CORE
Metadata, citation and similar papers at core.ac.uk
Provided by E-LIB Dokumentserver - Staats und Universitätsbibliothek Bremen

Dissertation

Multiple Testing under Copula Dependency Structures

Submitted by

André Neumann

In partial fulfillment of the requirements

for the degree of

Doktor der Naturwissenschaften

(Dr. rer. nat.)

Supervisor and referee: Prof. Dr. Thorsten Dickhaus

Second referee: Prof. Dr. Gilles Blanchard

University of Bremen, Institute for Statistics

May 2018

Universität Bremen

Contents

Nomenclature

(e)cdf	(empirical) cumulative distribution function
(p)FDR	(positive) false discovery rate
ANOVA	analysis of variance
FWER	family-wise error rate
LFC	least favorable configuration
MSE	mean squared error
VaR	value-at-risk
i.i.d.	independent and identically distributed
w.l.o.g.	without loss of generality
$(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}^{\otimes n}_{\boldsymbol{\vartheta}, C_{\boldsymbol{X}}} : \boldsymbol{\vartheta} \in \Theta))$	statistical model
$B_{\boldsymbol{K}}$	Bernstein operator
$B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)$	Bernstein copula of $C_{\boldsymbol{X}}$
$C^{\leftarrow}$	quantile of $u \mapsto C(u, \ldots, u)$
$C_{\boldsymbol{X}}$	copula of $\boldsymbol{X}$
F_{X_j}	j-th marginal cumulative distribution function of $\boldsymbol{X}$
H_0	global null hypothesis $\bigcap_{j=1}^{m} H_j$
$H_{\boldsymbol{X}}$	joint cumulative distribution function of $\boldsymbol{X}$
$I_0(\boldsymbol{\vartheta})$	index set of true null hypotheses under $\boldsymbol{\vartheta}$
$K_1, \ldots, K_m$	alternative hypotheses
$P_1, \ldots, P_m$	p-values
$\mathbb{E}$	expectation with respect to $\mathbb{P}$
$\mathbb{E}^*$	expectation with respect to $\mathbb{P}^*$
$\mathbb{P}$	probability measure on the elemental space Ω or probability measure on the product space $\mathcal{X}^\infty$

$\mathbb{P}^*$	bootstrap probability measure
$\mathbb{P}_{\boldsymbol{X}}$	distribution of $\boldsymbol{X}$
α	global significance level
$\alpha_{loc,1}, \ldots, \alpha_{loc,m}$	local significance levels
$\boldsymbol{T} = (T_1, \ldots, T_m)^\top$	vector of test statistics
$\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$	bootstrap resample of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$
$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$	i.i.d. sample of $\boldsymbol{X}$
$\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_m)^\top$	multiple test
$\boldsymbol{\vartheta}$	parameter of interest
$\boldsymbol{\vartheta}^*$	least favorable parameter configuration
$\hat{C}_{\boldsymbol{X},n}$	empirical copula of $\boldsymbol{X}$
$\hat{F}_{X_j,n}$	j-th marginal empirical cumulative distribution function of $\boldsymbol{X}$
$\hat{H}_{\boldsymbol{X},n}$	joint empirical cumulative distribution function of $\boldsymbol{X}$
$\hat{\pi}_0^{\mathrm{SS}}$	Schweder-Spjøtvoll estimator
$(C([0,1]^m), \lVert\cdot\rVert_\infty)$	space of uniformly continuous functions defined on $[0,1]^m$
$(\ell^\infty([0,1]^m), \lVert\cdot\rVert_\infty)$	space of uniformly bounded functions on $[0,1]^m$
$\mathbb{C}$	limit process of $\mathbb{C}_n$
$\mathbb{C}_n$	empirical copula process
$\mathcal{H} = \{H_1, \ldots, H_m\}$	set of null hypotheses
$\mathbb{1}_{\mathcal{A}}$	indicator function of the set $\mathcal{A}$
π_0	proportion of true null hypotheses
$\xrightarrow{\mathbb{P}}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
m	number of null hypotheses

m_0	number of true null hypotheses
m_1	number of false null hypotheses
n	sample size
o, O	Landau symbols
$\mathbb{P}_{\boldsymbol{\vartheta}}$	$\mathbb{P}_{\boldsymbol{\vartheta},C_{\boldsymbol{X}}}^{\otimes n}$
$\boldsymbol{k}/\boldsymbol{K}$	$\left(\frac{k_1}{K_1},\ldots,\frac{k_m}{K_m}\right)^\top$
$(-\boldsymbol{\infty},\boldsymbol{x}]$	$(-\infty,x_1]\times\ldots\times(-\infty,x_m]$
$\{\mathbf{0},\ldots,\boldsymbol{K}\}$	$\{0,...,K_1\}\times\ldots\times\{0,\ldots,K_m\}$
$\sum_{\boldsymbol{k}=\mathbf{0}}^{\boldsymbol{K}}$	$\sum_{k_1=0}^{K_1}\cdots\sum_{k_m=0}^{K_m}$

1 Synopsis

The key to multiple testing is to respect the dependencies between the marginal hypotheses tests. Multiple tests can range from basically performing the same test multiple times to tests with very complex interactions. Any dependency structure can be modeled by so-called copula functions. This makes copulas an interesting tool in multiple testing. In particular, multivariate multiple tests explicitly utilize the dependency structure of the data. This leads to the sub-class of copula-based multiple tests.
In this synopsis, I give a general overview about multiple testing and copula theory with emphasis on their connections to my own contributions. Furthermore, I present the ideas and challenges behind my own research.

Der Schlüssel zum multiplen Testen liegt im Berücksichtigen der Abhängigkeiten zwischen den Randtests. Multiple Tests können dabei von einem quasi mehrfach ausgeführten Test bis hin zu komplex interagierenden Tests reichen. Jede Abhängigkeitsstruktur kann durch sogenannte Copula-Funktionen beschrieben werden. Dies macht Copulas zu einem interessanten Hilfsmittel im multiplen Testen. Insbesondere wird bei den multivariaten multiplen Tests die Abhängigkeitsstruktur der Daten explizit verwendet. Dies führt zur Unterklasse der Copula-basierten multiplen Tests.
In meiner Synopsis gebe ich einen generellen Einblick in die Theorie der multiplen Tests und der Copulas. Die Betonung liegt dabei auf der Einordnung meiner Resultate in diese Theorien. Zudem gehe ich auf die Ideen und Herausforderungen ein, die hinter meinen Forschungsarbeiten stecken.

1.1 Multiple testing

The problem of multiple testing arises when we have to answer two or more questions considering only one data set. For example, in genetic association studies, one hypothesis is tested for every genetic marker. It is important to respect the interactions between genetic markers. Usually, these interactions are modeled as block dependency structures. Such dependency structures play a crucial role in multiple testing. In Section 1.2, we take a closer look how to model dependency structures and how to use them in multiple testing.

To clarify, a multiple test is not a simple tool to make scientific studies cheaper by testing more hypotheses on the same data. In order to successfully apply multiple testing frameworks, one should ask as few questions as possible. For a large number of hypotheses m, it is often helpful to reduce m. This can be achieved by applying selection or filtering methods first. Statistical learning algorithms trained on past data sets is one possibility. Additionally, multiple tests for high-dimensional data can be applied. Still, model

assumptions like sparsity of the data set are necessary to achieve sufficient performance. Hence, we must carefully choose an appropriate model for each multiple test problem.

Mathematically, we test a set $\mathcal{H} = \{H_1, \ldots, H_m\}$ of m null hypotheses. Each null hypothesis H_j is a (non-empty) subset of the parameter space Θ and is tested against an alternative hypothesis $K_j := \Theta \backslash H_j$. For convenience and consistency, the index j denotes always a number in $\{1, \ldots, m\}$. Likewise, the index i is always in $\{1, \ldots, n\}$, where n is the sample size. A multiple test $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_m)^\top : \mathcal{X}^n \to \{0,1\}^m$ is a function on the set of data samples $\mathcal{X}^n$, which maps the observed data sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to a decision vector in $\{0,1\}^m$. $\varphi_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1$ means rejection of the null hypothesis H_j.

It is convenient to think of multiple tests in terms of test statistics $T_1, \ldots, T_m$, which tend to larger values under alternatives, or p-values $P_1, \ldots, P_m$, which tend to smaller values under alternatives. The p-value P_j is basically a transformation of the test statistic T_j to the uniform scale $[0,1]$. Such transformations are easier to interpret in terms of significance. For example, in contrast, a test statistic corresponding to the average height of some peoples is (hopefully) much smaller than a test statistic corresponding to average their income. However, this does not mean that their income is significant. Additionally, many multiple tests can be easier described in detail using p-values. Nonetheless, test statistics are important for understanding test procedures on a general level. Since p-values tend to smaller values under alternatives, we are interested in the boundary values of significant p-values for which a chosen error rate is controlled. These boundary values are called the local significance levels $\alpha_{loc,1}, \ldots, \alpha_{loc,m}$. In the remainder, we perform multiple tests by means of p-values and local significance levels. Therefore, $\varphi_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1$ if and only if $P_j = P_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) < \alpha_{loc,j}$.

The most common error rates in multiple testing are the family-wise error rate (FWER) and the false discovery rate (FDR). The FWER is older than the FDR and much stricter in terms of false rejections. A family-wise error occurs when at least one true null hypothesis is rejected. The FDR was introduced by Benjamini and Hochberg (1995) in order to relax this strict behavior and is defined as the expected proportion of false rejections. This means that not too many false rejections are acceptable for each multiple test. Mathematically, we have $\mathrm{FDR}_{\boldsymbol{\vartheta}}(\boldsymbol{\varphi}) \leq \mathrm{FWER}_{\boldsymbol{\vartheta}}(\boldsymbol{\varphi})$ for all parameter $\boldsymbol{\vartheta} \in \Theta$. Therefore, the FDR is used especially for high-dimensional problems.

Classification of multiple tests

The book of Dickhaus (2014) contains a wide and well organized classification of multiple tests. Classifications are important to better understand the big picture and the starting points of my own research. In this section, we follow essentially Section 3 of Dickhaus (2014).

There are three main classes, namely marginal-based multiple tests, multivariate multiple tests and closed test procedures. Marginal-based multiple test procedures do not directly utilize the dependency structure. Instead, they work for a wide class of dependency structures. For example, the Bonferroni procedure (see Bonferroni (1935, 1936)) is marginal-based and one of the earliest contributions in multiple testing. The so-called Bonferroni correction sets each local significance level to $\alpha_{loc,j} = \alpha/m$. We call it correction because for each local test we correct the global significance level α. The global level $\alpha \in (0,1)$ is an upper bound for the chosen multiple test error rate and usually set to 0.05 (0.01 or 0.1). This procedure controls the FWER and works under arbitrary dependency structures. Since this method is very easy to apply in practice, Bonferroni is still widely used.

The so-called stepwise multiple tests are contained in this class as well. The basic idea is to order the p-values and to compare each p-value P_j with a local significance level depending on the rank of P_j. For simplicity, let us just consider two examples here. The famous procedure of Benjamini and Hochberg (1995) sets the local significance level to $\alpha_{loc,(j)} = j\alpha/m$, where (j) denotes the index in $\{1,\dots,m\}$ of the j-th smallest p-value. We search for the first p-value in descending order (say $P_{(k)}$) which fails to be larger than $\alpha_{loc,(k)}$ and reject all null hypotheses $H_{(1)},\dots,H_{(k)}$. This procedure controls the FDR at level $m_0/m \cdot \alpha$ and works under a specific class of dependency structures (see Table 5.1 in Dickhaus (2014)). Another example is the method of Holm (1979), which sets $\alpha_{loc,(j)} = \alpha/(m-j+1)$. In contrast to Benjamini and Hochberg (1995), we search for the first p-value in ascending order (again $P_{(k)}$) which fails to be smaller than or equal to $\alpha_{loc,(k)}$ and reject $H_{(1)},\dots,H_{(k-1)}$. This procedure controls the FWER and generally improves the Bonferroni method.

Contrarily to marginal-based multiple tests, multivariate multiple test procedures explicitly use the dependency structure of the data. Subclasses are resampling-based, central limit theorem based and copula-based methods. Let us just consider an example for the first subclass. The multivariate bootstrap (see Efron (1979)) creates resamples by drawing from the original sample with replacement. Notice that it is important to sample with replacement. Otherwise, test statistics like the sample mean would be constant. For each resample, we evaluate the test statistics. Hence, we obtain a sample of the test statistics. This allows us to empirically calibrate the local significance levels. The bootstrap works well in one sample problems for various test statistics. In one sample problems, all observed data originate from the same population. In these settings, the bootstrap procedure asymptotically approximates the distribution of the test statistics. More specifically, this approximation holds almost surely or in probability with respect to the distribution of the data. In terms of copula theory, we implicitly utilize the empirical copula in this proce-

dure. This means that there are connections between resampling-based and copula-based multiple tests. We refer to Westfall and Young (1993) for an algorithm-focused book on resampling-based multiple tests.

In central limit theorem based multiple tests, the test statistics are transformations of an asymptotically normally distributed point estimator. In this way, the asymptotic distribution of the test statistics can be derived. Examples are multiple linear regression models and generalized linear models.

The emphasis of my work lies on copula-based methods. In copula-based methods, we explicitly model the dependency structure in the most general framework possible. There are two main applications. First, the FWER can be represented by the copula of the test statistics. This will be discussed further in Section 1.2. Second, we can resample from an estimated copula of the data. This falls in the category of resampling-based methods.

Closed test procedures (see Marcus et al. (1976)) cannot be exactly assigned to one of the previous two classes. Under some assumptions, we can modify existing tests and these tests can be of either class. Let us assume for now that the set of null hypotheses $\mathcal{H}$ contains all intersection null hypotheses. Then, a multiple test $\boldsymbol{\varphi}'$ for $\mathcal{H}$ can be constructed by applying the so-called closure principle on a multiple test $\boldsymbol{\varphi}$ for $\mathcal{H}$. Mathematically, the test $\boldsymbol{\varphi}'$ is defined by $\varphi'_j := \min_{H_i \subseteq H_j} \varphi_i$. Any (coherent) test $\boldsymbol{\varphi}$ with local significance level set to $\alpha_{loc,j} = \alpha$ can be modified in this way to control the FWER at level α. Such a modified multiple test $\boldsymbol{\varphi}'$ is possibly more powerful than $\boldsymbol{\varphi}$. Notice that we can always construct an intersection-closed set $\mathcal{H}$. Hence, the main restriction is that we need local level α tests for all intersections.

Any of the mentioned classes above could be further refined by the used error rate (FWER versus FDR) or data (low-dimensional versus high-dimensional). In my contributions, we consider only the FWER and low-dimensional data in the sense that the dimension m is fixed.

1.2 Copula theory in multiple testing

The word copula means link and was introduced by Sklar (1959). A copula function links the marginal cumulative distribution functions (cdfs) together to a joint cdf. Therefore, a copula C can be seen as the dependency structure between the marginal cdfs. Mathematically, a copula $C : [0,1]^m \subset \mathbb{R}^m \to [0,1]$ is a joint distribution function of a uniformly distributed random vector with the domain restricted to $[0,1]^m$. This restriction is unproblematic because the probability mass of these random vectors is zero outside of $[0,1]^m$. Hence, there exists a one-to-one connection between copulas and joint cdfs of uniformly distributed random vectors. Sklar's theorem provides the relationship between the joint cdf, the marginal cdfs and the copula. This theorem is the foundation of statistical mod-

eling using copula functions.

Theorem 1.1 (Sklar (1959)). *Let $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$ be a random vector with values in $\mathbb{R}^m$ and joint cdf $H_{\boldsymbol{X}}$. Further, let $F_{X_1}, \ldots, F_{X_m}$ denote the marginal cdfs of $\boldsymbol{X}$. Then there exists a copula $C_{\boldsymbol{X}}$ such that*

$$H_{\boldsymbol{X}}(\boldsymbol{x}) = C_{\boldsymbol{X}}\left(F_{X_1}(x_1), \ldots, F_{X_m}(x_m)\right)$$

for all $\boldsymbol{x} \in \mathbb{R}^m$. If all marginal cdfs are continuous, then the copula $C_{\boldsymbol{X}}$ is unique.

Copula theory is focused mainly on the construction of suitable copula classes and the analysis of their structure. A standard reference for a good overview of copulas is the book of Nelsen (2006). For copula theory in the context of risk management, we refer to the books of Embrechts et al. (2003) and McNeil et al. (2005).

Since we are in a more general setting, classical linear dependency structures in form of correlations are contained as well. The Gaussian copula corresponds to the correlation matrix of a normally distributed random vector. Likewise, the t-copula corresponds to the dependency structure of a (standard) multivariate t-distribution. We could of course exchange these copulas and think of marginal t-distributions combined with a Gaussian copula. In this way, new multivariate distributions can be constructed. Therefore, copulas are a very flexible and general way of modeling joint cdfs.

Some important non-parametric copulas are the empirical copula and Bernstein copulas. Strictly speaking, the empirical copula fails to be continuous and therefore, is not a copula. Nonetheless, the empirical copula can be used, in particular, to construct proper copulas. Bernstein copulas are such examples. They play a crucial role in our paper Neumann et al. (forthcoming) about multiple testing based on non-parametric copula estimation (see Section 1.3).

Connection to multiple testing

In multiple testing, we can use Sklar's theorem to model the FWER. This enables us to think of the FWER in terms of the test statistics copula.

Lemma 1.2 (Dickhaus and Gierl (2013)). *Under some model assumptions, we have*

$$FWER_{\boldsymbol{\vartheta}, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) \leq 1 - C_{\boldsymbol{T}}\left(1 - \alpha_{loc,1}, \ldots, 1 - \alpha_{loc,m}\right),$$

where $\boldsymbol{\vartheta} \in \Theta$ is any parameter vector and $\alpha_{loc,1}, \ldots, \alpha_{loc,m}$ are the local significance levels.

In our multiple testing setup, we are interested in an parameter vector $\boldsymbol{\vartheta} \in \Theta$ corresponding to the marginal cdfs of the data. In this setting, the copula of the data $C_{\boldsymbol{X}}$ is an

infinite dimensional nuisance parameter and assumed to be independent of the parameter vector $\boldsymbol{\vartheta}$. On the other hand, the copula of the test statistics can depend on the parameter vector. Hence, the notation in Lemma 1.2 is somewhat imprecise. We assume that there exists a least favorable parameters configuration in the global null hypothesis $\bigcap_{j=1}^{m} H_j$. The notation $C_{\boldsymbol{T}}$ corresponds to this worst case. Often, only linear dependencies in the form of correlations are considered in multiple testing. We are interested in what we can achieve with this more general setup.

1.3 My contributions

Multivariate multiple test procedures based on non-parametric copula estimation[1]

The starting point for this paper are mainly two contributions. The first one is the statistical analysis of the so-called Bernstein copulas in Janssen et al. (2012) and the second one is the analysis of the FWER for parametric copula models in Stange et al. (2015). In this work, we have analyzed the FWER in a semi-parametric framework. More precisely, the hypotheses correspond to a finite dimensional parameter vector and the data copula is understood as an infinite dimensional nuisance parameter. The argumentation is similar as in Stange et al. (2015), but dropping the continuous differentiability assumption for the quantile $C_{\boldsymbol{T}}^{\leftarrow}$ of the test statistics copula $C_{\boldsymbol{T}}$ provided some extra challenges. To clarify, by quantile of a copula C I mean the quantile of the univariate function $u \mapsto C(u, \ldots, u)$. An estimator of $C_{\boldsymbol{T}}^{\leftarrow}$ is needed in order to estimate the local significance levels $\alpha_{loc,j} = \alpha_{loc} := 1 - C_{\boldsymbol{T}}^{\leftarrow}(1-\alpha)$.

