf) of all m marginal p-values.

The estimator $\hat{\pi}_0^{\text{Schweder}}$ also relies on the ecdf of $P_1, \ldots, P_m$ and on a tuning parameter $\lambda \in (0, 1)$, where the typical default value is $\lambda = 1/2$. The tuning parameter is chosen such that all p-values under alternatives are presumably smaller than λ. Denoting the ecdf of $P_1, \ldots, P_m$ by $\hat{F}_m$, $\hat{\pi}_0^{\text{Schweder}}$ is given by

$$\hat{\pi}_0^{\text{Schweder}} \equiv \hat{\pi}_0^{\text{Schweder}}(\lambda) = \frac{1 - \hat{F}_m(\lambda)}{1 - \lambda}.$$

There exist several heuristic motivations for the usage of $\hat{\pi}_0$. The simplest one considers a histogram of the marginal p-values with exactly two bins, namely $[0, \lambda]$ and $(\lambda, 1]$. Then, the height of the bin associated with $(\lambda, 1]$ equals $\hat{\pi}_0(\lambda)$ (Dickhaus 2014, fig. 3.2(a)). A graphical algorithm for computing $\hat{\pi}_0$ connects the point $(\lambda, \hat{F}_m(\lambda))$ with the point $(1, 1)$. The offset of the resulting straight line at $t = 0$ equals $\hat{\pi}_1 = \hat{\pi}_1(\lambda) = 1 - \hat{\pi}_0(\lambda)$ (Dickhaus 2014, fig. 3.2(b)). Both of these heuristic motivations implicitly assume that the ecdf of the p-values corresponding to true null hypotheses is close to the main diagonal in the unit square. However, under dependency this assumption is prone to be violated, because the p-values have the tendency to cluster. A drastic example of this behavior can be used to demonstrate that it is impossible to estimate π_0 based on $\hat{F}_m$ under arbitrary dependencies, even if the sample size tends to infinity.

Example 1. Assume that the copula of $\boldsymbol{P} = (P_1, \ldots, P_m)^\top$ is a Gumbel-Hougaard copula with copula parameter $\eta \geq 1$; see Stange et al. (2015) for justifications of this type of copula in the context of multiple tests related to extreme value theory. The value $\eta = 1$ corresponds to joint independence of all m p-values, while the strength of dependency among $P_1, \ldots, P_m$ increases with $\eta > 1$. Furthermore, assume that the p-values corresponding to true null hypotheses are marginally uniformly distributed on $[0, 1]$, while each P_j, $j > m_0$, is marginally uniformly distributed on $[0, \gamma_j]$ for uniformly selected values $\gamma_j < 1$.

In Figure 1, we present two computer simulations for $m = 100$, $m_0 = 50$, $\eta = 100$, and $\lambda = 1/2$. In both graphs displayed in Figure 1, the clustering of the p-values which is due to the large value of η can clearly be observed. The ecdf of $P_1, \ldots, P_m$ exhibits a large step at the realized value of the first p-value P_1, because all m_0 "true p-values" are almost totally dependent so that they take with very high probability essentially all the same value.

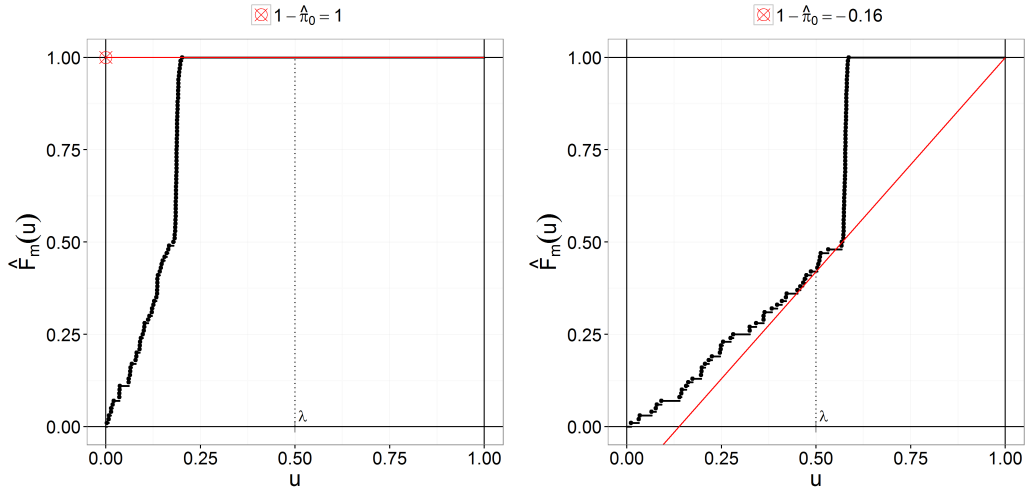

Figure 1: Computer simulations of the behavior of $\hat{\pi}_0^{\text{Schweder}}$ under a Gumbel-Hougaard copula with copula parameter $\eta = 100$. In the left graph, the p-values corresponding to true null hypotheses cluster around a value smaller than λ, while in the right graph they cluster around a value larger than λ.

Under this model, the behavior of $\hat{\pi}_0^{\text{Schweder}}$ (indicated by the straight lines) can be characterized as follows. If P_1 takes a value smaller than λ (as in the left graph), the main step of $\hat{F}_m$ is at a value smaller than λ, hence the estimated proportion of false hypotheses equals 1, meaning that we estimate m_0 to be equal to zero. On the other hand, if P_1 takes a value larger than λ (as in the right graph), the main step of $\hat{F}_m$ is at a value larger than λ, hence the estimated proportion of false hypotheses is less than or equal 0, meaning that we estimate m_0 to be larger than or equal to m. In practice, one may truncate the estimator at $m_0 = m$. In summary, the truncated Schweder-Spjøtvoll estimator for π_0 follows under very strong dependency a two-point distribution with two point masses in zero and one. It may be true that the point mass in one is large enough to make the (truncated) estimator mean conservative (i. e., upwardly biased), but its usage is inappropriate in practice. In particular, it is not consistent if $\pi_0 \in (0, 1)$. Finally, notice that the behavior of $\hat{\pi}_0^{\text{Schweder}}$ would remain exactly the same for a different value of m_0. Whether $\hat{\pi}_0^{\text{Schweder}}$

takes the value zero or the value one only depends on the realization of P_1, and this value is independent of the true value of m_0. In this sense, π_0 is not identified.