This makes it necessary to extend the results of the used non-parametric copula estimator to a suitable function space. In our theoretical analysis, we focused on Bernstein copulas. These copulas are smoothed versions of the empirical copula with Bernstein polynomials. In contrast to the empirical copula, Bernstein copulas are indeed copula functions. Our analysis is based on the results of Segers (2012) about the empirical copula in the function space of bounded functions. Previous works on Bernstein copulas in statistics have focused mainly on pointwise results for two-dimensional data (see, e.g., Janssen et al. (2012) and Belalia (2016)). Furthermore, we have extended these results to (fixed) higher dimensions $m > 2$. A general analysis of Bernstein copulas in higher dimensions has been done in Sancetta and Satchell (2004).

In order to deduce asymptotic normality for the FWER as $n \to \infty$, we have proven that for the quantiles of Bernstein copulas hold pointwise asymptotic normality (at point $1-\alpha$). Although this is a pointwise result, we utilize the uniform results for Bernstein copulas in the space of continuous functions. On the other hand, the consistency of the FWER

[1] Neumann et al. (forthcoming)

follows directly from the consistency of Bernstein copulas. The uniform consistency of Bernstein copulas is already known in two dimensions. Additionally, the argumentation is the same for fixed dimension $m > 2$. Hence, the core of our theoretical analysis is the asymptotic normality.

As mentioned before, we have reduced the assumptions. However, we additionally assume that the copula of the data can be transformed locally to the copula of the test statistics on the diagonal set $\{(u,\dots,u)^\top | u \in [0,1]\}$. In this setting, Bernstein copulas are used to approximate only the data copula C_X. Unfortunately, this assumption is hard to verify and to exploit in practice. In Bodnar and Dickhaus (2014) and Stange et al. (2015), the dependency structure among the test statistics or p-values is assumed to follow a parametric copula. Additionally, they utilize resampling methods to create an approximate sample of the test statistics (or p-values). In our paper, the strategy is similar in practice. However, we generate resamples by using a Bernstein copula of the data. After that, we calibrate the multiple test empirically. In terms of copula theory, this means that we use the empirical copula quantile of the resampled p-values for calibration.

Estimating the proportion of true null hypotheses under arbitrary dependency[2]

The idea for considering the proportion of true null hypotheses was to improve methods like the Benjamini-Hochberg procedure. This procedure controls the FDR at level $\pi_0\alpha \leq \alpha$, where $\pi_0 := m_0/m$ is the proportion of true null hypotheses and m_0 is the number of true null hypotheses. Of course, m_0 is unknown and an estimator of π_0 could be used to improve this procedure such that the error rate is (approximately) controlled at level α. However, in order to avoid confusion, we focused solely on this estimation problem. Besides, it can be helpful on its own to know the value of π_0. Some multiple testing methods perform better for larger (or smaller) values of π_0 than others. For example, our Bernstein procedure works better for smaller values of π_0. To clarify, in this manuscript, we did not use the connection between multiple tests and copulas in the sense of Lemma 1.2. We have estimated π_0 only in models where the dependency structure of the p-values is modeled by copulas.

The basic estimator of π_0 was introduced by Schweder and Spjøtvoll (1982). One of the assumptions for this estimator is the independence of the p-values under true null hypotheses. In multiple testing, such an assumption is often violated. There exists a vast literature on this topic but not in the context of copula theory. In the existing literature, independent p-values are still often assumed or models with some specific dependency structures are considered. For example, Tong et al. (2013) modified the Schweder-Spjøtvoll estimator for various patterns of the p-value histogram. Implicitly, these patterns

[2] Neumann et al. (preprint)

correspond to dependency structures among the p-values.

Our initial approach was to transform the p-values utilizing the copula directly. For example, in Archimedean copula models, we constructed an algorithm based on using the sampling procedure of Wu et al. (2007) backwards. Unfortunately, this only works under very restrictive assumptions. Therefore, instead, we have constructed new p-values by utilizing a marginal parametric bootstrap algorithm. This means that we split the original data sample and apply the univariate bootstrap on every marginal sample $x_{1,j}, \ldots, x_{n,j}$ given the estimated parameters. More specifically, the algorithm works as follows.

Algorithm 1.3 (Marginal parametric bootstrap)**.**

1. *Resample from the j-th marginal distribution function of the data given the estimated parameters.*
2. *Estimate the p-value P_j by using this Bootstrap sample.*
3. *Apply step 2 and 3 to every margin j and then estimate the ratio π_0.*
4. *Repeat the steps 2-4 B times.*
5. *Take the average over all B estimated values of π_0 as $\hat{\pi}_0$.*

The conditional nature (with respect to the observed data) of the bootstrap translates to the conditional independence of these bootstrap p-values. Assumptions of the bootstrap like suitable test statistics translate to our procedure as well. Additionally, we make a model assumption on the parameters for the marginal cdfs. This is necessary in order to split the sample without losing information about the parameters of interest. Fortunately, these assumptions are not hard to check.

Under a specific (mild) assumption, we have proven that $\hat{\pi}_0$ is an consistent estimator. If the assumption is not met, then the estimator is asymptotically positively biased. In multiple testing, this could mean that procedures based on this estimator are more conservative than in the unbiased case.

References

Belalia, M. (2016). On the asymptotic properties of the Bernstein estimator of the multivariate distribution function. *Stat. Probab. Lett. 110*, 249–256.

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

Bodnar, T. and T. Dickhaus (2014). False discovery rate control under Archimedean copula. *Electron. J. Stat. 8*(2), 2207–2241.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8*, 3–62.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, 13–60.

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

Dickhaus, T. and J. Gierl (2013). 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)*, pp. 75–80. Global Science and Technology Forum (GSTF).

Efron, B. (1979, jan). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Embrechts, P., F. Lindskog, and A. McNeil (2003). Modelling dependence with copulas and applications to risk management. In S. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384. Elsevier Science B.V.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat., Theory Appl. 6*, 65–70.

Janssen, P., J. Swanepoel, and N. Veraverbeke (2012). Large sample behavior of the Bernstein copula estimator. *J. Statist. Plann. Inference 142*(5), 1189–1197.

Marcus, R., E. Peritz, and K. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative risk management. Concepts, techniques, and tools.* Princeton, NJ: Princeton University Press.

Nelsen, R. B. (2006). *An introduction to copulas. 2nd ed.* Springer Series in Statistics. New York, NY: Springer.

Neumann, A., T. Bodnar, and T. Dickhaus (preprint). Estimating the proportion of true null hypotheses under copula dependency. *Stockholm University Research Report 2017*.

Neumann, A., T. Bodnar, D. Pfeifer, and T. Dickhaus (forthcoming). Multivariate multiple test procedures based on nonparametric copula estimation. *Biometrical Journal*.

Sancetta, A. and S. Satchell (2004). The bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory 20*(03), 535–562.

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

Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli 18*(3), 764–782.

Sklar, M. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

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

Tong, T., Z. Feng, J. S. Hilton, and H. Zhao (2013). Estimating the proportion of true null hypotheses using the pattern of observed *p*-values. *J. Appl. Stat. 40*(9), 1949–1964.

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.

Wu, F., E. Valdez, and M. Sherris (2007). Simulating from exchangeable Archimedean copulas. *Commun. Stat., Simulation Comput. 36*(5), 1019–1034.

2 Multivariate multiple test procedures based on non-parametric copula estimation

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

Multivariate multiple test procedures have received growing attention recently. This is due to the fact that data generated by modern applications typically are high-dimensional, but possess pronounced dependencies due to the technical mechanisms involved in the experiments. Hence, it is possible and often necessary to exploit these dependencies in order to achieve reasonable power. In the present paper, we express dependency structures in the most general manner, namely, by means of copula functions. One class of non-parametric copula estimators is constituted by Bernstein copulas. We extend previous statistical results regarding bivariate Bernstein copulas to the multivariate case and study their impact on multiple tests. In particular, we utilize them to derive asymptotic confidence regions for the FWER of multiple test procedures which are empirically calibrated by making use of Bernstein copulas approximations of the dependency structure among the test statistics. This extends a similar approach by Stange et al. (2015) in the parametric case. A simulation study quantifies the gain in FWER level exhaustion and, consequently, power which can be achieved by exploiting the dependencies, in comparison with common threshold calibrations like the Bonferroni or Šidák corrections. Finally, we demonstrate an application of the proposed methodology to real-life data from insurance.

Key words: Asymptotic oscillation behavior; Family-wise error rate; p-Value; Risk management

2.1 Introduction

Copula-based modeling of dependency structures has become a standard tool in applied multivariate statistics and quantitative risk management (see, e.g., Nelsen (2006), Joe (2014), Härdle and Okhrin (2010), Embrechts et al. (2003), and Chapter 5 of McNeil et al. (2005)). The estimation of an unknown copula is key to a variety of modern multivariate statistical methods. In particular, applications of copulas to the calibration and the analysis of multiple tests have been considered by Dickhaus and Gierl (2013), Bodnar and Dickhaus (2014), Stange et al. (2015), Cerqueti et al. (2012), Schmidt et al. (2014),

[1]Institute for Statistics, University of Bremen, Bibliothekstraße 1, D-28359 Bremen, Germany.
[2]Department of Mathematics, Stockholm University, Roslagsvägen 101, SE-10691 Stockholm, Sweden.
[3]Institute of Mathematics, Carl von Ossietzky University of Oldenburg, D-26111 Oldenburg, Germany.

and Schmidt et al. (2015); see also Sections 2.2.4 and 4.4 of Dickhaus (2014). Specifically, the copula-based construction of multiple test procedures developed by Dickhaus and Gierl (2013) and Stange et al. (2015) under parametric assumptions regarding the type of dependencies among test statistics considerably extends previous approaches as in Hothorn et al. (2008) which are confined to asymptotic Gaussianity and, consequently, linear dependencies.

In the case of a parametric copula, generic estimation techniques like the (generalized) method of moments or maximum likelihood estimation are established notions (cf. Section 3.2 of Stange et al. (2015) and references therein). The empirical copula as well as its asymptotic properties as a non-parametric estimator have been studied, among others, by Rüschendorf (1976), Deheuvels (1979), Stute (1984), and, more recently, by Bücher and Dette (2010), and Bouzebda and Zari (2013), to mention only a few references. However, similarly as multivariate histogram estimators, the empirical copula in dimension m has some undesirable properties. For example, it is discontinuous, and it typically assigns zero mass to large subsets of $[0,1]^m$, even if the sample size n is large, due to the concentration of measures phenomenon. One way to tackle these issues consists of smoothing of the empirical copula. In particular, Sancetta and Satchell (2004) proposed smoothing by Bernstein polynomials, leading to so-called Bernstein copulas. Approximation theory for Bernstein copulas has been derived by Cottin and Pfeifer (2014), and asymptotic statistical properties of Bernstein copula estimators in the bivariate case ($m = 2$) have been proven by Janssen et al. (2012) and Belalia (2016). Functional central limit theorems for empirical copula processes have been established by Segers (2012). Applications of Bernstein copulas to modeling dependencies in non-life insurance have been considered by Diers et al. (2012).

In the present work, we contribute to theory and applications of Bernstein copulas in the case of a general dimension $m \geq 2$. In Section 2.2, we extend the asymptotic theory regarding Bernstein copula estimators by proving its rate of convergence in infinity norm as well as its asymptotic normality in a function space, for arbitrary m. Also, we provide some justifications for the proposed smoothing approach. Section 2.3 is then devoted to applications of Bernstein copulas for multiple test procedures with control of the FWER, avoiding restrictive parametric dependency assumptions. The application of the central limit theorem derived in Section 2.2 allows for a precise quantification of the uncertainty about the realized FWER in the case that the copula of test statistics is pre-estimated prior to calibrating the significance thresholds of the multiple test procedure. This extends the results of Stange et al. (2015) to the case of non-parametric copula pre-estimation. Section 2.4 demonstrates by means of a simulation study that the latter pre-estimation approach leads to a better exhaustion of the FWER level and thus enhances the

power of the multiple test procedure compared with traditional approaches which only take univariate marginal distributions of test statistics into account. Finally, we apply the proposed multiple testing methodology to real-life data from insurance in Section 2.5, and we conclude with a discussion in Section 2.6. Lengthy proofs and some auxiliary results are deferred to Section 2.7.

2.2 Oscillation behavior of Bernstein copulas

In this section, asymptotic properties of (empirical) Bernstein copulas are studied. The main properties of Bernstein estimators are consistency (Theorem 2.1) and asymptotic normality (Theorem 2.4). The auxiliary lemmas can be found in Section 2.7. Nonetheless, the argumentation in this section is illustrated in some mathematical detail. More practically oriented readers might find Section 2.2.2 and the following sections more valuable. In Section 2.3, the methodology how to use this estimator in multiple testing is discussed and examples are given. The consistency of the realized FWER can be derived directly from the consistency of the Bernstein estimator. The asymptotic normality of the realized FWER follows indirectly from the asymptotic normality of the Bernstein estimator via Lemma 2.18.

Let $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$ be a random vector taking values in the probability space $(\mathcal{X}, \mathcal{F}, \mathbb{P}_{\boldsymbol{X}})$, where $\mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{F}$ is a σ-field over $\mathcal{X}$, and $\mathbb{P}_{\boldsymbol{X}}$ denotes the (joint) distribution of $\boldsymbol{X}$. The univariate marginal cdfs of $\boldsymbol{X}$ we denote by F_{X_j}, $1 \leq j \leq m$, whereas $C_{\boldsymbol{X}}$ stands for the copula related to the distribution $\mathbb{P}_{\boldsymbol{X}}$.

Assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are stochastically independent and identically distributed (i.i.d.) random vectors with $\boldsymbol{X}_1 \sim \mathbb{P}_{\boldsymbol{X}}$. Then, the marginal empirical cumulative distribution function (ecdf) $\hat{F}_{X_j,n}$ of $(X_{1,j}, \ldots, X_{n,j})^\top$ is given by $\hat{F}_{X_j,n}(x_j) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,x_j]}(X_{i,j})$, $1 \leq j \leq m$, and the joint ecdf is defined as $\hat{H}_{\boldsymbol{X},n}(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,\boldsymbol{x}]}(\boldsymbol{X}_i)$. The symbol $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function of set $\mathcal{A}$ and $(-\infty, \boldsymbol{x}] = (-\infty, x_1] \times \ldots \times (-\infty, x_m]$. We will use an analogous bold-face notation for vectors throughout the remainder. Finally, the empirical copula $\hat{C}_{\boldsymbol{X},n}$ pertaining to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is given by

$$\hat{C}_{\boldsymbol{X},n}(\boldsymbol{u}) = \hat{H}_{\boldsymbol{X},n}\left(\hat{F}^{\leftarrow}_{X_1,n}(u_1), \ldots, \hat{F}^{\leftarrow}_{X_m,n}(u_m)\right), \qquad \boldsymbol{u} \in [0,1]^m.$$

In this, $\hat{F}^{\leftarrow}_{X_j,n}$ denotes the generalized inverse of the marginal ecdf in coordinate $1 \leq j \leq m$.

2.2.1 Theoretical analysis

Denote the space of bounded functions on $[0,1]^m$, equipped with the supremum norm, by $(\ell^\infty([0,1]^m), \|\cdot\|_\infty)$, and the space of continuous (and bounded) functions defined on

$[0,1]^m$ by $(C([0,1]^m), \|\cdot\|_\infty)$, where $\|\cdot\|_\infty$ again denotes the supremum norm. The Bernstein copula estimation is based on the Bernstein polynomial approximation, which for a fixed copula C_X is given by the operator $B_{\boldsymbol{K}} : (\ell^\infty([0,1]^m), \|\cdot\|_\infty) \to (C([0,1]^m), \|\cdot\|_\infty)$ defined by

$$B_{\boldsymbol{K}}(f)(\boldsymbol{u}) := \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{K}} f(\boldsymbol{k}/\boldsymbol{K}) \prod_{j=1}^{m} P_{k_j,K_j}(u_j)$$

evaluated at the function $f = C_X$, where $\sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{K}} := \sum_{k_1=0}^{K_1} \cdots \sum_{k_m=0}^{K_m}$, $\boldsymbol{k}/\boldsymbol{K} := \left(\frac{k_1}{K_1}, \ldots, \frac{k_m}{K_m}\right)^\top$,

$$P_{k,K}(u) := \binom{K}{k} u^k (1-u)^{K-k},$$

and $K_1, \ldots, K_m$ are given positive integers. The Bernstein copula estimator for C_X is then defined by $B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right)$.

It is well known that continuous functions can be approximated using Bernstein polynomials. There are results on the convergence rate for continuous functions with bounded variation as well (see Chêng (1983)). For the special case of copula functions, it has been proved in Corollary 3.1 of Cottin and Pfeifer (2014) that any copula function can be approximated uniformly using Bernstein polynomials.

Theorem 2.1 establishes the consistency rate of Bernstein copula estimators for any copula function C_X. This result is known for the bivariate case (see Theorem 1 in Janssen et al. (2012)).

Theorem 2.1 (**Chung-Smirnov consistency rate**). *Let m be fixed. If $\boldsymbol{K} = \boldsymbol{K}(n)$ is such that $\sum_{j=1}^m K_j^{-1/2} = O\left(n^{-1/2}(\log\log n)^{1/2}\right)$, then it holds that*

$$\left\| B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right) - C_X \right\|_\infty = O\left(n^{-1/2}(\log\log n)^{1/2}\right) \quad \textit{almost surely,}$$

where $\|g\|_\infty := \sup_{\boldsymbol{u}\in[0,1]^m} |g(\boldsymbol{u})|$ for $g : [0,1]^m \longrightarrow \mathbb{R}$.