Example 1 demonstrates that some structural information about the dependency structure among the test statistics or p-values, respectively, it inevitable for the estimation of π_0. In this work, we assume that the dependency structure among $P_1, \ldots, P_m$ can be separated from the information that $P_1, \ldots, P_m$ carry about $\boldsymbol{\vartheta}$. Based on this structural assumption, we develop a marginal parametric bootstrap method for the estimation of π_0. We transform a bootstrap sample of the data into p-values $P_1^*, \ldots, P_m^*$, which approximately behave like realizations of jointly stochastically independent random variables. These p-values can then be used in $\hat{\pi}_0^{\text{Schweder}}$ instead of the original p-values. Applying this methodology to the situation considered in Example 1 leads to an accurate estimate of π_0, see Example 3 below. In contrast, the other approaches from the literature mentioned before are not suitable in this context. Namely, the model assumptions of Friguet and Causeur (2011) or Ostrovnaya and Nicolae (2012), respectively, are not fulfilled here. Application of the bounding function approach by Meinshausen and Bühlmann (2005) is difficult, because the p-values originated from one-sample problems. When applying the SLIM approach by Wang et al. (2011) with the recommended number of ten segments (i. e., subintervals of $[0, 1]$), we essentially encountered the same problems as for $\hat{\pi}_0^{\text{Schweder}}$, because their approach also relies on the ecdf $\hat{F}_m$. In every of the ten segments, we either obtained an estimated value for π_0 which exceeded one or which was equal to zero. It is to be expected that any ecdf-based estimator will suffer from the clustering effect of the p-values under null hypotheses.

The rest of the paper is structured as follows. In Section 2, we introduce the proposed bootstrap procedure. Theoretical properties of this procedure are analyzed in Section 3. A real data example from cancer research is presented in Section 4. Implications for multiple testing are discussed in Section 5, and we conclude with a discussion in Section 6.

2. ESTIMATION OF π_0 VIA MARGINAL PARAMETRIC BOOTSTRAP

For concreteness, we consider multiple test problems of the form $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (P_{\boldsymbol{\vartheta},C}^{\otimes n} : \boldsymbol{\vartheta} \in \Theta, C \in \mathcal{C}), \mathcal{H})$ for an independent and identically distributed (i.i.d.) sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_1 \sim \mathbf{X}$ and $\mathbf{X} = (X_1, \ldots, X_m)^\top$ is a random vector taking values in $\mathcal{X} \subseteq \mathbb{R}^m$ with an unknown absolutely continuous distribution $P \in \{P_{\boldsymbol{\vartheta},C} : \boldsymbol{\vartheta} \in \Theta, C \in \mathcal{C}\}$, and $\mathcal{F}$ is a σ-field over $\mathcal{X}$. The parameter $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_m)^\top$

is the main parameter of the model to which the family $\mathcal{H}$ of hypotheses $H_1, \ldots, H_m$ with corresponding alternatives $K_1, \ldots, K_m$ relates, and its parameter space Θ is a subset of $\mathbb{R}^m$. We assume that each ϑ_j is a parameter of the marginal distribution of X_j, for $1 \leq j \leq m$. The symbol C denotes the copula of $\mathbf{X}$, expressing the dependency structure among $(X_1, \ldots, X_m)^\top$. We mainly consider a semi-parametric setup in which $\mathcal{C}$ is a function space, although parametric copulae are utilized in some examples below. The hypotheses are interpreted as non-empty subsets of the parameter space Θ. We assume that hypotheses have the structure $H_j = \{\boldsymbol{\vartheta} \in \Theta \,|\vartheta_j = \theta_j\}$, where $(\theta_1, \ldots, \theta_m)^\top$ is a fixed element of Θ. This type of null hypotheses typically leads to uniformly distributed p-values under the null, while the latter is not fulfilled for general composite hypotheses (Dickhaus 2013).

Our proposed bootstrap method for estimating π_0 under arbitrary copula C of $\mathbf{X}$ is formalized in Algorithm 1.

Algorithm 1. *Let $\hat{\boldsymbol{\vartheta}}_n$ be a consistent estimator of $\boldsymbol{\vartheta}$, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the observed data sample and $k(n) \in \mathbb{N}$ the size of the bootstrap pseudo samples, where $k(n)$ is usually equal to n. We assume that for testing H_j a real-valued test statistic $T_j = T_{j,n}$ is at hand which tends to larger values under the alternative K_j, $1 \leq j \leq m$.*

1. *For every $1 \leq b \leq B$, $1 \leq j \leq m$ and $1 \leq i \leq k(n)$*

   (a) *sample $U_{i,j}^{*(b)}$ as standard uniformly distributed and independent random variable in i, j and b.*

   (b) *calculate $X_{i,j}^* := X_{i,j}^{*(b)} := F^{\leftarrow}_{X_j|\hat{\boldsymbol{\vartheta}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)}\left(U_{i,j}^{*(b)}\right)$, where $F^{\leftarrow}(u) := \inf\{x \in \mathbb{R} \,|F(x) \geq u\}$ denotes the generalized inverse of a cdf F.*

2. *For every $1 \leq b \leq B$ and $1 \leq j \leq m$*

   (a) *calculate $T_j^* := T_{j,n}^{*(b)} := T_{j,k(n)}\left(X_{1,j}^{*(b)}, \ldots, X_{k(n),j}^{*(b)}\right)$.*

   (b) *calculate $P_j^* := P_{j,n}^{*(b)} := 1 - F_{T_{j,k(n)}|\theta_j}\left(T_{j,n}^{*(b)}\right)$.*

3. *For every $1 \leq b \leq B$ calculate the Schweder-Spjøtvoll estimator*

$$\hat{\pi}_{0,n}^{*(b)}(\lambda) = \frac{1 - \hat{F}_m^{*(b)}(\lambda)}{1 - \lambda},$$

   where $\hat{F}_m^{*(b)}$ is the empirical distribution function of $P_{1,n}^{*(b)}, \ldots, P_{m,n}^{*(b)}$.

4. *Take the average $\overline{\hat{\pi}}_0^* := \overline{\hat{\pi}}_{0,n,B}^* := \frac{1}{B}\sum_{b=1}^{B} \hat{\pi}_{0,n}^{*(b)}$.*

In the first step, we generate for every $1 \leq j \leq m$ independently a bootstrap pseudo sample with the same marginal cdf as X_j under the estimated value of $\boldsymbol{\vartheta}$. Then, in step two, we calculate the test statistics and p-values based on these pseudo samples instead of the original data. In steps 3 and 4, we finally compute the Schweder-Spjøtvoll estimator. We show in Lemma 1 in Section 3 that the resulting bootstrap p-values are indeed conditionally independent given the data. Therefore, in contrast to Figure 1 we can expect the Schweder-Spjøtvoll estimator, applied to the bootstrapped p-values, to behave as in the case of joint independence of $X_1, \ldots, X_m$.

The following assumptions regarding the test statistics $T_1, \ldots, T_m$ are made throughout the remainder.

Assumption 1.

(a) The marginal parametric bootstrap works for the chosen test statistics, i. e., under true null hypotheses the differences between the marginal cdfs of the test statistics and the marginal cdfs of the bootstrap test statistics converge to zero uniformly, in probability.

(b) The marginal cdf of T_j only depends on ϑ_j and is continuous under true null hypotheses, for all $1 \leq j \leq m$.

Assumption (a) refers to the validity of the parametric bootstrap in a generic manner. Parametric bootstrap procedures have been considered in many fields, for example in gene expression analysis (Van Der Laan and Bryan 2001), in the analysis of variance (ANOVA) (Krishnamoorthy et al. 2007), for goodness-of-fit statistics (Cramr-von Mises, Kolmogorov-Smirnov) (Genest and Rémillard 2008), and for Wald statistics in dynamic factor models (Dickhaus and Pauly 2016). Assumption (b) formalizes the separation of the dependency structure in the data and their information about $\boldsymbol{\vartheta}$ which we have mentioned in the introduction. Continuity of the marginal cdfs is necessary for uniformly distributed p-values under true hypotheses. This property is essential for a reasonable behavior of the Schweder-Spjøtvoll estimator.

Remark 1. The marginal bootstrap is not suitable to approximate the null distribution of statistics like $\max_{1 \leq j \leq m} T_j$, which depend on the joint distribution of $\mathbf{X}$. For example, the maxT procedure of Westfall and Young (1993) for testing the global hypothesis $H_0 = \bigcap_{j=1}^{m} H_j$ uses the maximum of the test statistics to define adjusted p-values.

Before we analyze the theoretical properties of the proposed estimator $\overline{\hat{\pi}}_0^*$ in Section 3, the following examples illustrate the usage of Algorithm 1.

Example 2. Assume that $\mathbf{X}$ is normally distributed in $\mathbb{R}^m$ with expectation $\boldsymbol{\vartheta}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$, where the diagonal elements $\sigma_1^2, \ldots, \sigma_m^2$ are known, but the covariances are arbitrary. The null hypotheses $H_1, \ldots, H_m$ are given by $H_j := \{\boldsymbol{\vartheta} \in \mathbb{R}^m \,|\vartheta_j = \theta_j\}$, where $\boldsymbol{\theta} \in \mathbb{R}^m$. A consistent estimator $\hat{\vartheta}_j$ of the unknown parameter ϑ_j is given by the sample mean $\hat{\vartheta}_j(X_{1,j}, \ldots, X_{n,j}) = \bar{X}_j := n^{-1} \sum_{i=1}^n X_{i,j}$ in coordinate j. Based on this, we let $T_j = \sqrt{n}\left(\bar{X}_j - \theta_j\right)/\sigma_j$, $1 \leq j \leq m$. Since the data are assumed to be normally distributed, these test statistics are standard normally distributed under true null hypotheses. Thus, the p-values for the two-sided tests are given by $P_j = 2 \cdot (1 - \Phi(|T_j|))$, $1 \leq j \leq m$, where Φ denotes the cdf of the univariate standard normal distribution. Following Algorithm 1, we sample for every $1 \leq j \leq m$ the variates $X_{1,j}^*, \ldots, X_{k(n),j}^*$ i.i.d. from $\mathcal{N}\left(\hat{\vartheta}_j, \sigma_j^2\right)$ and plug their values into T_j. The resulting bootstrap test statistic T_j^* is equal to $\sqrt{k(n)}\left(\bar{X}_j^* - \theta_j\right)/\sigma_j$, and the bootstrap p-value P_j^* equals $2 \cdot \left(1 - \Phi\left(\left|T_j^*\right|\right)\right)$.

It is straightforward to check that our general Assumption 1 holds true for this example. Notice that part (b) of Assumption 1 is fulfilled by construction. Hence, it remains to check part (a), i. e., $\left\|F_{T_{j,n}} - F_{T_{j,n}^*}^*\right\|_\infty \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for all $1 \leq j \leq m_0$. To this end, notice that the term $\sqrt{k(n)}\left(\hat{\vartheta}_j - \theta_j\right)/\sigma_j = \sqrt{k(n)} \sum_{i=1}^n (x_{i,j} - \theta_j)/(\sigma_j n)$ converges to 0 for suitable choices of $k(n)$. For instance, we may choose $k(n)$ as the nearest integer to $n^{1-\varepsilon}$ with $\varepsilon > 0$. Then the bootstrap test statistics $T_{j,n}^* \sim \mathcal{N}\left(\sqrt{k(n)}\left(\hat{\vartheta}_j - \theta_j\right)/\sigma_j, 1\right)$, $1 \leq j \leq m$, are almost surely asymptotically standard normal under true null hypotheses by the dominated convergence theorem. By Pólyas uniform convergence theorem (Serfling 1980, sec. 1.5.3) the marginal cdfs converge uniformly.

Figure 2 displays the results of a computer simulation regarding the behavior of $\overline{\hat{\pi}}_0^*$ from Algorithm 1. The covariance matrix $\boldsymbol{\Sigma}$ has been generated randomly utilizing the function `genPositiveDefMat` from the `R` package `clusterGeneration`, and the means $\vartheta_{m_0+1}, \ldots, \vartheta_m \in (-10, 10)$ are chosen uniformly. The bootstrap estimate $\overline{\hat{\pi}}_0^* \approx 0.56$ is slightly larger than $\pi_0 = 0.5$. In general, the estimator $\overline{\hat{\pi}}_0^*$ is asymptotically positively biased as we will show in Theorem 1 below.