Proof. The proof can be done analogously to the proof of the bivariate case considered in Janssen et al. (2012). By the triangle inequality we split the convergence of the Bernstein copula estimators into an inner and outer convergence. It holds that

$$\left\| B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right) - C_X \right\|_\infty \leq \left\| B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right) - B_{\boldsymbol{K}}(C_X) \right\|_\infty + \|B_{\boldsymbol{K}}(C_X) - C_X\|_\infty . \tag{2.1}$$

For the outer convergence, we get from Lemma 2.17 and our assumption that

$$\|B_{\boldsymbol{K}}(C_X) - C_X\|_\infty = O\left(n^{-1/2}(\log\log n)^{1/2}\right).$$

The argumentation for the inner convergence is more complicated. For the first summand

in (2.1), we get

$$\left\|B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)-B_{\boldsymbol{K}}\left(C_{\boldsymbol{X}}\right)\right\|_{\infty} \leq \sup_{\boldsymbol{u}\in[0,1]^m}\sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{K}}\left|\hat{C}_{\boldsymbol{X},n}\left(\boldsymbol{k}/\boldsymbol{K}\right)-C_{\boldsymbol{X}}\left(\boldsymbol{k}/\boldsymbol{K}\right)\right|\prod_{j=1}^{m}P_{k_j,K_j}\left(u_j\right)$$
$$\leq \max_{\boldsymbol{k}\in\{\boldsymbol{0},\ldots,\boldsymbol{K}\}}\left|\hat{C}_{\boldsymbol{X},n}\left(\boldsymbol{k}/\boldsymbol{K}\right)-C_{\boldsymbol{X}}\left(\boldsymbol{k}/\boldsymbol{K}\right)\right|,$$

where $\{\boldsymbol{0},\ldots,\boldsymbol{K}\} := \{0,...,K_1\}\times\ldots\times\{0,\ldots,K_m\}$. Let $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$ be a sample of random vectors defined by $U_{i,j} := F_j\left(X_{i,j}\right)$, $1\leq i\leq n$, $1\leq j\leq m$. Application of the identity (see, e.g., Section 3 of Swanepoel (1986)) $\hat{F}^{\leftarrow}_{U_j,n}\left(u_j\right) = F_{X_j}\left(\hat{F}^{\leftarrow}_{X_j,n}\left(u_j\right)\right)$ leads to $\hat{C}_{\boldsymbol{X},n}\left(\boldsymbol{k}/\boldsymbol{K}\right) = \hat{H}_{\boldsymbol{U},n}\left(\hat{F}^{\leftarrow}_{U_1,n}\left(\frac{k_1}{K_1}\right),\ldots,\hat{F}^{\leftarrow}_{U_m,n}\left(\frac{k_m}{K_m}\right)\right)$ and

$$\left\|B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)-B_{\boldsymbol{K}}\left(C_{\boldsymbol{X}}\right)\right\|_{\infty} \leq \max_{\boldsymbol{k}\in\{\boldsymbol{0},\ldots,\boldsymbol{K}\}}\left|\hat{C}_{\boldsymbol{X},n}\left(\boldsymbol{k}/\boldsymbol{K}\right)-C_{\boldsymbol{X}}\left(\boldsymbol{k}/\boldsymbol{K}\right)\right|$$
$$\leq \max_{\boldsymbol{k}\in\{\boldsymbol{0},\ldots,\boldsymbol{K}\}}\left|\hat{C}_{\boldsymbol{X},n}\left(\boldsymbol{k}/\boldsymbol{K}\right)-C_{\boldsymbol{X}}\left(\hat{F}^{\leftarrow}_{U_1,n}\left(\frac{k_1}{K_1}\right),\ldots,\hat{F}^{\leftarrow}_{U_m,n}\left(\frac{k_m}{K_m}\right)\right)\right| \tag{2.2}$$
$$+\sum_{j=1}^{m}\max_{k_j\in\{0,\ldots,K_j\}}\left|\hat{F}^{\leftarrow}_{U_j,n}\left(\frac{k_j}{K_j}\right)-\frac{k_j}{K_j}\right|. \tag{2.3}$$

From Theorem 2 of Kiefer (1961), we get that the summand in (2.2) is of order $O\left(n^{-1/2}\left(\log\log n\right)^{1/2}\right)$ as well as that each summand in (2.3) is of order $O\left(n^{-1/2}\left(\log\log n\right)^{1/2}\right)$. This completes the proof. □

Remark 2.2. If m is not fixed, then the convergence rate in the last step of previous proof changes to $O\left(mn^{-1/2}\left(\log\log n\right)^{1/2}\right)$. Hence, we get almost surely

$$\left\|B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)-C_{\boldsymbol{X}}\right\|_{\infty} = O\left(mn^{-1/2}\left(\log\log n\right)^{1/2}\right).$$

The next theorem is taken from Whitt (2002) and will be useful in order to show asymptotic normality of Bernstein copula estimators.

Theorem 2.3 (Generalized Continuous Mapping Theorem). *Let g and g_n, $n\geq 1$, be measurable functions mapping (S,d) into (S',d'). Let the range (S',d') be separable. Further, let E be the set of x in S such that $g_n\left(x_n\right)\longrightarrow g\left(x\right)$ fails for some sequence $\{x_n : n\geq 1\}$ with $x_n\longrightarrow x$ in S. If $X_n\xrightarrow{d}X$, $n\longrightarrow\infty$, in (S,d) ($\xrightarrow{d}$ denotes the convergence in distribution) and $\mathbb{P}\left[X\in E\right]=0$, then $g_n\left(X_n\right)\xrightarrow{d}g\left(X\right)$, $n\longrightarrow\infty$, in (S',d').*

Furthermore, we need a result for the convergence of the empirical copula process $\mathbb{C}_n := n^{1/2}\left(\hat{C}_{\boldsymbol{X},n}-C_{\boldsymbol{X}}\right)$. Let $\boldsymbol{u}\mapsto\gamma\left(\boldsymbol{u}\right)$ be a $C_{\boldsymbol{X}}$-Brownian bridge, i.e., a zero mean Gaus-

sian process with (almost surely) continuous paths and covariance function given by

$$\mathrm{Cov}(\gamma(\boldsymbol{u}),\gamma(\boldsymbol{v})) = C_X(\boldsymbol{u}\wedge\boldsymbol{v}) - C_X(\boldsymbol{u})C_X(\boldsymbol{v})$$

for all $\boldsymbol{u},\boldsymbol{v}\in[0,1]^m$. Denote $\gamma_j(u_j) := \gamma(1,\ldots,1,u_j,1,\ldots,1)$. Then under some assumptions, the process $\mathbb{C}(\boldsymbol{u}) := \gamma(\boldsymbol{u}) - \sum_{j=1}^m \partial_j C_X(\boldsymbol{u})\gamma_j(u_j)$ is the weak limit of the empirical copula process $\mathbb{C}_n$ in $(\ell^\infty([0,1]^m), \|\cdot\|_\infty)$ as shown in Proposition 3.1 of Segers (2012). With these two arguments we can proof a functional central limit theorem for Bernstein copula estimators.

Theorem 2.4 (**Asymptotic normality**). *Let m be fixed. Assume that the first order partial derivatives of C_X exist and are continuous. If $\boldsymbol{K} = \boldsymbol{K}(n)$ is such that $n^{1/2}\sum_{j=1}^m K_j^{-1/2} \to 0$, $n\to\infty$, then it holds that*

$$n^{1/2}\cdot\left(B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right) - C_X\right) \xrightarrow{d} \mathbb{C} \text{ as } n\to\infty$$

in $(C([0,1]^m), \|\cdot\|_\infty)$.

Remark 2.5. The assumption of the existence and continuity of the first order partial derivatives on the boundaries can be weakened (see Condition 2.1 of Segers (2012)).

Proof. We split the Bernstein copula process $n^{1/2}\cdot\left(B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right) - C_X\right)$ into two parts. We get

$$\begin{aligned} n^{1/2}\cdot\left(B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right) - C_X\right) &= B_{\boldsymbol{K}}\left(n^{1/2}\left(\hat{C}_{X,n} - C_X\right)\right) + n^{1/2}\left(B_{\boldsymbol{K}}(C_X) - C_X\right) \\ &= B_{\boldsymbol{K}}(\mathbb{C}_n) + n^{1/2}\left(B_{\boldsymbol{K}}(C_X) - C_X\right). \end{aligned}$$

The second summand converges uniformly to zero because of Lemma 2.17 and our assumptions. The first summand is the empirical copula process $\mathbb{C}_n$ transformed by a family of operators $B_{\boldsymbol{K}}$, where $\boldsymbol{K} = \boldsymbol{K}(n)$.

We will use the Generalized Continuous Mapping Theorem 2.3. Let $(S,d) := (\ell^\infty([0,1]^m), \|\cdot\|_\infty)$ and $(S',d') := (C([0,1]^m), \|\cdot\|_\infty)$. Then (S',d') is a separable space, since the set of polynomials on $[0,1]^m$ with rational coefficients is a countable dense subset of S'. Further, let $g_n : S\to S'$ be defined by $g_n := B_{\boldsymbol{K}(n)}$ and $g : S\to S'$ be the identity function on S' and arbitrary on $S\backslash S'$. Notice that it does not matter, how g is defined on $S\backslash S'$, since we are interested in $g(\mathbb{C})$ and without loss of generality (w.l.o.g.) $\mathbb{C}$ takes values in S' (see Section 3 of Segers (2012)). Let E be the set of f in S such that $g_n(f_n)\to g(f)$ fails for some sequence $\{f_n : n\geq 1\}$ with $f_n\to f$ in S. Then $E\subseteq S\backslash S'$, since we can choose $f_n := f$ for $f\in S'$ and get uniform convergence

by Bernstein's theorem (or by using Corollary 3.1 of Cottin and Pfeifer (2014)). Hence, $\mathbb{P}[\mathbb{C} \in E] \leq \mathbb{P}[\mathbb{C} \in S \backslash S'] = 0$.

The last thing we need to check is the weak convergence of the empirical copula process $\mathbb{C}_n$ to $\mathbb{C}$ in $(\ell^\infty([0,1]^m), \|\cdot\|_\infty)$. As already mentioned, Segers (2012) has shown this convergence under assumptions only regarding the first order partial derivatives of C_X. Therefore, the proof is complete by using Proposition 3.1 of Segers (2012) and the generalized continuous mapping theorem. □

This result extends the pointwise central limit theorems of Janssen et al. (2012) and Belalia (2016) and works under weaker assumptions as well.

2.2.2 The effect of smoothing

This section is meant to be an addition to the simulation study of Omelka et al. (2009). Conducting such an extensive study ourselves would go beyond the scope of this paper. Nevertheless, it is an important question how precise the Bernstein estimator is compared to other copula estimators, and this should be discussed at least to some extent.

There exists a wide variety of methods to estimate copula functions non-parametrically. Usually, the empirical copula or some sort of smoothing method is used. Bernstein estimators studied in Section 2.2.1 are only one specific smoothing method among many others. Further examples comprise kernel (density) estimators (see Gijbels and Mielniczuk (1990)), and beta density estimators (see Chen (1999)). It is beyond the scope of the present work to compare all these competing approaches in detail. Generally speaking, the empirical copula is robust and universal, but it is not a copula in the strict sense, because it lacks continuity and does not have uniform margins. Bernstein copulas are differentiable estimators, but they converge rather slowly and cannot capture extreme tail dependencies (see Sancetta and Satchell (2004)). Recently, families of non-parametric copula estimators capable of modeling (positive) tail dependence have been studied by Pfeifer et al. (2017). Kernel methods suffer from a boundary bias, although several modifications like the mirror approach by Schuster (1985) exist to address this problem. Beta density estimators avoid the boundary bias, but the choice of their smoothing parameter is not trivial.

Let us briefly provide some numerical justifications for smoothing of the empirical copula. In Section 3 of Omelka et al. (2009) some kernel methods have been compared in simulations under two prototypical models (Model 1 and Model 2). In Model 1, the data follow a Frank copula with parameter corresponding to Kendall's $\tau = 0.25$. In Model 2, a Clayton copula corresponding to Kendall's $\tau = 0.75$ is used.

We have applied our proposed Bernstein estimator to these models as well. Figure 1 displays the results of a simulation study under these two models. The box plots demon-

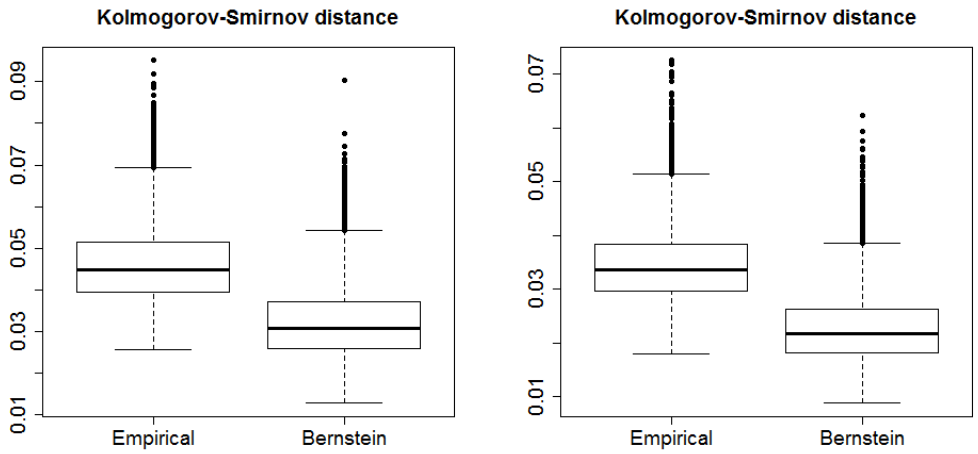

Figure 1: Comparison of the Bernstein copula and the empirical copula in the setting of Model 1 (left) and Model 2 (right) of Omelka et al. (2009) with respect to the supremum norm (Kolmogorov-Smirnov distance).

strate that the estimation accuracy (measured in terms of the Kolmogorov-Smirnov distance) can be improved by smoothing. Here, we only considered smoothing by means of Bernstein polynomials, but the simulation results for various kernel methods presented by Omelka et al. (2009) are very similar. Hence, in practice it may not be most important which smoothing method to choose, while it is recommendable to smooth at all. For a more detailed overview on copula estimation methods, see Charpentier et al. (2007).

2.3 Calibration of multivariate multiple test procedures

In this section, we assume that we have uncertainty about the distribution of $\mathbf{X}$. We thus consider a statistical model of the form $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_{\boldsymbol{\vartheta}, C_X} : \boldsymbol{\vartheta} \in \Theta, C_X \in C))$. The probability measure $\mathbb{P}_{\boldsymbol{\vartheta}, C_X}$ is indexed by two parameters. The parameter C_X denotes the copula of $\mathbf{X}$, and $\boldsymbol{\vartheta}$ is a vector of marginal parameters which refer to $F_{X_1}, \ldots, F_{X_m}$. The model for the i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ consequently reads as $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}^{\otimes n}_{\boldsymbol{\vartheta}, C_X} : \boldsymbol{\vartheta} \in \Theta, C_X \in C))$.

Based on this model, we consider multiple test problems of the form $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}^{\otimes n}_{\boldsymbol{\vartheta}, C_X} : \boldsymbol{\vartheta} \in \Theta, C_X \in C), \mathcal{H})$, where $\mathcal{H} = \{H_1, \ldots, H_m\}$ with $\emptyset \neq H_j \subset \Theta$ for all $1 \leq j \leq m$ denotes a family of m null hypotheses regarding the parameter $\boldsymbol{\vartheta}$. For notational convenience, we will write $\mathbb{P}_{\boldsymbol{\vartheta}, C_X}$ for $\mathbb{P}^{\otimes n}_{\boldsymbol{\vartheta}, C_X}$. The copula C_X is not the primary target of statistical inference, but a nuisance parameter in the sense that it does not depend on $\boldsymbol{\vartheta}$. This is a common setup in multiple test theory. We will mainly consider a semi-parametric situation, where Θ is of finite dimension, while C is a function space.

Remark 2.6. The assumption that the number of tests equals the dimension of $\boldsymbol{X}$ is only made for notational convenience. The case that these two quantities differ can be treated

with obvious modifications.

A multiple test for a given set of hypotheses $\mathcal{H}$ is a measurable mapping $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_m)^\top : \mathcal{X}^n \to \{0,1\}^m$, where $\varphi_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1$ for given data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ means rejection of the j-th null hypothesis H_j in favor of the alternative $K_j = \Theta \setminus H_j$, $1 \leq j \leq m$. We restrict our attention to multiple tests $\boldsymbol{\varphi}$ which are such that the hypotheses are rejected if the respective test statistics are large enough for given data, i.e., larger than their corresponding critical values. Notationally, this mean that

$$\varphi_j = \mathbb{1}_{(c_j, \infty)}(T_j), \qquad 1 \leq j \leq m, \tag{2.4}$$

where $\boldsymbol{T} = (T_1, \ldots, T_m)^\top : \mathcal{X}^n \to \mathbb{R}^m$ denotes a vector of real-valued test statistics which tend to larger values under alternatives, and $\mathbf{c} = (c_1, \ldots, c_m)^\top$ are the critical values. In many problems of practical interest, T_j will only use the marginal data $(x_{i,j})_{1 \leq i \leq n}$, for every $1 \leq j \leq m$. For example, this typically holds true if ϑ_j only corresponds to F_{X_j}, and H_j only concerns ϑ_j, for every $1 \leq j \leq m$.

For the calibration of $\mathbf{c}$, we aim at controlling the FWER in the strong sense. Strictly speaking, our procedure will only control the FWER under the global null hypothesis in the first place. However, strong control follows directly under Assumption 2.7 (a). For sufficient conditions of this assumption see Lemma 2.8.

For given $\boldsymbol{\vartheta} \in \Theta$ and $C_X \in C$, the FWER is defined as the probability for at least one false rejection (type I error) of $\boldsymbol{\varphi}$ under $\mathbb{P}_{\boldsymbol{\vartheta}, C_X}$, i.e.,

$$\mathrm{FWER}_{\boldsymbol{\vartheta}, C_X}(\boldsymbol{\varphi}) = \mathbb{P}_{\boldsymbol{\vartheta}, C_X}\left[\bigcup_{j \in I_0(\boldsymbol{\vartheta})} \{\varphi_j = 1\}\right],$$

where $I_0(\boldsymbol{\vartheta}) = \{1 \leq j \leq m : \boldsymbol{\vartheta} \in H_j\}$ denotes the index set of true null hypotheses under $\boldsymbol{\vartheta}$. The multiple test $\boldsymbol{\varphi}$ is said to control the FWER at level $\alpha \in [0,1]$, if

$$\sup_{\boldsymbol{\vartheta} \in \Theta, C_X \in C} \mathrm{FWER}_{\boldsymbol{\vartheta}, C_X}(\boldsymbol{\varphi}) \leq \alpha.$$

Notice that, although the trueness of the null hypotheses is determined by $\boldsymbol{\vartheta}$ alone, the FWER depends on $\boldsymbol{\vartheta}$ and C_X, because the dependency structure in the data typically influences the distribution of $\boldsymbol{\varphi}$ when regarded as a statistic with values in $\{0,1\}^m$.

Throughout the remainder, we assume that the following set of conditions is fulfilled.

Assumption 2.7.