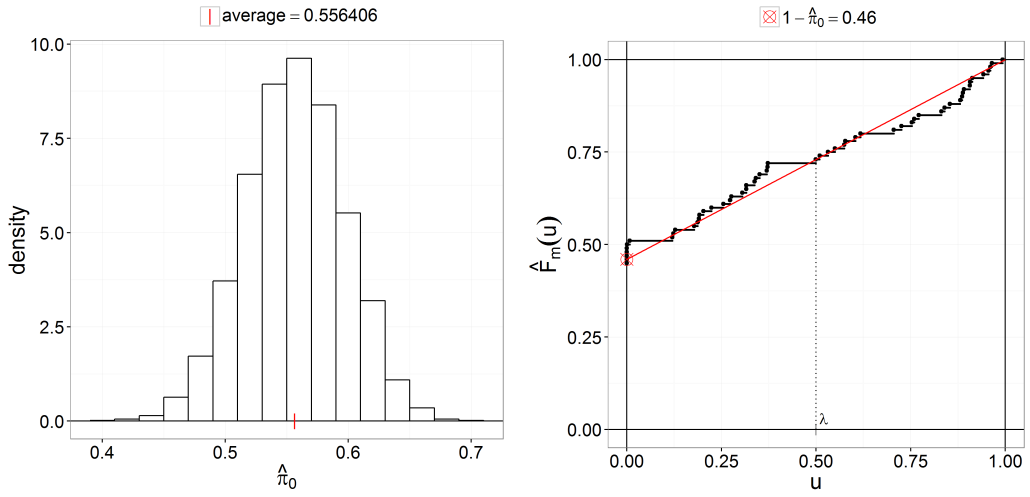

Figure 2: Computer simulation of $\overline{\hat{\pi}}_0^*$ in the setting of Example 2. The right graph displays one of the B Schweder-Spjøtvoll estimates based on the bootstrap p-values $P_1^*, \ldots, P_m^*$. The left graph displays the histogram of all estimates $\hat{\pi}_{0,n}^{*(1)}, \ldots, \hat{\pi}_{0,n}^{*(B)}$ with $B = 10{,}000$. In this simulation the bootstrap estimate of $\pi_0 = 0.5$ is $\overline{\hat{\pi}}_0^* = B^{-1} \sum_{b=1}^{B} \hat{\pi}_{0,n}^{*(b)} \approx 0.56$.

Example 3 (Example 1 continued.)**.** We consider the setup of Example 1, but now assume that the test statistics (instead of the p-values) are strongly dependent and possess a Gumbel-Hougaard copula with parameter η. Thus, the p-values cluster for large values of η, as in Example 1. Let $1 \leq j \leq m$ and assume that the sample $X_{1,j}, \ldots, X_{n,j}$ possesses the stochastic representation $X_{i,j} \overset{d}{=} \vartheta_j \cdot Z_{i,j}$, where $\vartheta_j > 0$ is unknown and $Z_{1,j}, \ldots, Z_{n,j}$ follow a beta distribution with fixed shape parameters $\alpha = 1$ and $\beta > 2$. For each $1 \leq j \leq m$ we want to test the null hypothesis $H_j = \{\boldsymbol{\vartheta} \in (0,\infty)^m \,|\vartheta_j = \theta_j\}$ versus the (one-sided) alternative $K_j := \{\boldsymbol{\vartheta} \in (0,\infty)^m \,|\vartheta_j > \theta_j\}$, where $\theta_j > 0$ is given. Assume that the dependence structure among $X_1, \ldots, X_m$ is given by a Gumbel-Hougaard copula with parameter η. According to Section 4.2 in Stange et al. (2015), suitable test statistics are given by $\tilde{T}_j = \max_{1 \leq i \leq n} X_{i,j}/\theta_j$, $1 \leq j \leq m$, and possess the same copula function. In order to get a non-degenerate limiting distribution function, we transform these test statistics to $T_j := \left(\tilde{T}_j - b_n\right)/a_n$, where $a_n = 1 - F_{Z_{1,1}}^{-1}\left(1 - n^{-1}\right)$ and $b_n \equiv 1$. Since a_n and b_n are deterministic quantities, the transformed test statistics follow the same Gumbel-Hougaard copula as well. The p-values for the one-sided hypotheses are given by $P_j := 1 - F_{\tilde{T}_j|\theta_j}\left(\tilde{T}_j\right) = 1 - F^n_{\mathrm{Beta}(\alpha,\beta)}(\tilde{T}_j)$. Since the random variables $Z_{i,j}$ are beta distributed, the expected value of $X_{1,j}$ is equal to $\vartheta_j \alpha/(\alpha+\beta)$. Hence, a consistent (method of moments) estimator $\hat{\vartheta}_j$ of ϑ_j is given

by $\hat{\vartheta}_j\left(X_{1,j},\ldots,X_{n,j}\right)=\bar{X}_j\left(\alpha+\beta\right)/\alpha$.

The plug-in rule of Algorithm 1 now yields bootstrap variates $X_{i,j}^{*}\overset{d}{=}\hat{\vartheta}_j\left(x_{1,j},\ldots,x_{n,j}\right)\cdot Z_{i,j}$, $1\leq i\leq n$, $1\leq j\leq m$, for observed data $\mathbf{x}_1,\ldots,\mathbf{x}_n$, with corresponding test statistics and p-values.

The validity of part (a) of Assumption 1 can be shown as follows. First, utilizing techniques from extreme value theory, we obtain that the marginal cdf of each original test statistic T_j converges under H_j to the (continuous) Weibull cdf with parameter $\beta>0$, i. e.,

$$\lim_{n\to\infty}F_{T_{j,n}}\left(x\right)=\lim_{n\to\infty}F_{Z_{1,1}}^{n}\left(a_nx+b_n\right)=G\left(x\right):=\begin{cases}\exp\left(-\left(-x\right)^{\beta}\right), & x<0,\\ 1, & x\geq 0.\end{cases}$$

In order to establish the limiting law of the bootstrapped test statistics, notice first that $T_j\overset{d}{=}\left(\max_{1\leq i\leq n}Z_{i,j}-b_n\right)/a_n$ under H_j. Let $a_n'\coloneqq a_n\hat{\vartheta}_{j,n}/\theta_j$ and $b_n'\coloneqq b_n\hat{\vartheta}_{j,n}/\theta_j$. We get that $a_n^{-1}a_n'$ converges to 1 $\mathbb{P}$-a.s. for $n\to\infty$. Now, assume for the moment that $a_n^{-1}\left(b_n'-b_n\right)$ converges to 0 $\mathbb{P}$-a.s. for $n\to\infty$. Then we get for all $1\leq j\leq m_0$ and $x\in\mathbb{R}$ that

$$\begin{aligned}\left|F_{T_{j,n}^{*}}^{*}\left(x\right)-G\left(x\right)\right|&=\left|F_{Z_{1,1}}^{n}\left(\frac{\theta_j}{\hat{\vartheta}_{j,n}}\left(a_nx+b_n\right)\right)-G\left(x\right)\right|\\&=\left|F_{Z_{1,1}}^{n}\left(a_n'x+b_n'\right)-G\left(x\right)\right|\\&=\left|F_{T_{j,n}}\left(a_n^{-1}a_n'x+a_n^{-1}\left(b_n'-b_n\right)\right)-G\left(x\right)\right|\\&\leq\left\|F_{T_{j,n}}-G\right\|_{\infty}+\left|G\left(a_n^{-1}a_n'x+a_n^{-1}\left(b_n'-b_n\right)\right)-G\left(x\right)\right|\\&\to 0\text{ for }n\to\infty\ \mathbb{P}-\text{a.s.}\end{aligned}$$

It remains to show that $a_n^{-1}\left(b_n'-b_n\right)\to 0$ almost surely as $n\to\infty$. To this end, notice that the convergence rate of $\hat{\vartheta}_{j,n}/\theta_j-1$ is arbitrarily close to $o\left(n^{-1/2}\right)$ $\mathbb{P}$-a.s. (Durrett 2010, thm. 2.5.8). For the chosen parameter values $\alpha=1$ and $\beta>2$, we get that $a_n=n^{-1/\beta}$ and $n^{(1/2-\epsilon)-1/\beta}\to\infty$ for any $\epsilon>0$ which is small enough. This means that $a_n^{-1}\left(b_n'-b_n\right)=o\left(1\right)/\left(n^{(1/2-\epsilon)-1/\beta}\right)$ indeed converges to 0 $\mathbb{P}$-a.s. for $n\to\infty$.