(a) Letting $H_0 = \bigcap_{j=1}^m H_j$ denote the global null hypothesis of $\mathcal{H}$, there exists a least

favorable configuration (LFC) $\boldsymbol{\vartheta}^* \in H_0$ *such that*

$$\forall \boldsymbol{\vartheta} \in \Theta \forall C_{\boldsymbol{X}} \in \mathcal{C} : FWER_{\boldsymbol{\vartheta}, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) \leq FWER_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}).$$

If this assumption is fulfilled, then weak FWER control implies strong FWER control. Notice that this assumption can be weakened by considering closed test procedures, where our proposed methodology is applied to every non-empty intersection hypothesis in $\mathcal{H}$ *(cf. Remark 1 of Stange et al. (2016) for details). However, in such a setting, the computation time for the multiple test can increase very fast with the number of hypotheses.*

(b) The vector of marginal cdfs of $\boldsymbol{T} = (T_1, \ldots, T_m)^\top$ *depends on* $\boldsymbol{\vartheta}$ *only, and is (at least asymptotically as* $n \to \infty$*) known under any LFC* $\boldsymbol{\vartheta}^*$*. We denote the vector of marginal cdfs of* $\boldsymbol{T} = (T_1, \ldots, T_m)^\top$ *under such an LFC* $\boldsymbol{\vartheta}^*$ *by* $\boldsymbol{F}_{\boldsymbol{T}} = (F_{T_1}, \ldots, F_{T_m})^\top$*.*

(c) Letting $C_{\boldsymbol{T}} = C_{\boldsymbol{T}, \boldsymbol{\vartheta}^*}$ *denote the copula of* $\boldsymbol{T}$ *under* $\boldsymbol{\vartheta}^*$ *from part (b), there exists a continuously differentiable function* $h : [0,1] \to [0,1]$ *such that* $C_{\boldsymbol{T}}(u, \ldots, u) = h(C_{\boldsymbol{X}}(u, \ldots, u))$ *for all* $u \in [0,1]$*, where* $C_{\boldsymbol{X}}$ *is the copula of* $\mathbf{X}$*. The function* h *may be unknown. Notice that, if* T_j *only uses the data* $(x_{i,j})_{1 \leq i \leq n}$*, for every* $1 \leq j \leq m$*, then the copula of* $\boldsymbol{T}$ *is independent of* $\boldsymbol{\vartheta}^*$*. The existence of* h *is guaranteed whenever plateaus of* $u \mapsto C_{\boldsymbol{X}}(u, \ldots, u)$ *occur on the same subset of* $[0,1]$ *as plateaus of* $u \mapsto C_{\boldsymbol{T}}(u, \ldots, u)$*. In particular,* h *exists if* $u \mapsto C_{\boldsymbol{X}}(u, \ldots, u)$ *is strictly increasing. The more crucial part of the assumption is that* h *needs to be continuously differentiable.*

The following lemma is useful in order to verify assumption (a).

Lemma 2.8. *Let* $H_j : \left\{\boldsymbol{\vartheta} \in \Theta \middle| \vartheta_j \in \Theta_j \subseteq \mathbb{R}\right\}$*,* $1 \leq j \leq m$*, such that the global null hypothesis* H_0 *is not empty and let the marginal distributions of the data in coordinate* j *depend on* ϑ_j *only. Further, assume that every test statistic* T_j *only uses the data* $(x_{i,j})_{1 \leq i \leq n}$*. Then for all* $\boldsymbol{\vartheta} \in \Theta$*,* $C_{\boldsymbol{X}} \in \mathcal{C}$ *and any multiple test* $\boldsymbol{\varphi}$ *which is as in (2.4), we can construct a parameter value* $\boldsymbol{\vartheta}^* \in H_0$ *with*

$$FWER_{\boldsymbol{\vartheta}, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) \leq FWER_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}).$$

In particular, this implies that the LFC is located in H_0*.*

Proof. Choose $\boldsymbol{\vartheta}^* \in H_0 \neq \emptyset$ with $\vartheta_j^* = \vartheta_j$ for $j \in I_0(\boldsymbol{\vartheta})$. Then it holds that

$$\mathbb{P}_{\boldsymbol{\vartheta}, C_{\boldsymbol{X}}}\left[\bigcup_{j \in I_0(\boldsymbol{\vartheta})} \{T_j > c_j\}\right] = \mathbb{P}_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}}\left[\bigcup_{j \in I_0(\boldsymbol{\vartheta})} \{T_j > c_j\}\right],$$

since it is assumed that the test statistics T_j, $j \in I_0(\boldsymbol{\vartheta})$, only utilize the data from that coordinate j. Hence,

$$\begin{aligned}
\mathrm{FWER}_{\boldsymbol{\vartheta},C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) &= \mathbb{P}_{\boldsymbol{\vartheta},C_{\boldsymbol{X}}}\left[\bigcup_{j\in I_0(\boldsymbol{\vartheta})}\{T_j > c_j\}\right] \\
&= \mathbb{P}_{\boldsymbol{\vartheta}^*,C_{\boldsymbol{X}}}\left[\bigcup_{j\in I_0(\boldsymbol{\vartheta})}\{T_j > c_j\}\right] \\
&\leq \mathbb{P}_{\boldsymbol{\vartheta}^*,C_{\boldsymbol{X}}}\left[\bigcup_{j=1}^{m}\{T_j > c_j\}\right] \\
&= \mathrm{FWER}_{\boldsymbol{\vartheta}^*,C_{\boldsymbol{X}}}(\boldsymbol{\varphi}).
\end{aligned}$$

□

More generally, the previous lemma holds if the test statistics satisfy the so-called subset pivotality condition (see Westfall and Young (1993) and Dickhaus and Stange (2013)). Before we start to explain the proposed method for the calibration of $\mathbf{c}$, let us illustrate prototypical example applications of our general setup.

Example 2.9.

(a) Let $\Theta = \mathbb{R}^m$ and assume that $\vartheta_j \in \mathbb{R}$ is the expected value of X_j for every $1 \leq j \leq m$. The j-th null hypothesis may be the one-sided null hypothesis $H_j = \{\vartheta_j \leq 0\}$ with corresponding alternative $K_j = \{\vartheta_j > 0\}$. Assume that the variance of the marginal distribution of each X_j is known and w.l.o.g. equal to one. A suitable test statistic T_j is then given by $T_j(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) = \sum_{i=1}^n X_{i,j}/\sqrt{n}$. From Lemma 2.8 it follows that the LFC lies in H_0. Since the test statistics tend to get larger with increasing values of $\boldsymbol{\vartheta}$, the LFC $\boldsymbol{\vartheta}^*$ equals $\mathbf{0}$. Under $\boldsymbol{\vartheta}^*$, we have that $F_{T_j} = \Phi$ (the cdf of the standard normal law on $\mathbb{R}$) is the cdf of the (asymptotic) null distribution of T_j for every $1 \leq j \leq m$. If the considered copula family C consists of multivariate stable copulas (meaning that the observables follow a multivariate stable distribution), then the copula $C_{\boldsymbol{T}}$ is of the same type as $C_{\boldsymbol{X}}$, hence all parts of Assumption 2.7 are fulfilled.

(b) Let $\mathcal{X} = [0,\infty)$ and assume that the stochastic representations $X_j \overset{d}{=} \vartheta_j Z_j$ with $\vartheta_j > 0$ hold true for all $1 \leq j \leq m$, where Z_j is a random variable taking values in $[0,1]$. The parameter of interest in this problem is $\boldsymbol{\vartheta} \in (0,\infty)^m$. For each coordinate j, we consider the pair of hypotheses $H_j : \{\vartheta_j \leq \vartheta_j^*\}$ versus $K_j : \{\vartheta_j > \vartheta_j^*\}$, where the LFC $\boldsymbol{\vartheta}^* \in (0,\infty)^m$ (same argumentation as in (a)) is identical to the hypothesized upper bounds for the supports (or right end-points of the distributions) of

the X_j's. This has applications in the context of stress testing in actuarial science and financial mathematics (cf., e.g., Longin (2000)). Suitable test statistics are given by the component-wise maxima of the observables, i.e., $T_j(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) = \max_{1\le i\le n} X_{i,j}/\vartheta_j^*$, $1 \le j \le m$. Assuming that the tail behavior of each X_j is known such that the marginal (limiting) extreme value distribution of T_j under $\boldsymbol{\vartheta}^*$ can be derived and letting C consist of max-stable copulas, all parts of Assumption 2.7 are fulfilled here, too.

Let us remark here that these two examples have been treated under the restrictive assumption of one-parametric copula families C by Stange et al. (2015). The following lemma is taken from Dickhaus and Gierl (2013) and connects the FWER with the test statistics copula $C_{\boldsymbol{T}}$.

Lemma 2.10. *Let Assumption 2.7 be fulfilled. Then we have that*

$$FWER_{\boldsymbol{\vartheta},C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) \le 1 - C_{\boldsymbol{T}}\left(1-\alpha_{loc}^{(1)},\ldots,1-\alpha_{loc}^{(m)}\right),$$

where $\alpha_{loc}^{(j)} = 1 - F_{T_j}(c_j(\alpha))$ denotes a local significance level for the j-th marginal test problem. In practice, it is convenient to carry out the multiple test procedure in terms of p-values $P_j = 1 - F_{T_j}(T_j)$ such that $\varphi_j = \mathbb{1}_{[0,\alpha_{loc}^{(j)})}(P_j)$.

Proof. The assertion follows from Assumption 2.7 (a) and Sklar's Theorem, since it holds that

$$\begin{aligned}\mathrm{FWER}_{\boldsymbol{\vartheta},C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) &\le \mathrm{FWER}_{\boldsymbol{\vartheta}^*,C_{\boldsymbol{X}}}(\boldsymbol{\varphi})\\ &= 1 - C_{\boldsymbol{T}}\left(F_{T_1}(c_1(\alpha)),\ldots,F_{T_m}(c_m(\alpha))\right)\\ &= 1 - C_{\boldsymbol{T}}\left(1-\alpha_{loc}^{(1)},\ldots,1-\alpha_{loc}^{(m)}\right).\end{aligned}$$

□

Lemma 2.10 shows that the problem of calibrating the local significance levels corresponding to $\boldsymbol{c}$ is equivalent to the problem of estimating the contour line of $C_{\boldsymbol{T}}$ at contour level $1-\alpha$. Any point on that contour line defines a valid set of local significance levels. Thus, one may weight the m hypotheses for importance by choosing particular points on the contour line. If all m hypotheses are equally important it is natural to choose equal local levels $\alpha_{loc}^{(j)} \equiv \alpha_{loc}$ for all $1 \le j \le m$. This amounts to finding the point of intersection of the contour line of $C_{\boldsymbol{T}}$ at contour level $1-\alpha$ and the "main diagonal" in the m-dimensional unit hypercube. Assumption 2.7 (c) is tailored towards this strategy and should be modified accordingly if a different weighting scheme is used.

Recall that we assume that C_X and, consequently, C_T are unknown. Based on our investigations in Section 2.2 and making use of Assumption 2.7 (c), we thus propose to calibrate $\boldsymbol{\varphi}$ empirically. If h is known, this can be done by solving the equation

$$h\left(B_K\left(\hat{C}_{X,n}\right)(1-\alpha_{loc},\ldots,1-\alpha_{loc})\right) = 1-\alpha \tag{2.5}$$

for α_{loc}. Note that this assumption is formulated for equally important hypotheses and has to be modified for different situations. If for a given α the solution of (2.5) is not unique, one should choose the smallest set of local significance levels such that (2.5) holds. We denote the solution of (2.5) by $\hat{\alpha}_{loc,n}$. This leads to the representation

$$\hat{\alpha}_{loc,n} = 1 - B_K\left(\hat{C}_{X,n}\right)^{\leftarrow}(h^{\leftarrow}(1-\alpha)),$$

where $B_K\left(\hat{C}_{X,n}\right)^{\leftarrow}$ is the quantile of $u \mapsto B_K\left(\hat{C}_{X,n}\right)(u,\ldots,u)$. Since $B_K\left(\hat{C}_{X,n}\right)$ depends on the data, $\hat{\alpha}_{loc,n}$ is a random variable and

$$\widehat{\mathrm{FWER}}_{\vartheta^*,C_X}(\boldsymbol{\varphi}) = 1 - C_T\left(1-\hat{\alpha}_{loc,n},\ldots,1-\hat{\alpha}_{loc,n}\right)$$

is a random variable, too, which is distributed around the target FWER level α. The following theorem is the main result of this section and quantifies the uncertainty about the realized FWER if the empirical calibration of $\boldsymbol{\varphi}$ is performed via (2.5).

Theorem 2.11. *Let Assumption 2.7 be fulfilled. Then the realized FWER has the following properties.*

a) Consistency:

$$\forall C_X \in \mathcal{C} : \widehat{FWER}_{\vartheta^*,C_X}(\boldsymbol{\varphi}) \to \alpha \text{ almost surely as } n \to \infty.$$

b) Asymptotic Normality:

$$\forall C_X \in \mathcal{C} : \sqrt{n}\left(\widehat{FWER}_{\vartheta^*,C_X}(\boldsymbol{\varphi}) - \alpha\right) \xrightarrow{d} \mathcal{N}(0,\sigma_\alpha^2) \text{ as } n \to \infty,$$

where

$$\sigma_\alpha^2 = \frac{\sigma^2\left(C_T(1-\alpha),\ldots,C_T(1-\alpha)\right)}{\left(C_X'\left(C_T(1-\alpha)\right)\right)^2} \cdot \left(C_T'\left(C_T(1-\alpha)\right)\right)^2,$$

$\sigma^2(\boldsymbol{u}) := \mathbb{V}[\mathbb{C}(\boldsymbol{u})]$, and C_X', C_T' denotes the first derivative of the univariate functions $u \mapsto C_X(u,\ldots,u)$, $u \mapsto C_T(u,\ldots,u)$, respectively.

c) Asymptotic Confidence Region:

$$\forall \delta \in (0,1) \forall C_{\boldsymbol{X}} \in \mathcal{C} : \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}} \left[\sqrt{n} \frac{\widehat{FWER}_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) - \alpha}{\hat{\sigma}_n} \leq z_{1-\delta} \right] = 1 - \delta,$$

where $\hat{\sigma}_n^2 : \mathcal{X}^n \to (0,\infty)$ is a consistent estimator of the asymptotic variance σ_α^2. In this, $z_\beta = \Phi^{-1}(\beta)$ denotes the β-quantile of the standard normal distribution on $\mathbb{R}$.

Proof.

a) Let $C_{\boldsymbol{X}} \in \mathcal{C}$ be arbitrary, but fixed. Since h is continuously differentiable, h is also Lipschitz-continuous with Lipschitz constant $L > 0$. Therefore, with Theorem 2.1 we get

$$\begin{aligned}
&\left| \widehat{\mathrm{FWER}}_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) - \alpha \right| \\
&= \left| 1 - \alpha - C_{\boldsymbol{T}} \left(1 - \hat{\alpha}_{loc,n}, \ldots, 1 - \hat{\alpha}_{loc,n} \right) \right| \\
&= \left| h\left(B_{\boldsymbol{K}} \left(\hat{C}_{\boldsymbol{X},n} \right) \left(1 - \hat{\alpha}_{loc,n}, \ldots, 1 - \hat{\alpha}_{loc,n} \right) \right) - h\left(C_{\boldsymbol{X}} \left(1 - \hat{\alpha}_{loc,n}, \ldots, 1 - \hat{\alpha}_{loc,n} \right) \right) \right| \\
&\leq \left\| h\left(B_{\boldsymbol{K}} \left(\hat{C}_{\boldsymbol{X},n} \right) \right) - h\left(C_{\boldsymbol{X}} \right) \right\|_\infty \\
&\leq L \cdot \left\| B_{\boldsymbol{K}} \left(\hat{C}_{\boldsymbol{X},n} \right) - C_{\boldsymbol{X}} \right\|_\infty \\
&= O\left(n^{-1/2} (\log\log n)^{1/2} \right) \text{ almost surely.}
\end{aligned}$$

b) Letting $p := h^{\leftarrow}(1-\alpha)$, Lemma 2.18 yields that

$$\begin{aligned}
\sqrt{n} \left(1 - \hat{\alpha}_{loc,n} - C_{\boldsymbol{X}}^{\leftarrow}(p) \right) &= \sqrt{n} \left(B_{\boldsymbol{K}} \left(\hat{C}_{\boldsymbol{X},n} \right)^{\leftarrow}(p) - C_{\boldsymbol{X}}^{\leftarrow}(p) \right) \\
&\xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2 \left(C_{\boldsymbol{X}}^{\leftarrow}(p), \ldots, C_{\boldsymbol{X}}^{\leftarrow}(p) \right)}{\left(C_{\boldsymbol{X}}' \left(C_{\boldsymbol{X}}^{\leftarrow}(p) \right) \right)^2} \right).
\end{aligned}$$

Therefore, applying the Delta Method to $u \mapsto C_{\boldsymbol{T}}(u, \ldots, u)$, we have that

$$\begin{aligned}
&\sqrt{n} \left(\widehat{\mathrm{FWER}}_{\boldsymbol{\vartheta}^*, C_{\boldsymbol{X}}}(\boldsymbol{\varphi}) - \alpha \right) \\
&= -\sqrt{n} \left(C_{\boldsymbol{T}} \left(1 - \hat{\alpha}_{loc,n}, \ldots, 1 - \hat{\alpha}_{loc,n} \right) - (1 - \alpha) \right) \\
&= -\sqrt{n} \left(C_{\boldsymbol{T}} \left(1 - \hat{\alpha}_{loc,n}, \ldots, 1 - \hat{\alpha}_{loc,n} \right) - C_{\boldsymbol{T}} \left(C_{\boldsymbol{X}}^{\leftarrow}(p), \ldots, C_{\boldsymbol{X}}^{\leftarrow}(p) \right) \right) \\
&\xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2 \left(C_{\boldsymbol{X}}^{\leftarrow}(p), \ldots, C_{\boldsymbol{X}}^{\leftarrow}(p) \right)}{\left(C_{\boldsymbol{X}}' \left(C_{\boldsymbol{X}}^{\leftarrow}(p) \right) \right)^2} \cdot \left(C_{\boldsymbol{T}}' \left(C_{\boldsymbol{X}}^{\leftarrow}(p) \right) \right)^2 \right).
\end{aligned}$$

The result follows from the definition of p.

c) Since $\hat{\sigma}_n \to \sigma_\alpha$ almost surely and particularly, in distribution for $n \to \infty$, the assertion follows directly from part b) using Slutsky's Theorem.

□

If the function h is unknown, one may approximate the value of $\hat{\alpha}_{loc,n}$ with high precision by a Monte Carlo simulation for a given number M of Monte Carlo repetitions. To this end, generate $M \times n$ pseudo-random vectors which follow the estimated (joint) distribution of $\boldsymbol{X}$ under $\boldsymbol{\vartheta}^*$, by combining $B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)$ and the marginal cdfs $F_{X_1}, \ldots, F_{X_m}$ of $X_1, \ldots, X_m$ under the global null hypothesis. From these, calculate a pseudo-sample $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_M$ from the distribution of $\boldsymbol{T}$ under $\boldsymbol{\vartheta}^*$. Then, $\boldsymbol{F}_{\boldsymbol{T}_1}(\boldsymbol{T}_1), \ldots, \boldsymbol{F}_{\boldsymbol{T}_M}(\boldsymbol{T}_M)$ constitutes a pseudo-random sample from the estimator of $C_{\boldsymbol{T}}$, and the empirical equi-coordinate $(1-\alpha)$-quantile of this pseudo-sample approximates $\hat{\alpha}_{loc,n}$. Since the number M of pseudo-random vectors to be generated is in principle unlimited, Theorem 2.11 continues to hold true if this strategy is pursued. We will make use of this approach in the more involved examples studied in Section 2.4 and Section 2.5.