Finally, Pólyas uniform convergence theorem yields that

$$\left\|F_{T_j}-F_{T_j^{*}}^{*}\right\|_{\infty}\to 0$$

$\mathbb{P}$-a.s. for $n\to\infty$ and all $1\leq j\leq m_0$, since the limiting cdf G is continuous.

In analogy to Figure 2, Figure 3 displays the results of a computer simulation employing Algorithm 1 in this example, where $\eta=100$. One may compare the

right graph in Figure 3 with Figure 1 for a demonstration of the improvement of estimation accuracy obtained by applying Algorithm 1.

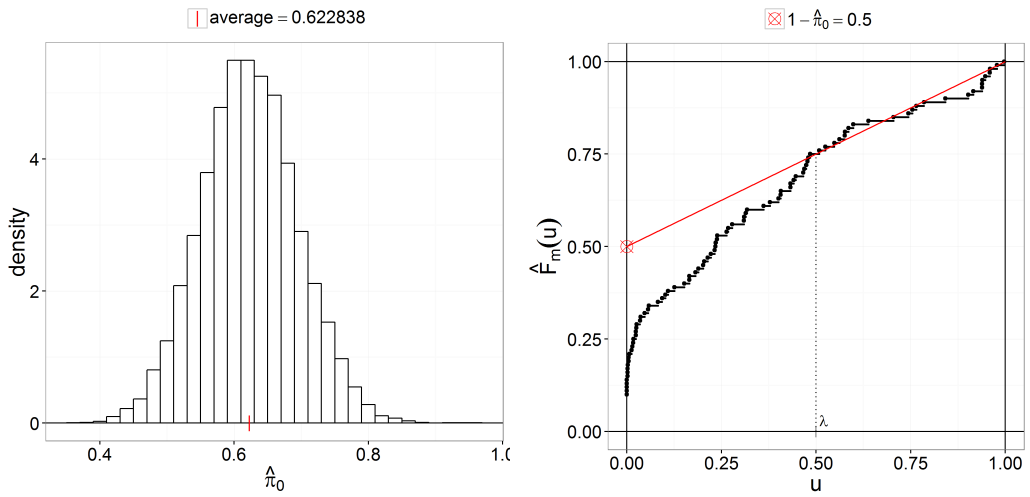

Figure 3: Computer simulation of $\overline{\hat{\pi}}_0^*$ in the setting of Example 3 with $\alpha = 1$, $\beta = 4$, $n = 100$, $m = 100$, $m_0 = 50$ and $\eta = 100$. The parameter values $\vartheta_{m_0+1}, \ldots, \vartheta_{100} \in (1.5, 2.5)$ have been chosen uniformly. The right graph displays one of the B Schweder-Spjøtvoll estimates based on the bootstrapped p-values $P_1^*, \ldots, P_m^*$. The left graph displays the histogram of all estimates $\hat{\pi}_{0,n}^{*(1)}, \ldots, \hat{\pi}_{0,n}^{*(B)}$ with $B = 10{,}000$. In this simulation the bootstrap estimate of $\pi_0 = 0.5$ is $\overline{\hat{\pi}}_0^* = B^{-1} \sum_{b=1}^{B} \hat{\pi}_{0,n}^{*(b)} \approx 0.62$.

3. THEORETICAL ANALYSIS

In this section we analyze the theoretical properties of our bootstrap estimator $\overline{\hat{\pi}}_0^*$. For the ease of notation let $\mathbb{P} = P_{\boldsymbol{\vartheta},C}^{\otimes n}$ denote the true distribution of the data sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and let $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ denote the probability space related to the bootstrap random variables for fixed data.

First, we prove that the bootstrap p-values are indeed independent.

Lemma 1. *Let $n \in \mathbb{N}$ and the observed data sample be fixed. Then the following assertions hold true.*

1. *The bootstrapped p-values $P_{1,n}^{*(b)}, \ldots, P_{m,n}^{*(b)}$ are stochastically independent with respect to $\mathbb{P}^*$ for every $1 \leq b \leq B$.*

2. *The estimators $\hat{\pi}_{0,n}^{*(1)}, \ldots, \hat{\pi}_{0,n}^{*(B)}$ are i.i.d. with respect to $\mathbb{P}^*$.*

Proof. For fixed data our bootstrap sample $\mathbf{X}_1^{*(b)}, \ldots, \mathbf{X}_n^{*(b)}$ consists of random variables $X_{i,j}^{*(b)}$, which are independent in i, j and b by construction. They are also identically distributed in i and b. In Algorithm 1 we just transform this sample to $T_{j,n}^{*(b)}$, $P_{j,n}^{*(b)}$ and $\hat{\pi}_{0,n}^{*(b)}(\lambda)$ for every $1 \leq b \leq B$. These measurable transformations depend on j, but not on b. Therefore, the assertions follow. □

Remark 2. Under the assumptions of Lemma 1 we get by the strong law of large numbers that

$$\overline{\hat{\pi}}_0^* \to \mathbb{E}^*\left[\hat{\pi}_{0,n}^{*(1)}\right] \ \mathbb{P}^*\text{-a.s. for } B \to \infty.$$

Lemma 2. *Let $b \in \{1, \ldots, B\}$ be fixed. Then*

$$\left\| F_{P_{j,n}} - F^*_{P_{j,n}^{*(b)}} \right\|_\infty \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty$$

for all $1 \leq j \leq m_0$.

Proof. From part (a) of Assumption 1 we get that

$$\left\| F_{T_{j,n}} - F^*_{T_{j,n}^{*(b)}} \right\|_\infty \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty$$

for each $1 \leq j \leq m_0$. Since the p-values are measurable transformations of the test statistics, the assertion follows. □

The following theorem is the main result of this section. Note that we can choose B as large as we want, if we have enough computing power. Therefore, the assertion of Theorem 1 is mainly an asymptotic property with respect to the sample size $n \to \infty$.

Theorem 1. *We have*

$$\lim_{n\to\infty} \lim_{B\to\infty} \overline{\hat{\pi}}^*_{0,n,B} \geq \pi_0$$

$\mathbb{P}^*$-a.s. and in probability with respect to $\mathbb{P}$.

Proof. To proof this theorem, we combine Remark 2 with Lemma 2. Let $(n_k)_{k\in\mathbb{N}} = (n_{\ell_k})_{k\in\mathbb{N}}$ be an arbitrary subsequence of a subsequence $(n_\ell)_{\ell\in\mathbb{N}}$ of integers. Remark 2 yields that for every observed data sample it holds $\mathbb{P}^*$-a.s. that

$$\begin{aligned}
\lim_{B\to\infty} \overline{\hat{\pi}}^*_{0,n_k,B} &= \mathbb{E}^*\left[\hat{\pi}_{0,n_k}^{*(1)}\right] \\
&= \mathbb{E}^*\left[\frac{1 - \hat{F}_m^{*(1)}(\lambda)}{1-\lambda}\right] \\
&= \frac{1 - \frac{1}{m}\sum_{j=1}^m \mathbb{P}^*\left[P^*_{j,n_k} \leq \lambda\right]}{1-\lambda}.
\end{aligned}$$

Furthermore, from Lemma 2 it follows that

$$\forall 1 \leq j \leq m_0 : \mathbb{P}^* \left[P^*_{j,n_k} \leq \lambda \right] \to \text{Prob}(U \leq \lambda) = \lambda \text{ as } n_k \to \infty$$

$\mathbb{P}$-a.s., where U denotes a standard uniform variate.

Hence, we get $\mathbb{P}$-a.s. that

$$\begin{aligned}
\lim_{n_k \to \infty} \frac{1 - m^{-1} \sum_{j=1}^{m} \mathbb{P}^* \left[P^*_{j,n_k} \leq \lambda \right]}{1 - \lambda} &= \lim_{n_k \to \infty} \frac{1 - \pi_0 \lambda - m^{-1} \sum_{j=m_0+1}^{m} \mathbb{P}^* \left[P^*_{j,n_k} \leq \lambda \right]}{1 - \lambda} \\
&\geq \frac{1 - \pi_0 \lambda - (1 - \pi_0)}{1 - \lambda} \\
&= \pi_0.
\end{aligned}$$

Thus, the assertion follows by the subsequence principle. □

Theorem 1 shows that the bootstrap estimator $\overline{\hat{\pi}}^*_0$ is asymptotically non-negatively biased (i. e., mean conservative). The theorem also shows that we achieve an asymptotically unbiased estimator of π_0 whenever $\mathbb{P}^* \left[P^*_{j,n} \leq \lambda \right]$ tends to one for $n \to \infty$ under alternatives. The latter condition is weaker than the assumption of a consistent multiple test (Troendle 2000).

Remark 3. For fixed $1 \leq b \leq B$ and varying m, the bootstrap estimator $\hat{\pi}^{*(b)}_{0,n} = \hat{\pi}^{*(b)}_{0,n}(m)$ converges to a number greater than or equal to π_0 in probability with respect to $\mathbb{P}^*$ and $\mathbb{P}$ for $n \to \infty$ and $m \to \infty$, if $\pi_0 := \lim_{m \to \infty} m_0(m)/m$ exists. This follows from the weak law of large numbers for triangular arrays applied to $\left\{ \mathbb{1}_{\{P^*_j \leq \lambda\}} : 1 \leq j \leq m, m \in \mathbb{N} \right\}$ (Lehmann and Romano 2005, lem. 15.4.1). Hence, $\hat{F}^*_m(\lambda)$ converges to its expected value in probability with respect to $\mathbb{P}^*$. The rest of the argumentation is analogous to the second part of the proof of Theorem 1.

4. REAL DATA ANALYSIS

In order to demonstrate the relevance of the proposed methodology in practice, we apply Algorithm 1 to a real-life gene expression dataset from Notterman et al. (2001). The goal of the study was to identify differentially expressed gene, RNA, and DNA profiles. The authors compared tumor tissues with paired healthy tissues from $n = 18$ adenocarcinomic cancer patients. On total, expression levels for 7457 RNA, DNA and gene entities were determined for every tissue. The available data matrix $(X_{j,i} : 1 \leq j \leq m, 1 \leq i \leq 2n)$ consists of m rows and $2n$ columns, where the first n columns contain the measurements for tumor tissues and the column $i + n$ contains

the paired measurement for healthy tissue, $1 \leq i \leq n$. Our model follows part (b) of Definition 10.1 of Dickhaus (2014). For every entity j under consideration, we test the null hypothesis of no differential expression between tumor tissue and healthy tissue. To this end, we take the logarithmic differences $D_{i,j} := \log(X_{i,j}) - \log(X_{i+n,j})$, $1 \leq i \leq n$, $1 \leq j \leq m$, and calculate the test statistic $T_j(\mathbf{X}_1, \ldots, \mathbf{X}_n) := \sqrt{n} \cdot \overline{D}_j / S_j$, where $\overline{D}_j$ is the sample mean and $S_j^2 := (n-1)^{-1} \sum_{i=1}^{n} \left(D_{i,j} - \overline{D}_j\right)^2$ is the sample variance of the logarithmic differences in coordinate j. Under the assumption that $D_{1,j}, \ldots, D_{n,j}$ are i.i.d. normally distributed with mean ϑ_j and finite variance, T_j follows under the null hypothesis $H_j : \{\vartheta_j = 0\}$ a Student's t-distribution with $n - 1 = 17$ degrees of freedom, for every $1 \leq j \leq m$. Corresponding two-sided p-values are given by $P_j := 2 \cdot (1 - F_{t_{17}}(|T_j|))$, $1 \leq j \leq m$.

In such type of data, there are often pronounced dependencies among the p-values due to the biological mechanism of co-regulation of genes. In the dataset at hand, for example, entity 4035 exhibits strong dependencies with a block of 57 interacting entities. Considering this block of entities for exemplary purposes, we obtain a multiple test problem with $m = 58$. As displayed in Figure 4, the m observed p-values were all extremely small, indicating that presumably all corresponding null hypotheses are false.

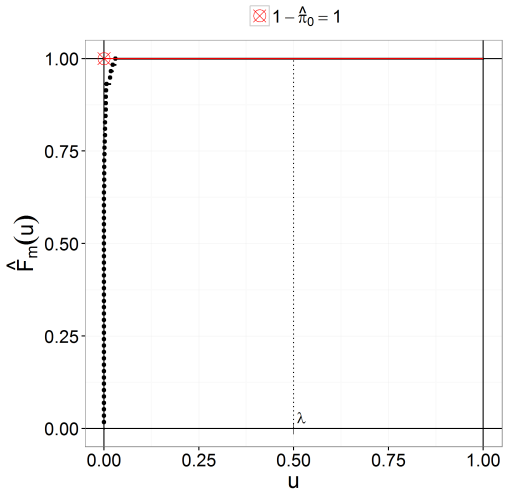

Figure 4: Empirical cdf of a block of $m = 58$ p-values calculated from the gene expression dataset of Notterman et al. (2001).