2.4 Simulation study

In this section we report the results of a simulation study regarding the FWER and the power of multiple tests which are empirically calibrated as proposed in Section 2.3. Assume w.l.o.g. that $I_0(\boldsymbol{\vartheta}) := \{1, \ldots, m_0\}$ and let $m_1 := m - m_0$. The empirical FWER is given by the relative frequency over the L simulation runs of the occurrence of at least one false rejection, i.e.,

$$\mathrm{eFWER}(\boldsymbol{\varphi}) := L^{-1} \sum_{\ell=1}^{L} \mathbb{1}_{\bigcup_{j=1}^{m_0} \left\{\varphi_j^{(\ell)}=1\right\}} \left(\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_n^{(\ell)}\right).$$

Likewise, the empirical power is defined as the average proportion of true rejections, i.e.,

$$\mathrm{ePower}(\boldsymbol{\varphi}) := L^{-1} \sum_{\ell=1}^{L} \left(m_1^{-1} \sum_{j=m_0+1}^{m} \mathbb{1}_{\left\{\varphi_j^{(\ell)}=1\right\}} \left(\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_n^{(\ell)}\right) \right),$$

where $\left(\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_n^{(\ell)}\right) \in \mathcal{X}^n$ denotes the pseudo-sample in the ℓ-th simulation run.

The setting is as follows. We simulate from various one-parametric copula models (namely, Frank, Clayton, Gumbel, Student's t with four degrees of freedom, and the product copula) with parameters corresponding to weak (Kendall's $\tau \approx 0.25$) and strong dependence (Kendall's $\tau \approx 0.75$), respectively. In the case of t_4-copulas we restrict our attention to the case of equi-correlation, and the parameter is the equi-correlation coefficient. For convenience (and without loss of generality), the data are marginally normally

distributed with all marginal variances equal to one. In the inference procedures, however, we assume these variances to be unknown, leading to Studentized test statistics. For each $1 \leq j \leq m$, we let ϑ_j be the mean in coordinate j. In all simulation settings, ϑ_j is set to 0.4 under alternatives. The null hypotheses are given by $H_j : \{\vartheta_j = \vartheta_j^* = 0\}$, with two-sided alternatives. Hence, marginal two-sided t-tests are performed with multiplicity corrected local significance level. Our Bernstein procedure is compared with the widely used Bonferroni and Šidák methods.

Notice that Assumption 2.7 is fulfilled. From Lemma 2.8 we get that the LFC is indeed $\boldsymbol{\vartheta}^* = (0, \ldots, 0)^\top$. Further, the marginal distribution functions of the test statistics are known (even for finite n) and the function h exists, since $u \mapsto C_X(u, \ldots, u)$ is strictly increasing for the choices of C_X in this simulation study. However, the function h is unknown in contrast to the examples in Section 2.3.

The calculation of the Bernstein copula has been performed as in Example 4.2 of Cottin and Pfeifer (2014), which uses $K_j := n$ for all $j \in \{1, \ldots, m\}$. This choice fulfills the assumption of Theorem 2.1. In order to meet the assumptions of Theorem 2.4 it would be necessary to choose K_j slightly larger. Notice, however, that we consider small sample sizes $n \in \{20, 100\}$ in our simulations, such that asymptotic considerations do not apply here. Instead, some preliminary simulations indicated that the choice $K_j \equiv n$ is appropriate. The choice of n was motivated by the purpose to demonstrate how accurately the Bernstein estimator performs in a small sample scenario. For instance, the real data example that we will present in Section 2.5 has a sample size of $n = 20$. With the simulations presented here, we can thus evaluate the appropriateness of the application of the proposed methodology in this real data example.

Since the function h is assumed unknown here, we calibrate the proposed multiple test with the following algorithm which was outlined at the end of Section 2.3.

Algorithm 2.12.

1. *Choose a number M of Monte Carlo repetitions.*
2. *For each $b = 1, \ldots, M$ draw a sample $\boldsymbol{U}_1^{\#b}, \ldots, \boldsymbol{U}_n^{\#b}$ of $B_{\boldsymbol{K}}\left(\hat{C}_{X,n}\right)$ and calculate*

$$X_{i,j}^{\#b} = \hat{\sigma}_j \cdot \Phi_j^{-1}\left(U_{i,j}^{\#b}\right) + \vartheta_j^*, \qquad 1 \leq i \leq n, 1 \leq j \leq m,$$

   where $\hat{\sigma}_j$ is the sample standard deviation of $X_{1,j}, \ldots, X_{n,j}$.

3. *For all $1 \leq j \leq m$, compute*

$$T_j^{\#b} = T_j\left(\boldsymbol{X}_1^{\#b}, \ldots, \boldsymbol{X}_n^{\#b}\right) = \left| \sqrt{n} \cdot \frac{\frac{1}{n}\sum_{i=1}^n X_{i,j}^{\#b} - \vartheta_j^*}{\hat{\sigma}_j^{\#b}} \right|$$

and obtain the pseudo-sample

$$V_j^{\#b} = 2F_{t_{n-1}}\left(T_j^{\#b}\right) - 1$$

from the copula of $\boldsymbol{T}$.

4. *Finally, calibrate* $\hat{\boldsymbol{\alpha}}_{loc,n} = \left(\hat{\alpha}_{loc,n}^{(1)}, \ldots, \hat{\alpha}_{loc,n}^{(m)}\right)^\top$ *by solving*

$$\#\left\{b \middle| V_j^{\#b} \leq 1 - \hat{\alpha}_{loc,n}^{(j)} \text{ for all } 1 \leq j \leq m\right\} = \lceil (1-\alpha) M \rceil . \tag{2.6}$$

Notice that in (2.6), we implicitly weight the hypotheses. This means that the weights corresponding to the obtained $\hat{\boldsymbol{\alpha}}_{loc,n}$ depend on the simulation data, for convenience of implementation. In comparison, the classical Bonferroni and Šidák corrected local significance levels are given by

$$\alpha_{loc}^{(j)} = \frac{\alpha}{m} \text{ and } \alpha_{loc}^{(j)} = 1 - (1-\alpha)^{1/m}, \qquad 1 \leq j \leq m,$$

respectively.

The results are displayed in Table 1 (weak dependence with Kendall's $\tau \approx 0.25$) and Table 2 (strong dependence with Kendall's $\tau \approx 0.75$). They reveal that in this simulation study the Bernstein method performs best in the case that M is large and the proportion of true null hypotheses π_0 is not too large, i.e., in these cases its empirical FWER is closer to α and its empirical power is higher than those of the generic calibrations. Under strong dependence the power of the Bernstein method increases even further. On the other hand, if all hypotheses are true then the empirical FWER for the Bernstein method can be above $\alpha = 5\%$ and M needs to be large in order to improve the empirical FWER. Surprisingly, the sample size n does not have a clear positive impact in this simulation study.

2.5 Real data analysis

In this section, we analyze insurance claim data from $m = 19$ adjacent geographical regions (see Table 5). For every region $j \in \{1, \ldots, 19\}$ these claims have, for confidentiality reasons, been adjusted to a neutral monetary scale. The claim amounts and types have been aggregated to full years, such that temporal dependencies are considered negligible. However, strong non-linear spatial dependencies are likely to be present in the data. Hence, we treat each of the $n = 20$ rows in Table 5 as an independent repetition $\boldsymbol{X}_i = \boldsymbol{x}_i$ of an m-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$, where $1 \leq i \leq 20$ is the time index in years and $m = 19$ refers to the regions.

Table 1: Comparison of empirical FWER and power regarding Bonferroni, Šidák and Bernstein corrections under various weak dependency structures with $m = 20$, $\pi_0 = m_0/m \in \{0.5, 1\}$, $\alpha = 0.05$, $L = 1000$, $M \in \{200, 1000\}$, and $n \in \{20, 100\}$.

Family	π_0	M	n	eFWER			ePower		
				Bonferroni	Šidák	Bernstein	Bonferroni	Šidák	Bernstein
Frank(2)	0.5	200	20	1.8%	1.8%	6.1%	7.5%	7.7%	14.7%
			100	1.6%	1.8%	6.9%	81.4%	81.7%	86.1%
		1000	20	2.6%	2.6%	4.2%	7.8%	8.0%	10.6%
			100	2.6%	2.6%	4.0%	82.0%	82.2%	84.6%
	1	200	20	5.2%	5.2%	14.6%			
			100	3.0%	3.1%	13.1%			
		1000	20	5.5%	5.5%	7.8%			
			100	4.6%	4.9%	6.6%			
Gumbel(2)	0.5	200	20	2.5%	2.5%	6.4%	8.4%	8.6%	19.4%
			100	1.3%	1.5%	6.0%	80.5%	80.7%	89.9%
		1000	20	1.3%	1.3%	3.8%	7.1%	7.2%	12.7%
			100	2.1%	2.2%	4.8%	80.6%	80.9%	88.1%
	1	200	20	1.9%	1.9%	9.8%			
			100	2.6%	2.6%	10.0%			
		1000	20	2.7%	2.7%	5.3%			
			100	2.2%	2.2%	6.4%			
Clayton(1)	0.5	200	20	2.2%	2.2%	7.0%	7.0%	7.1%	14.3%
			100	2.1%	2.1%	6.0%	81.3%	81.5%	88.0%
		1000	20	2.4%	2.4%	4.3%	7.0%	7.1%	9.6%
			100	1.8%	1.8%	3.9%	81.3%	81.5%	86.4%
	1	200	20	3.3%	3.4%	12.6%			
			100	4.6%	4.6%	14.7%			
		1000	20	3.5%	3.7%	5.6%			
			100	3.6%	3.7%	7.4%			
$t_4(0.4)$	0.5	200	20	2.8%	2.8%	7.6%	6.7%	6.8%	13.4%
			100	2.0%	2.1%	8.0%	81.7%	82.0%	87.4%
		1000	20	2.3%	2.3%	3.6%	7.3%	7.5%	10.2%
			100	2.9%	3.0%	4.0%	81.3%	81.5%	85.0%
	1	200	20	5.1%	5.1%	15.0%			
			100	4.1%	4.1%	12.8%			
		1000	20	4.4%	4.5%	7.6%			
			100	3.3%	3.3%	6.9%			
Independence	0.5	200	20	2.5%	2.6%	8.2%	7.4%	7.6%	13.5%
			100	3.4%	3.4%	7.6%	81.8%	81.9%	86.0%
		1000	20	2.9%	2.9%	3.9%	7.0%	7.2%	8.8%
			100	2.1%	2.2%	3.6%	81.4%	81.6%	82.8%
	1	200	20	5.3%	5.3%	14.3%			
			100	5.7%	5.8%	15.5%			
		1000	20	4.0%	4.1%	6.9%			
			100	4.2%	4.2%	7.6%			

Table 2: Comparison of empirical FWER and power regarding Bonferroni, Šidák and Bernstein corrections under various strong dependency structures with $m = 20$, $\pi_0 = m_0/m \in \{0.5, 1\}$, $\alpha = 0.05$, $L = 1000$, $M \in \{200, 1000\}$, and $n \in \{20, 100\}$.

Family	π_0	M	n	eFWER			ePower		
				Bonferroni	Šidák	Bernstein	Bonferroni	Šidák	Bernstein
Frank (14)	0.5	200	20	0.8%	0.8%	6.8%	8.1%	8.2%	22.5%
			100	0.6%	0.6%	7.0%	81.7%	81.9%	94.4%
		1000	20	1.0%	1.0%	3.2%	7.9%	8.0%	18.3%
			100	0.9%	0.9%	4.1%	81.2%	81.4%	92.3%
	1	200	20	0.9%	1.0%	7.5%			
			100	1.0%	1.0%	8.8%			
		1000	20	1.4%	1.4%	5.2%			
			100	1.1%	1.1%	5.3%			
Gumbel (4)	0.5	200	20	1.5%	1.6%	7.1%	7.7%	7.8%	23.3%
			100	0.6%	0.6%	6.2%	81.6%	81.8%	94.9%
		1000	20	0.5%	0.5%	2.2%	7.6%	7.7%	18.1%
			100	1.1%	1.1%	4.3%	80.9%	81.1%	93.6%
	1	200	20	1.3%	1.3%	6.2%			
			100	0.9%	0.9%	7.9%			
		1000	20	1.5%	1.5%	4.1%			
			100	1.4%	1.4%	6.3%			
Clayton (6)	0.5	200	20	0.9%	0.9%	4.8%	7.2%	7.3%	22.0%
			100	1.2%	1.3%	7.6%	81.3%	81.5%	94.9%
		1000	20	0.8%	0.8%	3.5%	7.0%	7.1%	15.9%
			100	0.9%	0.9%	4.2%	80.8%	81.0%	93.0%
	1	200	20	1.3%	1.5%	5.9%			
			100	1.3%	1.3%	8.7%			
		1000	20	1.4%	1.4%	4.1%			
			100	1.0%	1.0%	5.0%			
t_4 (0.9)	0.5	200	20	1.6%	1.6%	6.9%	8.3%	8.4%	22.2%
			100	0.7%	0.8%	6.8%	80.9%	81.1%	94.3%
		1000	20	1.0%	1.0%	2.3%	7.4%	7.6%	16.1%
			100	1.0%	1.0%	4.8%	81.4%	81.5%	93.0%
	1	200	20	1.8%	1.8%	7.8%			
			100	0.9%	1.0%	9.1%			
		1000	20	1.5%	1.6%	4.1%			
			100	1.4%	1.4%	5.7%			

An important quantity for regulators and risk managers is the region-specific value-at-risk (VaR). The VaR at level p for region j is defined as the p-quantile of the (marginal) distribution of X_j, i.e.,

$$\mathrm{VaR}_j(p) := F_{X_j}^{\leftarrow}(p).$$

In insurance mathematics, typically considered values of p are close to one. Here, we chose $p = 0.995$. Our goal is to derive multiplicity-corrected confidence intervals for $\vartheta_j = \mathrm{VaR}_j(0.995)$, $1 \leq j \leq m = 19$ which are compatible with (i.e., dual to) the Bonferroni, Šidák and Bernstein copula-based correction methods discussed before. To this end, let auxiliary point hypotheses be defined as $H_{\vartheta_j^*} : \left\{\vartheta_j = \vartheta_j^*\right\}$ for fixed $\vartheta_j^* > 0$. According to the Extended Correspondence Theorem (see Section 1.3 of Dickhaus (2014)), the set of all values ϑ_j^* for which $H_{\vartheta_j^*}$ is retained by a multiple test at FWER level α (leading to a local significance level $\alpha_{loc}^{(j)}$ in coordinate j) constitutes a confidence region at simultaneous confidence level $1-\alpha$ for ϑ_j, $1 \leq j \leq m$. We set $\alpha = 5\%$.

These model assumptions are analogous to those from the examples in the previous sections. It can be shown (cf. our argumentation in Example 2.9 (a)) that Assumption 2.7 (a) and (b) are fulfilled. On the other hand, it is difficult to check Assumption 2.7 (c) in many applications. For example, in the simulation study reported in Section 2.4 we used the fact that the data were simulated under some suitable copula families.

In quantitative risk management, it is common practice to model the excess distribution of X_j over some given threshold u_j by a generalized Pareto distribution (GPD) (cf., e.g., Section 7.2.2 of McNeil et al. (2005)).

Definition 2.13 (Definition 7.16 of McNeil et al. (2005))**.** For shape parameter $\xi \in \mathbb{R}$ and scale parameter $\beta > 0$, the cdf of the GPD is given by

$$G_{\xi,\beta}(x) = \begin{cases} 1-(1+\xi x/\beta)^{-1/\xi} & ,\xi \neq 0, \\ 1-\exp(-x/\beta) & ,\xi = 0, \end{cases}$$

where $x \geq 0$ if $\xi \geq 0$ and $0 \leq x \leq -\beta/\xi$ if $\xi < 0$.

In the remainder, we make the following assumption.

Assumption 2.14. *For every $1 \leq j \leq m = 19$ there exists a threshold u_j and parameter values ξ_j and β_j such that*

$$\mathbb{P}\left[X_j - u_j \leq x \middle| X_j > u_j\right] \approx G_{\xi_j,\beta_j}(x)$$

for all $x \geq 0$.

Under Assumption 2.14, an approximation of the VaR at level p for region j is given by

$$\mathrm{VaR}_{\xi_j,\beta_j}(p) \approx u_j + \frac{\beta_j}{\xi_j}\left(\left(\frac{1-p}{1-F_{X_j}(u_j)}\right)^{-\xi_j} - 1\right) =: q_j\left(\xi_j,\beta_j\right),$$

provided that $p \geq F_{X_j}(u_j)$. For ease of notation, we let $\vartheta_j = q_j\left(\xi_j,\beta_j\right)$ in the sequel.

For computational convenience, we carried out the test for $H_{\vartheta_j^*}$ as a confidence-region test in the sense of Aitchison (1964) based on the family

$$\left(H_{\xi_j^*,\beta_j^*} : \left\{\xi_j = \xi_j^*, \beta_j = \beta_j^*\right\} \middle| \beta_j^* > 0, \xi_j^* \in \mathbb{R}\right) \tag{2.7}$$

of point hypotheses. Namely, the test procedure works as follows.

Algorithm 2.15.

1. *Test each $H_{\xi_j^*,\beta_j^*}$ by an arbitrary level $\alpha_{loc}^{(j)}$ test, where $\alpha_{loc}^{(j)}$ denotes a multiplicity-corrected significance level based on the Bonferroni, Šidák or Bernstein copula calibration, respectively.*

2. *Let a confidence region $C_{\xi_j,\beta_j}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ at confidence level $1-\alpha_{loc}^{(j)}$ for (ξ_j,β_j) be defined as the set of all parameter values (ξ_j^*,β_j^*) for which $H_{\xi_j^*,\beta_j^*}$ is retained.*

3. *Reject $H_{\vartheta_j^*}$ at level $\alpha_{loc}^{(j)}$, if the set $\{(\xi_j^*,\beta_j^*) : q_j\left(\xi_j^*,\beta_j^*\right) = \vartheta_j^*\}$ has an empty intersection with $C_{\xi_j,\beta_j}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$.*

Due to Algorithm 2.15, it suffices to construct point hypothesis tests for (2.7). A standard technique for testing parametric hypotheses is to perform a likelihood ratio test. In the risk management context, this method is described in Appendix A.3.5 of McNeil et al. (2005). Define the random variable $N_{u_j} := \#\left\{1 \leq i \leq n \middle| X_{i,j} > u_j\right\}$ and let $\tilde{X}_{1,j},\ldots,\tilde{X}_{N_{u_j},j}$ denote the corresponding sub-sample for region j. Then the excesses $Y_{1,j},\ldots,Y_{N_{u_j},j}$ over u_j are defined by

$$Y_{i,j} := \tilde{X}_{i,j} - u_j.$$

The test statistic for testing $H_{\xi_j^*,\beta_j^*}$ is then given by

$$T_j\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi_j^*,\beta_j^*\right) := -2\log\Lambda\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi_j^*,\beta_j^*\right),$$

where the likelihood ratio Λ is defined by

$$\Lambda\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi_j^*,\beta_j^*\right) := \frac{L\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi_j^*,\beta_j^*\right)}{\sup_{(\xi,\beta)} L\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi,\beta\right)}$$

with log-likelihood function

$$\log L\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi,\beta\right) = -N_{u_j}\log\beta - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{N_{u_j}}\log\left(1+\xi\frac{Y_{i,j}}{\beta}\right).$$

Under $H_{\xi_j^*,\beta_j^*}$, T_j is asymptotically χ^2-distributed with two degrees-of-freedom. This means that the (asymptotic) confidence interval $C_{\xi_j,\beta_j}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ in the second step of Algorithm 2.15 is given by

$$C_{\xi_j,\beta_j}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) = \left\{\left(\xi_j^*,\beta_j^*\right) : T_j\left(Y_{1,j},\ldots,Y_{N_{u_j},j};\xi_j^*,\beta_j^*\right) \leq F_{\chi_2^2}^{-1}\left(1-\alpha_{loc}^{(j)}\right)\right\}. \tag{2.8}$$

Utilizing (2.8), the confidence region $[\vartheta_j^{\text{lower}},\vartheta_j^{\text{upper}}]$ for ϑ_j based on the third step of Algorithm 2.15 is constructed by finding the minimum value $\vartheta_j^{\text{lower}} = \min q_j(\xi_j^*,\beta_j^*)$ and the maximum value $\vartheta_j^{\text{upper}} = \max q_j(\xi_j^*,\beta_j^*)$, where (ξ_j^*,β_j^*) are located on the boundary of $C_{\xi_j,\beta_j}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$.

A graphical method for the determination of a suitable threshold u_j is based on the mean excess plot in coordinate j (cf. Section 7.2.2 of McNeil et al. (2005) for details). Namely, all possible values u of u_j are plotted against the mean of the values of $Y_{1,j},\ldots,Y_{N_u,j}$. If the GPD model is appropriate, the plot should yield an approximately linear graph for arguments exceeding u_j. Usually the few largest values of u are ignored, because they lead to very small values of N_u.

For example, Figure 2 shows the mean excess plots for the two regions 2 and 4. The mean excess plot for region 2 is approximately linear when ignoring the three smallest and the four largest values of u. This means that a suitable threshold u_2 would be between 18.815 and 28.316. Similarly, the mean excess plot for region 4 is approximately linear when ignoring the two largest values of u, hence $u_4 < 0.321$. Based on such considerations, we chose the thresholds $\boldsymbol{u} = (u_1,\ldots,u_{19})^\top$ given by

$$\boldsymbol{u} := (1.0, 28.0, 9.0, 0.3, 0.2, 0.4, 2.6, 1.2, 0.4, 1.1, 0.1, 0.2, 22.5, 1.6, 3.2, 0.2, 12.5, 1.2, 0.5)^\top.$$

Finally, it remains to determine the local significance levels $\left(\alpha_{loc}^{(j)}\right)_{1\leq j\leq 19}$. In the case of the Bonferroni or the Šidák method, this is trivial. To calibrate the local significance levels with the Bernstein method, we employed a modified version of Algorithm 2.12 based on the empirical excess distribution. Algorithm 2.16 yields a resampling-based approximation of the copula of the vector $\boldsymbol{T} = (T_1,\ldots,T_m)^\top$ of the region-specific likelihood ratio test statistics.

Algorithm 2.16.

1. *For every* $1 \leq j \leq m$, *estimate the parameters* ξ_j *and* β_j *of the excess distribution of*

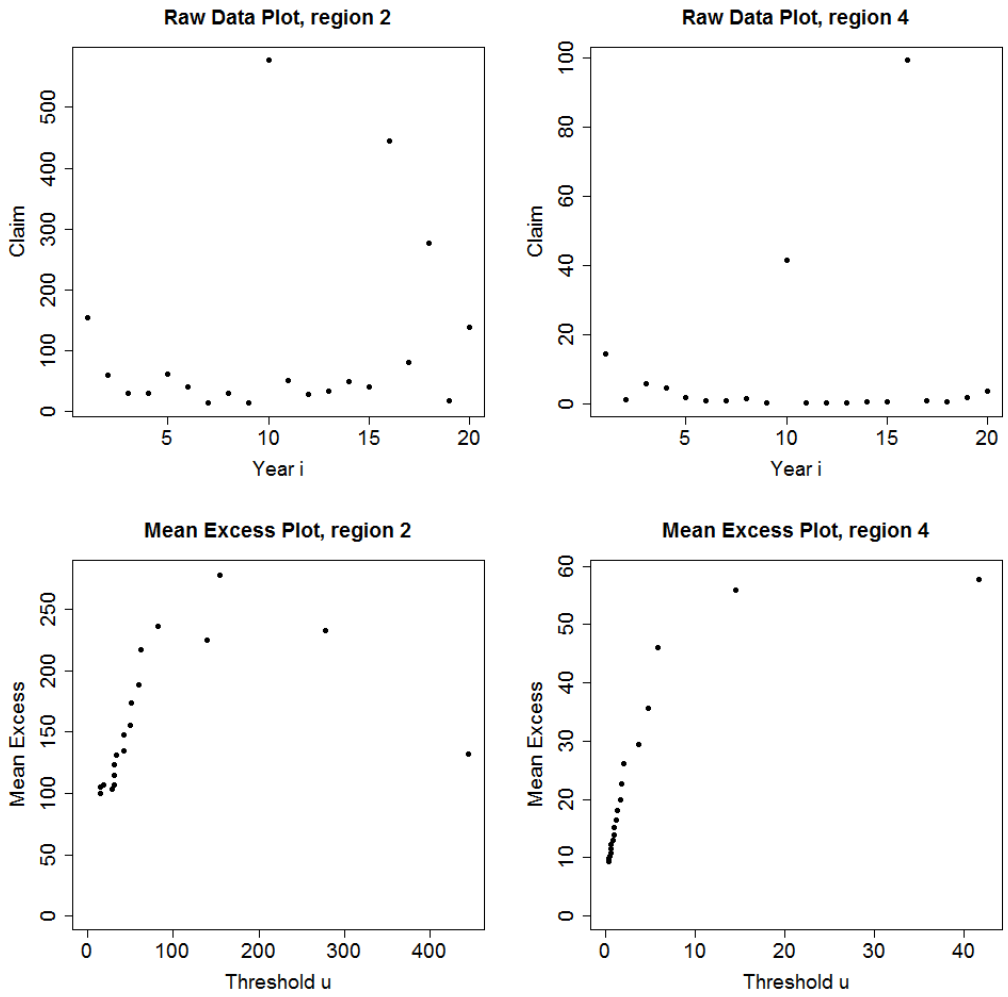

Figure 2: Raw data and mean excess plots for regions 2 and 4. The graphs in the upper panel display the data from Table 5 for $j \in \{2,4\}$, respectively. The graphs in the lower panel show the corresponding mean excess plots.

Table 3: Estimated parameters $\hat{\xi}_j$, $\hat{\beta}_j$, $1 \le j \le 19$, for the region-specific GPD models. Estimation has been performed via maximum likelihood.

$\hat{\xi}_j$	0.41	1.17	0.75	1.43	0.87	1.51	1.10	0.30	0.49	0.79
	0.56	0.98	1.00	0.73	0.47	0.81	1.08	0.60	0.89	
$\hat{\beta}_j$	19.59	22.21	18.41	0.82	1.10	1.56	4.57	9.75	2.91	6.46
	0.64	0.99	5.12	3.42	20.34	4.52	6.98	1.96	1.64	

X_j via maximum likelihood and calculate N_{u_j}.

2. *Choose a number M of Monte Carlo repetitions.*

3. *For each $1 \le b \le M$ draw a pseudo sample $\boldsymbol{U}_1^{\#b}, \ldots, \boldsymbol{U}_n^{\#b}$ from the (empirical) Bernstein copula $B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)$ and calculate the corresponding GPD excesses*

$$Y_{i,j}^{\#b} = G^{\leftarrow}_{\hat{\xi}_j,\hat{\beta}_j}\left(U_{(i),j}^{\#b}\right), 1 \le i \le N_{u_j}, 1 \le j \le m,$$

where $U_{(i),j}^{\#b}$ denotes the i-th reverse order statistic of $\left(U_{i,j}^{\#b}\right)_{1\le i\le n}$.

4. *For each $1 \le j \le m$, compute $T_j^{\#b} = T_j\left(Y_{1,j}^{\#b}, \ldots, Y_{N_{u_j},j}^{\#b}; \hat{\xi}_j, \hat{\beta}_j\right)$, and obtain the pseudo-sample*

$$V_j^{\#b} = \hat{G}_{j,M}\left(T_j^{\#b}\right), 1 \le j \le m$$

from the copula of $\boldsymbol{T}$.

5. *Finally, calibrate $\hat{\boldsymbol{\alpha}}_{loc,n} = \left(\hat{\alpha}_{loc,n}^{(1)}, \ldots, \hat{\alpha}_{loc,n}^{(m)}\right)^{\top}$ by solving*

$$\#\left\{b \middle| V_j^{\#b} \le 1 - \hat{\alpha}_{loc,n}^{(j)} \text{ for all } 1 \le j \le m\right\} = \lceil (1-\alpha) M \rceil .$$

Table 3 displays the parameter estimates for the region-specific GPD models, and Table 4 displays the lower bounds $\left(\vartheta_j^{\text{lower}}\right)_{1\le j\le m}$ of the region-specific confidence intervals for the 99.5% VaR obtained by the Bonferroni, Šidák and Bernstein copula calibration, respectively.

Similarly as in Algorithm 2.12, an implicit weighting has been employed for the determination of the local significance levels $\left(\alpha_{loc}^{(j)}\right)_{1\le j\le m}$ in Algorithm 2.16. Therefore, the confidence bounds obtained with the Bernstein copula method are not guaranteed to be more informative (i.e., larger) than the ones obtained by the Bonferroni or the Šidák methods for all regions. However, we observe improvements in almost all regions j. It is remarkable that this expected behavior of the Bernstein copula calibration can already be

Table 4: Lower confidence bounds $\vartheta_j^{\text{lower}}$ for the 99.5% VaR, $1 \leq j \leq 19$, obtained by the Bonferroni, the Šidák and the Bernstein copula method, respectively. The results for the Bernstein method rely on $M = 1{,}000$ Monte Carlo repetitions in Algorithm 2.16.

Bonferroni	89.08	283.30	126.20	19.41	10.00	36.68	62.57	39.45	14.62	51.14
	3.74	10.13	53.74	25.43	101.62	37.11	84.99	12.79	14.80	
Šidák	89.22	284.03	126.46	19.48	10.03	36.81	62.75	39.51	14.64	51.25
	3.75	10.15	53.82	25.47	101.78	37.20	85.20	12.81	14.83	
Bernstein	91.59	287.32	127.61	19.81	10.13	37.37	63.54	38.82	14.73	51.74
	3.78	10.25	52.89	26.27	99.90	37.58	82.71	12.91	14.98	

verified for the rather moderate sample size of $n = 20$, because the likelihood ratio tests and the Bernstein copula calibration are both based on asymptotic considerations.

We omitted the values of $\left(\vartheta_j^{\text{upper}}\right)_{1\leq j\leq m}$, because they are uninformative (extremely large). This is in line with the fact that all scale parameter estimates $\hat{\xi}_j$ in Table 3 are positive. For $\xi \geq 0$, the GPD has infinite support, thus the modeled 99.5% VaR tends to be very large.

2.6 Discussion

We have derived a non-parametric approach to the calibration of multiple test procedures which take the joint distribution of test statistics into account. In contrast to previous approaches which were restricted to cases with low-dimensional copula parameters, the Bernstein copula-based approximation of the local significance levels proposed in the present work can be applied under almost no assumptions regarding the dependency structures among test statistics or p-values, respectively. This makes the proposed methodology an attractive choice for data the dependency structure of which has not been explicitly modeled prior to the statistical analysis. Furthermore, our empirical results on simulated as well as on real-life data indicate the gain in power which is possible by the consideration of the dependency structure among test statistics in the calibration of the multiple test. This is particularly important for modern applications with high dimensionality of, but also pronounced dependencies in the data.

On the other hand, Theorem 2.11 provides a precise asymptotic performance guarantee for the empirically calibrated multiple test, meaning that a sharp upper bound for its realized FWER can be obtained, at least asymptotically for large sample sizes. This is in contrast to most of the existing resampling-based multiple test procedures like the 'max T' and 'min P' tests proposed by Westfall and Young (1993), which are obvious competitors of our approach.

Future work shall explore the case that some qualitative assumptions regarding the

Table 5: Insurance claim data from 19 adjacent geographical regions over 20 years.

Raw data $x_{i,j}$		region j							
		1	2	3	4	5	6	7	8
	1	23.664	154.664	40.569	14.504	10.468	7.464	22.202	17.682
	2	1.080	59.545	3.297	1.344	1.859	0.477	6.107	7.196
	3	21.731	31.049	55.973	5.816	14.869	20.771	3.580	14.509
	4	28.990	31.052	30.328	4.709	0.717	3.530	6.032	6.512
	5	53.616	62.027	57.639	1.804	2.073	4.361	46.018	22.612
	6	29.950	41.722	12.964	1.127	1.063	4.873	6.571	11.966
	7	3.474	14.429	10.869	0.945	2.198	1.484	4.547	2.556
	8	10.020	31.283	21.116	1.663	2.153	0.932	25.163	3.222
	9	5.816	14.804	128.072	0.523	0.324	0.477	3.049	7.791
year i	10	170.725	576.767	108.361	41.599	20.253	35.412	126.698	71.079
	11	21.423	50.595	4.360	0.327	1.566	64.621	5.650	1.258
	12	6.380	28.316	3.740	0.442	0.736	0.470	3.406	7.859
	13	124.665	33.359	14.712	0.321	0.975	2.005	3.981	4.769
	14	20.165	49.948	17.658	0.595	0.548	29.350	6.782	4.873
	15	78.106	41.681	13.753	0.585	0.259	0.765	7.013	9.426
	16	11.067	444.712	365.351	99.366	8.856	28.654	10.589	13.621
	17	6.704	81.895	14.266	0.972	0.519	0.644	8.057	18.071
	18	15.550	277.643	26.564	0.788	0.225	1.230	26.800	64.538
	19	10.099	18.815	9.352	2.051	1.089	6.102	2.678	4.064
	20	8.492	138.708	46.708	3.680	1.132	1.698	165.600	7.926

9	10	11	12	13	14	15	16	17	18	19
12.395	18.551	1.842	4.100	46.135	14.698	44.441	7.981	35.833	10.689	7.299
1.436	3.720	0.429	1.026	7.469	7.058	4.512	0.762	14.474	9.337	0.740
17.175	87.307	0.209	2.344	22.651	4.117	26.586	3.920	13.804	2.683	3.026
0.682	3.115	0.521	0.696	31.126	1.878	29.423	6.394	18.064	1.201	0.894
1.581	11.179	2.715	1.327	40.156	4.655	104.691	28.579	17.832	1.618	3.402
15.676	24.263	4.832	0.701	16.712	11.852	29.234	7.098	17.866	5.206	5.664
0.456	1.137	0.268	0.580	11.851	2.057	11.605	0.282	16.925	2.082	1.008
1.581	5.477	0.741	0.369	3.814	1.869	8.126	1.032	14.985	1.390	1.703
4.079	7.002	0.524	6.554	5.459	3.007	8.528	1.920	5.638	2.149	2.908
21.762	64.582	9.882	6.401	106.197	44.912	191.809	90.559	154.492	36.626	36.276
0.626	3.556	1.052	8.277	22.564	8.961	19.817	16.437	25.990	2.364	6.434
0.894	3.591	0.136	0.364	28.000	7.574	3.213	1.749	12.735	1.744	0.558
2.006	1.973	1.990	15.176	57.235	23.686	110.035	17.373	7.276	2.494	0.525
2.921	6.394	0.630	0.762	25.897	3.439	8.161	3.327	24.733	2.807	1.618
2.180	3.769	0.770	15.024	36.068	1.613	6.127	8.103	12.596	4.894	0.822
9.589	19.485	0.287	0.464	24.211	38.616	51.889	1.316	173.080	3.557	11.627
5.515	13.163	0.590	2.745	16.124	2.398	20.997	2.515	5.161	2.840	3.002
2.637	80.711	0.245	0.217	12.416	4.972	59.417	3.762	24.603	7.404	19.107
2.373	2.057	0.415	0.351	10.707	2.468	10.673	1.743	27.266	1.368	0.644
2.972	5.237	0.566	0.708	22.646	6.652	14.437	63.692	113.231	7.218	2.548

dependency structure are at hand. For example, it will be interesting to quantify the uncertainty of the FWER of a multiple test procedure which is calibrated by assuming an Archimedean p-value copula as in Bodnar and Dickhaus (2014). In this case, non-parametric estimation of the copula generator function as for instance proposed by Lambert (2007) will lead to an empirical calibration of the multiple test.

2.7 Auxiliary results

In this section two auxiliary lemmas are formulated and proved. The first lemma is used in the proofs of Theorem 2.1 and Theorem 2.4. The second lemma follows from Theorem 2.4 and is used in Theorem 2.11.

Lemma 2.17. *It holds that*

$$\|B_{\boldsymbol{K}}(C_{\boldsymbol{X}}) - C_{\boldsymbol{X}}\|_\infty \leq \frac{1}{2}\sum_{j=1}^{m} K_j^{-1/2},$$

where $\|g\|_\infty := \sup_{\boldsymbol{u}\in[0,1]^m} |g(\boldsymbol{u})|$ *for* $g : [0,1]^m \longrightarrow \mathbb{R}$.

Proof. We get

$$\begin{aligned}
\|B_{\boldsymbol{K}}(C_{\boldsymbol{X}}) - C_{\boldsymbol{X}}\|_\infty &\leq \sup_{\boldsymbol{u}\in[0,1]^m} \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{K}} |C_{\boldsymbol{X}}(\boldsymbol{k}/\boldsymbol{K}) - C_{\boldsymbol{X}}(\boldsymbol{u})| \prod_{j=1}^{m} P_{k_j,K_j}(u_j) \\
&\leq \sup_{\boldsymbol{u}\in[0,1]^m} \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{K}} \sum_{j_1=1}^{m} \left| \frac{k_{j_1}}{K_{j_1}} - u_{j_1} \right| \cdot \prod_{j_2=1}^{m} P_{k_{j_2},K_{j_2}}(u_{j_2}) \\
&\leq \frac{1}{2}\sum_{j=1}^{m} K_j^{-1/2},
\end{aligned}$$

where the second inequality follows from the Lipschitz property of multivariate copula (see Section 2 of Sancetta and Satchell (2004)). For the last inequality we use the fact that $P_{k_j,K_j}(u_j)$ is the probability function of the binomial distribution for each $u_j \in [0,1]$ and

$1 \leq j \leq m$. Therefore, by the Jensen inequality it follows that

$$\begin{aligned}
&\sup_{\boldsymbol{u}\in[0,1]^m} \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{K}} \sum_{j_1=1}^{m} \left| \frac{k_{j_1}}{K_{j_1}} - u_{j_1} \right| \cdot \prod_{j_2=1}^{m} P_{k_{j_2},K_{j_2}}\left(u_{j_2}\right) \\
&= \sum_{j=1}^{m} \sup_{u_j\in[0,1]} \sum_{k_j=0}^{K_j} \left| \frac{k_j}{K_j} - u_j \right| P_{k_j,K_j}\left(u_j\right) \\
&\leq \sum_{j=1}^{m} \sup_{u_j\in[0,1]} \left(\sum_{k_j=0}^{K_j} \left(\frac{k_j}{K_j} - u_j \right)^2 P_{k_j,K_j}\left(u_j\right) \right)^{1/2} \\
&= \sum_{j=1}^{m} \sup_{u_j\in[0,1]} \left(\frac{u_j\left(1-u_j\right)}{K_j} \right)^{1/2} = \frac{1}{2} \sum_{j=1}^{m} K_j^{-1/2}.
\end{aligned}$$

□

Lemma 2.18. *Let $p \in (0,1)$. Suppose that $C_{\boldsymbol{X}}'\left(C_{\boldsymbol{X}}^{\leftarrow}(p)\right) > 0$ exists, then*

$$n^{1/2}\left(B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)^{\leftarrow}(p) - C_{\boldsymbol{X}}^{\leftarrow}(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2\left(C_{\boldsymbol{X}}^{\leftarrow}(p),\ldots,C_{\boldsymbol{X}}^{\leftarrow}(p)\right)}{\left(C_{\boldsymbol{X}}'\left(C_{\boldsymbol{X}}^{\leftarrow}(p)\right)\right)^2}\right),$$

where $\sigma^2(\boldsymbol{u}) = \mathbb{V}[\mathbb{C}(\boldsymbol{u})]$, $C_{\boldsymbol{X}}'$ is the first derivative of $u \mapsto C_{\boldsymbol{X}}(u,\ldots,u)$, and $C_{\boldsymbol{X}}^{\leftarrow}$, $B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)^{\leftarrow}$ is the quantile of $u \mapsto C_{\boldsymbol{X}}(u,\ldots,u)$, $u \mapsto B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)(u,\ldots,u)$, respectively.