This allows for a data transformation yielding a pseudo ground truth for the evaluation of the accuracy of the considered estimators of π_0. Namely, we randomly chose $m_0 = 29$ of the m considered entities and subtracted from the corresponding raw measurements their respective group-specific mean, imitating the case of no dif-

ferential expression between the groups. Thereafter, we took the signed logarithm (see page 75 of Zumel and Mount (2014)) of the resulting measurements. This transformation does not change the dependency structure. With the so-transformed data, we re-computed the m p-values and applied the original Schweder-Spjøtvoll estimator as well as the proposed new estimator from Algorithm 1. Figure 5 demonstrates that our proposed algorithm can deal with the pronounced dependencies among the p-values much better than the original Schweder-Spjøtvoll estimator.

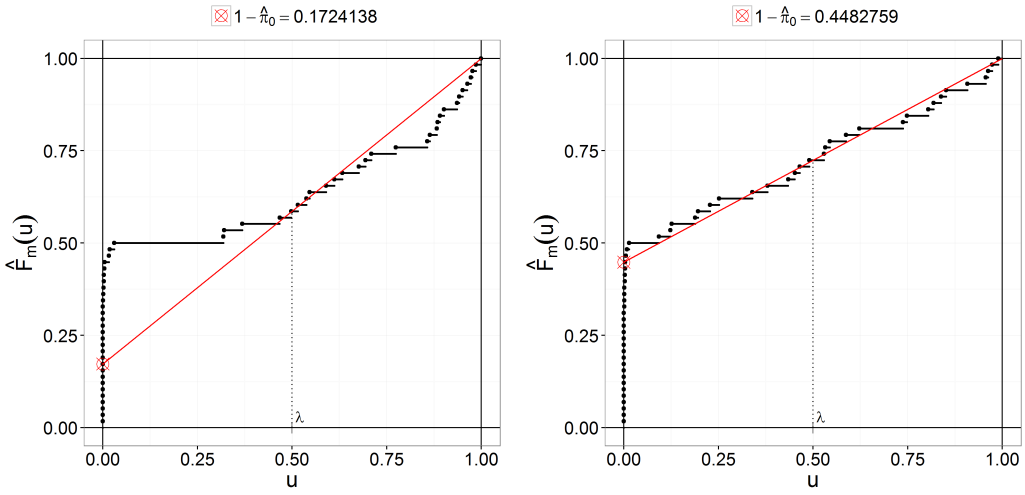

Figure 5: Results of the analysis of the transformed data from the gene expression dataset of Notterman et al. (2001). The left graph corresponds to the original Schweder-Spjøtvoll estimator. The right graph corresponds to one particular Monte Carlo run of Algorithm 1. In both groups of tissues the group-specific sample mean was subtracted from the original data for $m_0 = 29$ randomly chosen entities.

5. IMPLICATIONS FOR MULTIPLE TESTING

A multiple test for testing $\mathcal{H}$ is a measurable mapping $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_m) : \mathcal{X}^n \to \{0,1\}^m$, where $\varphi_j(\mathbf{x}_1, \ldots, \mathbf{x}_n) = 1$ means rejection of the j-th null hypothesis H_j in favor of the alternative K_j, $1 \leq j \leq m$. We restrict our attention to multiple tests which operate on the p-values $P_1, \ldots, P_m$ and are such that $\varphi_j = \mathbb{1}_{[0,\alpha_j)}(P_j)$, where $\alpha_1, \ldots, \alpha_m$ denote local (marginal) significance levels. The traditional type I error measure in multiple hypotheses testing is the family-wise error rate (FWER). For given values of $\boldsymbol{\vartheta}$ and C, it is defined by $\mathrm{FWER}_{\boldsymbol{\vartheta},C}(\boldsymbol{\varphi}) := \mathbb{P}_{\boldsymbol{\vartheta},C}\left(\bigcup_{j=1}^{m_0}\{\varphi_j = 1\}\right)$. A multiple test $\boldsymbol{\varphi}$ controls the FWER at level α in the strong sense if $\sup_{\boldsymbol{\vartheta}\in\Theta, C\in\mathcal{C}} \mathrm{FWER}_{\boldsymbol{\vartheta},C}(\boldsymbol{\varphi}) \leq \alpha$.

A simple, but often conservative method for achieving strong FWER control

under arbitrary copula C is the Bonferroni correction, meaning that the local significance levels are given by $\alpha_j = \alpha/m$ for all $1 \leq j \leq m$. In the sequel, we choose $\alpha = 0.05$. One approach to improve upon the Bonferroni correction is the so-called Bonferroni plug-in (BPI) method, where m is replaced by $\hat{m}_0 = m \cdot \hat{\pi}_0$ in the denominator of the α_j's, where $\hat{\pi}_0$ is some estimate of π_0.

For a comparison, Figure 6 displays histograms of the empirical powers (Dickhaus 2014, def. 1.4) of the original Bonferroni method and its plug-in version utilizing $\overline{\hat{\pi}}_0^*$ for the normal means problem considered in Example 2. The histograms are based on 100 simulation runs with $B = 10{,}000$ Monte Carlo repetitions in Algorithm 1 within each of these runs. For each simulation run the covariance matrix $\mathbf{\Sigma}$ with fixed diagonal elements $\sigma_1^2 = \ldots = \sigma_m^2 = 100$ has been generated randomly using the function `genPositiveDefMat` from the `R` package `clusterGeneration` (Joe 2006) and the means $\vartheta_1, \ldots, \vartheta_m$ are uniformly taken from $(-10, 10)$ under alternatives. The values $m = 100$, $m_0 = 20$, $n = 40$, and $\theta_1 = \ldots = \theta_m = 0$ have been fixed throughout the whole simulation. In every of the 100 simulation runs, the empirical power of the BPI method was larger than or equal to that of the original Bonferroni method, and the average improvement in empirical power was approximately 5.6%, see the right graph in Figure 6.