Remark 2.19. In order to prove this lemma, we need a slightly extended version of Theorem 2.4. Let $\boldsymbol{u} \in [0,1]^m$ and $\boldsymbol{u}_n := \boldsymbol{u} \pm \boldsymbol{\epsilon}_n$, where $\boldsymbol{\epsilon}_n \to \boldsymbol{0}$ for $n \to \infty$, such that $\boldsymbol{u}_n \in [0,1]^m$ for all $n \in \mathbb{N}$. Then under the assumptions of Theorem 2.4 it holds that

$$n^{1/2} \cdot \left(B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)(\cdot \pm \boldsymbol{\epsilon}_n) - C_{\boldsymbol{X}}(\cdot \pm \boldsymbol{\epsilon}_n)\right) \xrightarrow{d} \mathbb{C}$$

in $(C([0,1]^m), \|\cdot\|_\infty)$.

The proof is essentially the same. Notice that Lemma 2.17 and Bernstein's theorem hold uniformly. This means that we can use Lemma 2.17 directly again and Bernstein's theorem with an additional argument. We used Bernstein's theorem to show the uniform convergence of $g_n(f) \to g(f)$ for $n \to \infty$ and all $f \in S'$. Recall that f is any continuous function on the compact set $[0,1]^m$. We need to show that $g_n(f) \to g(f)$ for $n \to \infty$ still holds uniformly when we transform the argument $\boldsymbol{u}$ of $g_n(f)$ to $\boldsymbol{u}_n$. We get that

$$\begin{aligned}
\sup_{\boldsymbol{u}\in[0,1]^m} \left| g_n(f)(\boldsymbol{u}_n) - g(f)(\boldsymbol{u}) \right| &= \sup_{\boldsymbol{u}\in[0,1]^m} \left| B_{\boldsymbol{K}(n)}(f)(\boldsymbol{u}_n) - f(\boldsymbol{u}) \right| \\
&\leq \left\| B_{\boldsymbol{K}(n)}(f) - f \right\|_\infty + \sup_{\boldsymbol{u}\in[0,1]^m} \left| f(\boldsymbol{u}_n) - f(\boldsymbol{u}) \right|.
\end{aligned}$$

The first summand again converges to zero because of Bernstein's theorem. The second summand converges to zero because of the uniform continuity of f. The function g_n' defined by $g_n'(f)(\boldsymbol{u}) := g_n(f)(\boldsymbol{u}_n) = B_{\boldsymbol{K}(n)}(f)(\boldsymbol{u}_n)$ is then used in the generalized continuous mapping theorem instead of g_n.

Proof. We argue similarly to the proof of Theorem A in Section 2.3.3 of Serfling (1980). Fix $p \in (0,1)$ and let

$$G_n(t) := \mathbb{P}\left[\frac{n^{1/2}\left(B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)^{\leftarrow}(p) - C_{\boldsymbol{X}}^{\leftarrow}(p)\right)}{\tilde{\sigma}} \leq t\right],$$

where $\tilde{\sigma} := \frac{\sigma\left(C_{\boldsymbol{X}}^{\leftarrow}(p),\ldots,C_{\boldsymbol{X}}^{\leftarrow}(p)\right)}{C_{\boldsymbol{X}}'\left(C_{\boldsymbol{X}}^{\leftarrow}(p)\right)}$. Let $u_n := t\tilde{\sigma}n^{-1/2} + C_{\boldsymbol{X}}^{\leftarrow}(p)$. We have

$$\begin{aligned} G_n(t) &= \mathbb{P}\left[B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)^{\leftarrow}(p) \leq u_n\right] \\ &= \mathbb{P}\left[p \leq B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)(u_n,\ldots,u_n)\right] \end{aligned}$$

Put $c_{nt} := \frac{n^{1/2}(C_{\boldsymbol{X}}(u_n,\ldots,u_n)-p)}{\sigma(u_n,\ldots,u_n)}$. Then it holds that

$$G_n(t) = \mathbb{P}[-c_{nt} \leq Z_n],$$

where $Z_n := \frac{n^{1/2}\left(B_{\boldsymbol{K}}\left(\hat{C}_{\boldsymbol{X},n}\right)(u_n,\ldots,u_n)-C_{\boldsymbol{X}}(u_n,\ldots,u_n)\right)}{\sigma(u_n,\ldots,u_n)}$. Furthermore, we get

$$\begin{aligned} \Phi(t) - G_n(t) &= \mathbb{P}[Z_n < -c_{nt}] - (1-\Phi(t)) \\ &= \mathbb{P}[Z_n < -c_{nt}] - \Phi(-c_{nt}) + \Phi(t) - \Phi(c_{nt}) \end{aligned} \tag{1}$$

Since $C_{\boldsymbol{X}}$ and $\partial_j C_{\boldsymbol{X}}$, $1 \leq j \leq m$, are continuous, we have

$$\begin{aligned} \lim_{n\to\infty} c_{nt} &= \lim_{n\to\infty}\left(t \cdot \frac{\tilde{\sigma}}{\sigma(u_n,\ldots,u_n)} \cdot \frac{C_{\boldsymbol{X}}(u_n,\ldots,u_n) - C_{\boldsymbol{X}}\left(C_{\boldsymbol{X}}^{\leftarrow}(p),\ldots,C_{\boldsymbol{X}}^{\leftarrow}(p)\right)}{t\tilde{\sigma}n^{-1/2}}\right) \\ &= t \cdot \frac{\tilde{\sigma}}{\sigma\left(C_{\boldsymbol{X}}^{\leftarrow}(p),\ldots,C_{\boldsymbol{X}}^{\leftarrow}(p)\right)} \cdot C_{\boldsymbol{X}}'\left(C_{\boldsymbol{X}}^{\leftarrow}(p)\right) \\ &= t. \end{aligned}$$

Next, we utilize Remark 2.19 (restricted to the point $\boldsymbol{u} := \left(C_{\boldsymbol{X}}^{\leftarrow}(p),\ldots,C_{\boldsymbol{X}}^{\leftarrow}(p)\right)$ with $\boldsymbol{u}_n := (u_n,\ldots,u_n)$) and Polya's Theorem (see Section 1.5.3 of Serfling (1980)) to show uniform convergence of the distribution function of Z_n to the standard normal distribution function.

Since Φ is continuous, we have

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{P}[Z_n \leq x] - \Phi(x)| = 0.$$

Using these two properties, (1) results in

$$\begin{aligned}\lim_{n\to\infty} |\Phi(t) - G_n(t)| &\leq \lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{P}[Z_n < x] - \Phi(x)| + \lim_{n\to\infty} |\Phi(t) - \Phi(c_{nt})| \\ &= 0.\end{aligned}$$

□

References

Aitchison, J. (1964). Confidence-region tests. *J. R. Stat. Soc., Ser. B 26*, 462–476.

Belalia, M. (2016). On the asymptotic properties of the Bernstein estimator of the multivariate distribution function. *Stat. Probab. Lett. 110*, 249–256.

Bodnar, T. and T. Dickhaus (2014). False discovery rate control under Archimedean copula. *Electron. J. Stat. 8*(2), 2207–2241.

Bouzebda, S. and T. Zari (2013). Strong approximation of empirical copula processes by Gaussian processes. *Statistics 47*(5), 1047–1063.

Bücher, A. and H. Dette (2010). A note on bootstrap approximations for the empirical copula process. *Stat. Probab. Lett. 80*(23-24), 1925–1932.

Cerqueti, R., M. Costantini, and C. Lupi (2012). A copula-based analysis of false discovery rate control under dependence assumptions. Economics & Statistics Discussion Paper 065/12, Università degli Studi del Molise, Dipartimento di Scienze Economiche, Gestionali e Sociali (SEGeS).

Charpentier, A., J.-D. Fermanian, and O. Scaillet (2007). The estimation of copulas: Theory and practice. In J. Rank (Ed.), *Copulas: From Theory to Application in Finance*, pp. 35–62. London: Risk Books.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal. 31*(2), 131–145.

Chêng, F. H. (1983). On the rate of convergence of Bernstein polynomials of functions of bounded variation. *J. Approx. Theory 39*(3), 259–274.

Cottin, C. and D. Pfeifer (2014). From Bernstein polynomials to Bernstein copulas. *J. Appl. Funct. Anal. 9*(3-4), 277–288.

Deheuvels, P. (1979). La fonction de dependance empirique et ses propriétés. Un test non paramètrique d'independance. *Bull. Cl. Sci., V. Sér., Acad. R. Belg. 65*, 274–292.

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

Dickhaus, T. and J. Gierl (2013). 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)*, pp. 75–80. Global Science and Technology Forum (GSTF).

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 Statist. Assoc. Bull. 65*(257-260), 123–144.

Diers, D., M. Eling, and S. D. Marek (2012). Dependence modeling in non-life insurance using the Bernstein copula. *Insur. Math. Econ. 50*(3), 430–436.

Embrechts, P., F. Lindskog, and A. McNeil (2003). Modelling dependence with copulas and applications to risk management. In S. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384. Elsevier Science B.V.

Gijbels, I. and J. Mielniczuk (1990). Estimating the density of a copula function. *Comm. Statist. Theory Methods 19*(2), 445–464.

Härdle, W. K. and O. Okhrin (2010). De copulis non est disputandum - Copulae: an overview. *AStA Adv. Stat. Anal. 94*(1), 1–31.

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

Janssen, P., J. Swanepoel, and N. Veraverbeke (2012). Large sample behavior of the Bernstein copula estimator. *J. Statist. Plann. Inference 142*(5), 1189–1197.

Joe, H. (2014). *Dependence modeling with copulas.* Boca Raton, FL: CRC Press.

Kiefer, J. (1961). On large deviations of the empiric D. F. of vector chance variables and a law of the iterated logarithm. *Pacific J. Math. 11*, 649–660.

Lambert, P. (2007). Archimedean copula estimation using Bayesian splines smoothing techniques. *Comput. Stat. Data Anal. 51*(12), 6307–6320.

Longin, F. M. (2000). From value at risk to stress testing: The extreme value approach. *Journal of Banking & Finance 24*, 1097–1130.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative risk management. Concepts, techniques, and tools.* Princeton, NJ: Princeton University Press.

Nelsen, R. B. (2006). *An introduction to copulas. 2nd ed.* Springer Series in Statistics. New York, NY: Springer.

Omelka, M., I. Gijbels, and N. Veraverbeke (2009). Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing. *Ann. Statist. 37*(5B), 3023–3058.

Pfeifer, D., A. Mändle, and O. Ragulina (2017). New copulas based on general partitions-of-unity and their applications to risk management (part II). *Depend. Model. 5*(1), 246–255.

Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Stat. 4*, 912–923.

Sancetta, A. and S. Satchell (2004). The bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory 20*(03), 535–562.

Schmidt, R., A. Faldum, and J. Gerß (2015). Adaptive designs with arbitrary dependence structure based on Fisher's combination test. *Stat. Methods Appl. 24*(3), 427–447.

Schmidt, R., A. Faldum, O. Witt, and J. Gerß (2014). Adaptive designs with arbitrary dependence structure. *Biom. J. 56*(1), 86–106.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Comm. Statist. A—Theory Methods 14*(5), 1123–1136.

Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli 18*(3), 764–782.

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

Stange, J., T. Bodnar, and T. Dickhaus (2015). Uncertainty quantification for the family-wise error rate in multivariate copula models. *AStA Adv. Stat. Anal. 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. *Stat. Probab. Lett. 111*, 32–40.

Stute, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab. 12*(2), 361–379.

Swanepoel, J. W. H. (1986). A note on proving that the (modified) bootstrap works. *Comm. Statist. A—Theory Methods 15*(11), 3193–3203.

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.

Whitt, W. (2002). *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer Series in Operations Research. Springer-Verlag, New York.

3 Estimating the proportion of true null hypotheses under arbitrary dependency

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

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.

Key words: Bootstrap; Copula; Family-wise error rate; p-Value; Schweder-Spjøtvoll estimator.

3.1 Introduction

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 (see Section 3.1.3 in Dickhaus (2014)) incorporate an estimate $\hat{\pi}_0$ into their decision rules in order to optimize power (see Langaas et al. (2005), Finner and Gontscharuk (2009), Celisse and Robin (2010) and 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

[1]Institute for Statistics, University of Bremen, Bibliothekstraße 1, D-28359 Bremen, Germany.
[2]Department of Mathematics, Stockholm University, Roslagsvägen 101, SE-10691 Stockholm, Sweden.

Schweder-Spjøtvoll estimator $\hat{\pi}_0^{\text{SS}}$ for π_0 has originally been proposed by Schweder and Spjøtvoll (1982). Theoretical properties of $\hat{\pi}_0^{\text{SS}}$ 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 and Chen (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 $\hat{\pi}_0^{\text{SS}}$. 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 (see 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 (see Sklar (1996)). Unfortunately, as we will demonstrate in Example 3.1 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 ecdf of all m marginal p-values.

The estimator $\hat{\pi}_0^{\text{SS}}$ 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 $\hat{F}_m$, $\hat{\pi}_0^{\text{SS}}$ is given by

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

This form of the Schweder-Spjøtvoll estimator has been mentioned by Storey (2002); Storey et al. (2004). There exist several heuristic motivations for the usage of $\hat{\pi}_0^{\text{SS}}$. 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 $\hat{\pi}_0^{\text{SS}}(\lambda)$ (see Figure 3.2 (a) in Dickhaus (2014)). A graphical algorithm for computing $\hat{\pi}_0^{\text{SS}}$ connects the point $(\lambda, \hat{F}_m(\lambda))$ with the point $(1,1)$. The offset of the resulting straight line at $t = 0$ equals

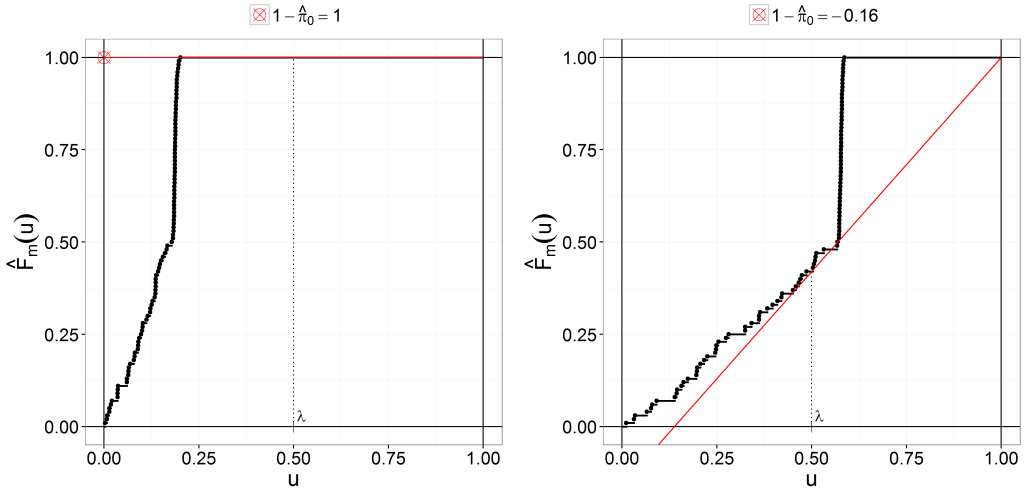

Figure 3: Computer simulations of the behavior of $\hat{\pi}_0^{SS}$ under a Gumbel-Hougaard copula with copula parameter $\eta = 100$. In the left graph, the p-values corresponding to true null hypotheses cluster around a value smaller than λ, while in the right graph they cluster around a value larger than λ.

$\hat{\pi}_1^{SS} := \hat{\pi}_1^{SS}(\lambda) := 1 - \hat{\pi}_0^{SS}(\lambda)$ (see Figure 3.2 (b) in Dickhaus (2014)). 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 worst case 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. For a more realistic setups see Section 3.4.

Example 3.1. Assume that the copula of $\boldsymbol{P} = (P_1, \ldots, P_m)^\top$ is a Gumbel-Hougaard copula with copula parameter $\eta \geq 1$ (cf. 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 3, we present two computer simulations for $m = 100$, $m_0 = 50$, $\eta = 100$, and $\lambda = 1/2$. In both graphs displayed in Figure 3, 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.

Under this model, the behavior of $\hat{\pi}_0^{SS}$ (indicated by the straight lines) can be char-

acterized 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 $\hat{\pi}_0^{\mathrm{SS}}$ would remain exactly the same for a different value of m_0. Whether $\hat{\pi}_0^{\mathrm{SS}}$ 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 3.1 demonstrates that some structural information about the dependency structure among the test statistics or p-values, respectively, is 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 $\hat{\pi}_0^{\mathrm{SS}}$ instead of the original p-values. Applying this methodology to the situation considered in Example 3.1 leads to an accurate estimate of π_0, see Example 3.5 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 $\hat{\pi}_0^{\mathrm{SS}}$, 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 manuscript is structured as follows. In Section 3.2, we introduce the proposed bootstrap procedure. Theoretical properties of this procedure are analyzed in Section 3.3. The sensitivity of this procedure is discussed for a simulation study in Section

3.4. A real data example from cancer research is presented in Section 3.5. We conclude with a discussion in Section 3.6.

3.2 Estimation of π_0 via marginal parametric bootstrap

We consider multiple test problems of the form as in Section 2.3. Additionally, we assume that hypotheses have the structure $H_j = \left\{\boldsymbol{\vartheta} \in \Theta \middle| \vartheta_j = \theta_j\right\}$, 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 (see Dickhaus (2013)).

Our proposed bootstrap method for estimating π_0 under arbitrary copula C_X of $\boldsymbol{X}$ is formalized in Algorithm 3.2.