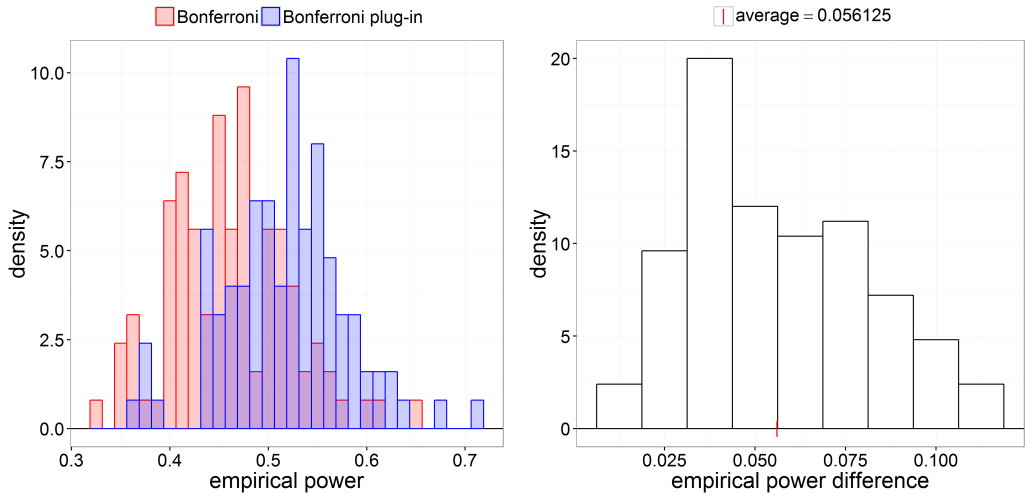

Figure 6: The left graph displays the histograms of the Bonferroni and the Bonferroni plug-in method resulting from Algorithm 1 over 100 simulation runs. The right graph displays a histogram of their power differences.

6. DISCUSSION

We have presented a method for estimating the proportion of true null hypotheses under arbitrary copula dependence. In contrast to multivariate multiple test procedures which explicitly exploit the dependencies in the data in order to relax the multiplicity adjustment in comparison with the independent case (Dickhaus and Stange 2013), addressing the estimation problem considered in this work profits from neglecting the dependencies, meaning that in the proposed marginal bootstrap procedure the true copula of the data is replaced by the independence copula.

There are a couple of potential modifications and extensions of the present statistical model which can be treated in an analogous manner. For example, consider the problem of "all pairs" comparisons (Tukey contrasts) in the balanced one-factorial ANOVA with k groups and n observational units per group. Here, the multiplicity of the multiple test problem equals $m = k(k+1)/2$, such that the dimension of ϑ and the multiplicity m do not coincide. Furthermore, we do not observe dependent data, but the dependencies in the test statistics are induced by utilizing the same data points in several of the test statistics (which are the scaled group-specific mean differences). However, this problem can easily be converted to our setup by re-organizing the data. Namely, one may construct a matrix $(X_{j,i}) : 1 \leq j \leq m, 1 \leq i \leq 2n$, where every row contains the data for exactly two of the k groups. With this construction, Algorithm 1 may readily be applied, and the dependency-inducing issue that data from one and the same group appear repeatedly (i. e., in more than one row) in the constructed matrix is addressed by our proposed marginal bootstrap method which only utilizes the estimated mean differences.

The obvious limitation of our approach is that only marginal parameters can be tested. We do not see any way of getting rid of part (b) of Assumption 1 in the case of a completely unspecified copula C. One could, however, consider special (parametric) model classes for C and design whitening procedures which exploit these parametric assumptions regarding the dependencies.

References

Celisse, A. and S. Robin (2010). A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference 140*(11), 3132–3147.

Cheng, Y., D. Gao, and T. Tong (2015). Bias and variance reduction in estimating the proportion of true-null hypotheses. *Biostatistics 16*(1), 189–204.

Dickhaus, T. (2013). Randomized p-values for multiple testing of composite null hypotheses. *Journal of Statistical Planning and Inference 143*(11), 1968–1979.

Dickhaus, T. (2014). *Simultaneous Statistical Inference with Applications in the Life Sciences*. Springer-Verlag Berlin Heidelberg.

Dickhaus, T. and M. Pauly (2016). Simultaneous statistical inference in dynamic factor models. In I. Rojas and H. Pomares (Eds.), *Time Series Analysis and Forecasting*, pp. 27–45. Springer.

Dickhaus, T. and J. Stange (2013). 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.

Dickhaus, T., K. Strassburger, D. Schunk, C. Morcillo-Suarez, T. Illig, and A. Navarro (2012). How to analyze many contingency tables simultaneously in genetic association studies. *Statistical Applications in Genetics and Molecular Biology 11*(4), Article 12.

Durrett, R. (2010, 007). *Probability: Theory and Examples* (4 ed.). Cambridge: Cambridge University Press.

Finner, H. and V. Gontscharuk (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society B 71*(5), 1031–1048.

Friguet, C. and D. Causeur (2011). Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Comput. Statist. Data Anal. 55*(9), 2665–2676.

Genest, C. and B. Rémillard (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. In *Annales de l'IHP Probabilités et statistiques*, Volume 44, pp. 1096–1127.

Hwang, Y.-T., H.-C. Kuo, C.-C. Wang, and M. F. Lee (2014). Estimating the number of true null hypotheses in multiple hypothesis testing. *Stat. Comput. 24*(3), 399–416.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis 97*(10), 2177–2189.

Krishnamoorthy, K., F. Lu, and T. Mathew (2007). A parametric bootstrap approach for anova with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis 51*(12), 5731–5742.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B, Statistical Methodology 67*(4), 555–572.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (3 ed.). Springer Texts in Statistics, New York, NY: Springer.

Meinshausen, N. and P. Bühlmann (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika 92*(4), 893–907.

Nguyen, V. H. and C. Matias (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Stat. 41*(4), 1167–1194.

Notterman, D. A., U. Alon, A. J. Sierk, and A. J. Levine (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer research 61*(7), 3124–3130.

Ostrovnaya, I. and D. L. Nicolae (2012). Estimating the proportion of true null hypotheses under dependence. *Statist. Sinica 22*(4), 1689–1716.

Oyeniran, O. (2016). *Estimating the Proportion of True Null Hypotheses in Multiple Testing Problems*. Ph. D. thesis, Bowling Green State University, Statistics.

Schweder, T. and E. Spjøtvoll (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika 69*, 493–502.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York.

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

Stange, J., T. Bodnar, and T. Dickhaus (2015). Uncertainty quantification for the family-wise error rate in multivariate copula models. *Advances in Statistical Analysis 99*(3), 281–310.

Stange, J., T. Dickhaus, A. Navarro, and D. Schunk (2016). Multiplicity- and dependency-adjusted p-values for control of the family-wise error rate. *Statistics & Probability Letters 111*, 32–40.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 66*(1), 187–205.

Troendle, J. F. (2000). Stepwise normal theory multiple test procedures controlling the false discovery rate. *Journal of Statistical Planning and Inference 84*(1-2), 139–158.

Van Der Laan, M. J. and J. Bryan (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics 2*(4), 445–461.

Wang, H.-Q., L. K. Tuominen, and C.-J. Tsai (2011). Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics 27*(2), 225.

Westfall, P. H. and S. S. Young (1993). *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.

Zumel, N. and J. Mount (2014). *Practical Data Science with R*. Manning, Shelter Island.