Algorithm 3.2. *For all $1 \leq j \leq m$, let $\hat{\vartheta}_{j,n}$ be a consistent estimator of ϑ_j, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ 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 $1 \leq b \leq B$ and $1 \leq j \leq m$*
   (a) *draw a bootstrap sample $X_{1,j}^{*(b)}, \ldots, X_{k(n),j}^{*(b)}$ of size $k(n)$ from the marginal distribution of X_j with estimated parameters $\hat{\vartheta}_{j,n}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.*
   (b) *calculate the bootstrap test statistics $T_j^* = T_{j,n}^{*(b)} := T_{j,k(n)}\left(X_{1,j}^{*(b)}, \ldots, X_{k(n),j}^{*(b)}\right)$.*
   (c) *calculate the bootstrap p-values $P_j^* = P_{j,n}^{*(b)} := 1 - F_{T_{j,k(n)}|\theta_j}\left(T_{j,n}^{*(b)}\right)$.*
2. *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 $P_{1,n}^{*(b)}, \ldots, P_{m,n}^{*(b)}$.
3. *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 ϑ_j. Then, we calculate the test statistics and p-values based on these pseudo samples instead of the original data. In steps 2 and 3, we finally compute the Schweder-Spjøtvoll estimator. We will show in Lemma 3.6 that the resulting bootstrap p-values are indeed conditionally independent given the data. Therefore, in contrast to Figure 3 we can expect the Schweder-Spjøtvoll

estimator, applied to the bootstrapped p-values, to behave as in the case of joint independence of $X_1, \ldots, X_m$. Hence, it is important to perform the bootstrap marginally. If one would use a multivariate bootstrap procedure instead, then the estimator would still suffer under dependency. For example, Lu and Perkins (2007) observed this for strong correlation in microarray data.

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

Assumption 3.3.

(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 (see Van Der Laan and Bryan (2001)), in the analysis of variance (ANOVA) (see Krishnamoorthy et al. (2007)), for goodness-of-fit statistics (Cramér-von Mises, Kolmogorov-Smirnov) (see Genest and Rémillard (2008)), and for Wald statistics in dynamic factor models (see 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 3.4. 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 $\boldsymbol{X}$. For example, the 'max T' 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 3.3, the following example illustrates the applicability of Algorithm 3.2 in extreme situations like in Example 3.1. In Section 3.4 we take a look at more realistic settings.

Example 3.5 (*Example 3.1 continued.*)**.** We consider the setup of Example 3.1, 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 3.1. 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 $H_j = \left\{\boldsymbol{\vartheta} \in (0,\infty)^m \middle| \vartheta_j = \theta_j\right\}$ versus the (one-sided) alternative $K_j := \left\{\boldsymbol{\vartheta} \in (0,\infty)^m \middle| \vartheta_j > \theta_j\right\}$, 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 $a_n = 1 - F^{-1}_{Z_{1,1}}\left(1 - n^{-1}\right)$ 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 $\hat{\vartheta}_{j,n}$ of ϑ_j is given by $\hat{\vartheta}_{j,n}\left(X_{1,j}, \ldots, X_{n,j}\right) = \bar{X}_j\left(\alpha+\beta\right)/\alpha$.

The plug-in rule of Algorithm 3.2 now yields bootstrap variates $X^*_{i,j} \overset{d}{=} \hat{\vartheta}_{j,n}(x_{1,j}, \ldots, x_{n,j}) \cdot Z_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m$, for observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with corresponding test statistics and p-values.

The validity of part (a) of Assumption 3.3 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.,

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

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 $a'_n := a_n \hat{\vartheta}_{j,n}/\theta_j$ and $b'_n := b_n \hat{\vartheta}_{j,n}/\theta_j$. We get that $a_n^{-1} a'_n$ converges to 1 almost surely for $n \to \infty$. Now, assume for the moment that $a_n^{-1}\left(b'_n - b_n\right)$ converges to 0 almost surely for $n \to \infty$. Then we get for all $1 \leq j \leq m_0$ and $x \in \mathbb{R}$ that

$$\begin{aligned}
\left|F^*_{T^*_{j,n}}(x) - G(x)\right| &= \left|F^n_{Z_{1,1}}\left(\frac{\theta_j}{\hat{\vartheta}_{j,n}}(a_n x + b_n)\right) - G(x)\right| \\
&= \left|F^n_{Z_{1,1}}\left(a'_n x + b'_n\right) - G(x)\right| \\
&= \left|F_{T_{j,n}}\left(a_n^{-1} a'_n x + a_n^{-1}\left(b'_n - b_n\right)\right) - G(x)\right| \\
&\leq \left\|F_{T_{j,n}} - G\right\|_\infty + \left|G\left(a_n^{-1} a'_n x + a_n^{-1}\left(b'_n - b_n\right)\right) - G(x)\right| \\
&\to 0 \text{ almost surely for } n \to \infty.
\end{aligned}$$

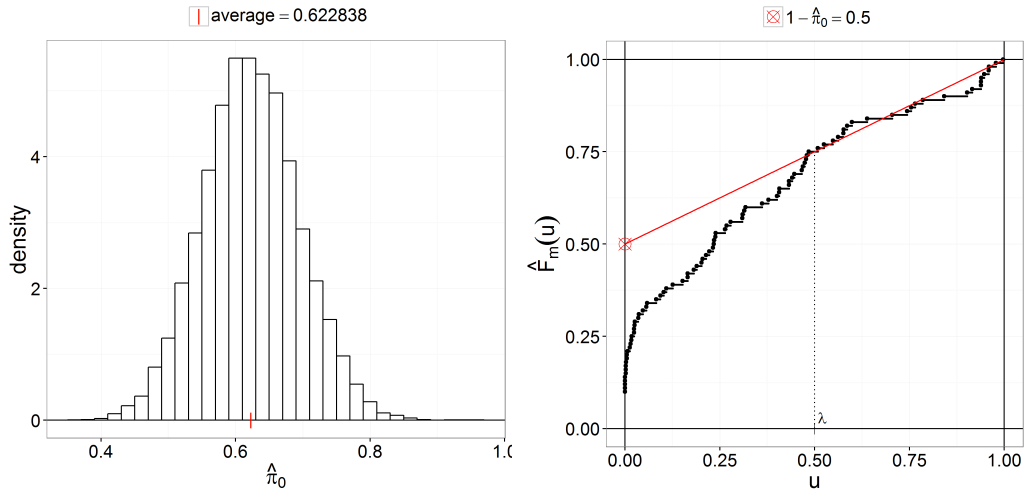

Figure 4: Computer simulation of $\overline{\hat{\pi}}_0^*$ in the setting of Example 3.5 with $\alpha = 1$, $\beta = 4$, $n = 100$, $m = 100$, $m_0 = 50$ and $\eta = 100$. The parameter values $\vartheta_{m_0+1}, \ldots, \vartheta_{100} \in (1.5, 2.5)$ have been chosen uniformly. The right graph displays 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 $\overline{\hat{\pi}}_0^* = B^{-1} \sum_{b=1}^{B} \hat{\pi}_{0,n}^{*(b)} \approx 0.62$.

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 $o\left(n^{-1/2}\right)$ almost surely (see Theorem 2.5.8 in Durrett (2010)). 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 $a_n^{-1}\left(b_n' - b_n\right) = o(1)/\left(n^{(1/2-\epsilon)-1/\beta}\right)$ indeed converges to 0 almost surely for $n \to \infty$.

Finally, Pólya's uniform convergence theorem yields that

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

almost surely for $n \to \infty$ and all $1 \le j \le m_0$, since the limiting cdf G is continuous.

Figure 4 displays the results of a computer simulation employing Algorithm 3.2 in this example, where $\eta = 100$. One may compare the right graph in Figure 4 with Figure 3 for a demonstration of the improvement of estimation accuracy obtained by applying Algorithm 3.2.

3.3 Theoretical analysis

In this section, we analyze the theoretical properties of our bootstrap estimator $\overline{\hat{\pi}}_0^*$. For the ease of notation, let $\mathbb{P} = \mathbb{P}_{\boldsymbol{\vartheta}, C_X}^{\otimes\infty}$ denote the true distribution of the data sample 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 3.6. *Let $m, n \in \mathbb{N}$ 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 $\boldsymbol{X}_1^{*(b)}, \ldots, \boldsymbol{X}_n^{*(b)}$ 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 3.2 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 3.7. Under the assumptions of Lemma 3.6 we get by the strong law of large numbers that

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

Lemma 3.8. *Let $b \in \{1, \ldots, B\}$ and $m \in \mathbb{N}$ 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 3.3 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 3.9 is mainly an asymptotic property with respect to the sample size $n \to \infty$.

Theorem 3.9. *Let $m \in \mathbb{N}$ be fixed. We have*

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

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

Proof. To proof this theorem, we combine Remark 3.7 with Lemma 3.8. 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 3.7 yields that for every observed data sample it holds $\mathbb{P}^*$-almost surely that

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

Furthermore, from Lemma 3.8 it follows that

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

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

Hence, we get $\mathbb{P}$-almost surely that

$$\begin{aligned}\lim_{n_k\to\infty}\frac{1-m^{-1}\sum_{j=1}^m \mathbb{P}^*\left[P^*_{j,n_k}\le\lambda\right]}{1-\lambda} &= \lim_{n_k\to\infty}\frac{1-\pi_0\lambda-m^{-1}\sum_{j=m_0+1}^m \mathbb{P}^*\left[P^*_{j,n_k}\le\lambda\right]}{1-\lambda} \\ &\ge \frac{1-\pi_0\lambda-(1-\pi_0)}{1-\lambda} \\ &= \pi_0.\end{aligned} \tag{3.1}$$

Thus, the assertion follows by the subsequence principle. □

Theorem 3.9 shows that the bootstrap estimator $\overline{\hat{\pi}}^*_0$ is asymptotically non-negatively biased (i.e., mean conservative).

Corollary 3.10. *Theorem 3.9 also shows that we achieve an asymptotically unbiased estimator of π_0 whenever $\mathbb{P}^*\left[P^*_{j,n}\le\lambda\right]$ tends to one for $n\to\infty$ under alternatives, because in such cases inequality (3.1) becomes an equality. This means that for example every consistent multiple test (see Troendle (2000)) leads to an asymptotically unbiased estimator.*

Let us demonstrate the assertion of Corollary 3.10 in a small simulation. Assume a multivariate normal model with equi-correlation coefficient $\rho = 0.7$ and known variances equal to 1. The marginal means are tested against zero and take absolute values between 0.5 and 2 under alternatives; see Section 3.4 for a detailed description of this model. The assumption of consistency of the marginal tests is fulfilled here. The number of hypotheses m is equal to 100 and π_0 is set to 0.5. We only present results for $B = 100$

Table 6: Bias and MSE of the marginal bootstrap procedure for fixed $B = 100$ and varying values of n.

n	bias	MSE
10	0.069	0.00661
30	0.039	0.00214
50	0.020	0.00082
100	0.005	0.00032
1000	-0.006	0.00013
10000	-0.002	0.00006

here. For larger values of B, simulation results were similar. In Table 6, the performance of the proposed estimator is displayed in terms of its bias and its mean squared error (MSE), where the sample size n takes values from 10 to 10,000. It is clearly visible that bias and MSE approach zero for $n \to \infty$ when B is large enough. Throughout the remainder, we will evaluate the performance of concurring estimators by means of bias and MSE.

3.4 Simulation study

In this section, we present the results of a simulation study in order to (i) investigate the performance of the proposed marginal bootstrap procedure for finite n, and (ii) compare the proposed procedure with existing approaches for estimating π_0 taken from the literature.

To this end, we consider an equi-correlated multivariate normal model. In order to analyze the sensitivity of the procedures with respect to varying parameters, we used various values for the equi-correlation ρ, the number of hypotheses m, the ratio of true null hypotheses π_0 and the sample size n. There exist several methods for estimating π_0, which are implemented in `R`. The container method `estim.pi0` in the package `cp4p` was used, which includes methods from the packages `limma` and `qvalue`.

Since our bootstrap algorithm is a strategy to deal with dependencies, it can easily be combined with all of these methods. Most of them assume independent p-values. In total nine methods, all relying on p-values $P_1, \ldots, P_m$, were compared with their respective marginal bootstrap version. Namely, the methods from Schweder and Spjøtvoll (1982), Storey and Tibshirani (2003), Storey et al. (2004), Langaas et al. (2005), Nettleton et al. (2006), Pounds and Cheng (2006), Jiang and Doerge (2008), Wang et al. (2011) and Phipson (2013) have been taken into account.

The model is as follows. Assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $n \in \{10, 30, 50, 100\}$, is a sample of normally distributed random vectors in $\mathbb{R}^m$, $m \in \{10, 50, 100\}$, with covariance matrix $\Sigma = (1-\rho)I_m + \rho \mathbf{1}_m \mathbf{1}_m^\top$ of $\boldsymbol{X}_1$, with equi-correlation coefficient $\rho \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$.

In the data analysis, the marginal variances have been assumed to be unknown. The proportion of true null hypotheses π_0 takes values in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The parameters of interest are the marginal expectations $\vartheta_1, \ldots, \vartheta_m$ of $\boldsymbol{X}_1$ and we carry out two-sided t-tests for the marginal hypotheses $H_j = \left\{\boldsymbol{\vartheta} \in \mathbb{R}^m \middle| \vartheta_j = \theta_j\right\}$, $1 \leq j \leq m$, where $\theta_1, \ldots, \theta_m$ are set to zero. Under alternatives the θ_j have been drawn from a uniform distribution, with absolute values between 0.5 and 2. The marginal p-values have been computed utilizing Student's t-distribution with $n-1$ degrees of freedom.

Under this model, Assumption 3.3 can be checked as follows. Notice that part (b) of the assumption is fulfilled by construction. With respect to part (a), notice that taking $k(n) = n$ in Algorithm 3.2 would lead to a violation of the assumption. Namely, the numerator $\sqrt{k(n)}\left(\bar{X}_j^* - \theta_j\right)$ of the j-th bootstrap test statistic is conditionally (given the data) normally distributed with variance $\hat{\sigma}_{j,n}$ and expectation $\sqrt{k(n)} \cdot \left(\hat{\vartheta}_{j,n} - \theta_j\right)$. For the validity of part (a) of Assumption 3.3, the latter expectation needs to converge to zero in probability when regarded as a random variable. However, this is not the case if $k(n) = n$, because $\sqrt{n} \cdot \left(\hat{\vartheta}_{j,n} - \theta_j\right)$ has a non-degenerate limit distribution. However, taking $k(n) = n^{1-\varepsilon}$ for arbitrary $\varepsilon > 0$ leads to convergence of $\sqrt{k(n)} \cdot \left(\hat{\vartheta}_{j,n} - \theta_j\right)$ to zero in probability as $n \longrightarrow \infty$. On the basis of preliminary simulations, we chose $\varepsilon = 1/2$, meaning that we set $k(n)$ to the nearest integer smaller than or equal to $\sqrt{n}$. For $k(n) = \lfloor \sqrt{n} \rfloor$ every bootstrap test statistic conditionally follows a t-distribution with $k(n) - 1$ degrees of freedom under H_j. Since the t-distribution converges to the standard normal distribution for $n \longrightarrow \infty$, the bootstrap distribution functions converge pointwise to the true distribution functions under H_j. This convergence holds uniformly by Pólya's uniform convergence theorem (see Section 1.5.3 in Serfling (1980)) and therefore, assumption (a) is fulfilled. The downside of this choice of $k(n)$ is that the convergence of our method is much slower and a larger sample size is required to get the same precision.

Table 7 and Table 8 show a small part of the comparison. Due to space constraints we restrict ourselves in these tables to the Schweder-Spjøtvoll estimate with tuning parameter $\lambda = 0.5$ (SS), the proposed bootstrap version of it (BSS), and the standard methods of the packages `limma` (Phipson, P) and `qvalue` (Storey-Tibshirani, ST). The complete simulation results can be found in the supplementary material. In order to complete this large-scale simulation study within a reasonable amount of time, we had to restrict the number of Monte Carlo repetitions per simulation setting to 100.

The Schweder-Spjøtvoll estimator maintains a small bias for almost all settings, but the MSE increases considerably with stronger dependencies and larger π_0. In comparison, our bootstrap version maintains a very small MSE in all settings. On the other hand, its bias can be considerably larger for small π_0 and moderate values of n. For large π_0 the bootstrap estimates can be negatively biased, but this bias gets closer to zero with

increasing n. This confirms what we have shown in Theorem 3.9 and observed in the small simulation at the end of the last section. Overall, the estimate is conservative in most cases and can be favorable even for moderate sample sizes. If one can choose larger values for $k(n)$ for different test statistics, then our method can even be applied for small sample sizes. The Phipson estimate is often not conservative and underestimates the true π_0, but it has a small bias and MSE in most cases. The method gets worse with stronger dependencies and larger π_0. The Storey-Tibshirani estimate has a large bias and MSE. It is by far the most conservative estimate for weaker dependencies here. But this estimate considerably underestimate the true π_0 for stronger dependencies and larger π_0 as well.

This simulation shows that our method can be applied in all settings, leading to mostly conservative estimates. It is clearly favorable to use the marginal bootstrap for stronger dependencies, since the other methods lead to a large bias or MSE in these situations.

Remark 3.11. To confirm our theoretical reasoning, we also performed some simulations in the case of $k(n) = n$, which are not presented here. These simulations confirmed bad results, i.e., much larger bias.

3.5 Real data analysis

One application of the estimation of π_0 is the calculation of posterior probabilities for the validity of null hypotheses in an empirical Bayes model. In the context of control of the positive false discovery rate (pFDR), such posterior probabilities have been referred to as q-values by Storey (2003). The (p)FDR is nowadays a standard type I error criterion for large-scale multiple test problems. The estimation of π_0 in this context has been treated, e.g., by Lai (2007), Lu and Perkins (2007), Tong and Zhao (2008), Hunt et al. (2009), Tong et al. (2013), Cheng et al. (2015), and Singh et al. (2015). Most of the latter references consider methods for estimating π_0 which are specifically targeted towards microarray data.

In this section, we compare different estimators of π_0 on the basis of the gene expression dataset from Alon et al. (1999), which has been analyzed by Tong and Zhao (2008). As mentioned by Tong et al. (2013), for example, the independence assumption is often not justified in microarray gene expression data due to co-regulation. We include in our comparison the methods from Section 3.4, and additionally the method `SamS` from Lu and Perkins (2007), because an implementation thereof is publicly available, in contrast to the other methods mentioned before.

The data are available in various `R` packages, for example in the package `plsgenomics` using the command `data(Colon)`. The dataset contains $n_1 = 22$ normal colon samples and $n_2 = 40$ colon tumor samples for $m = 2000$ genes. The aim of this study was to identify genes, which exhibit significant differential expression between the groups. Fol-

Table 7: b = bias, M = MSE, SS = Schweder-Spjøtvoll, BSS = bootstrap Schweder-Spjøtvoll, P = Phipson, ST = Storey-Tibshirani. The (mean) bias and MSE are calculated over 100 repetitions and the bootstrap repetitions are set to $B = 100$.

ρ	m	n	π_0	b(SS)	M(SS)	b(BSS)	M(BSS)	b(P)	M(P)	b(ST)	M(ST)
0	50	30	0.1	0.000	0.002	0.119	0.014	0.002	0.001	0.227	0.064
0	50	30	0.3	0.003	0.005	0.109	0.012	0.005	0.002	0.226	0.098
0	50	30	0.5	-0.009	0.011	0.036	0.002	-0.013	0.003	0.202	0.096
0	50	30	0.7	-0.002	0.014	0.015	0.001	-0.014	0.004	0.091	0.056
0	50	30	0.9	0.026	0.016	-0.039	0.002	-0.013	0.003	0.026	0.016
0	50	100	0.1	-0.002	0.002	0.036	0.001	-0.003	0.001	0.232	0.068
0	50	100	0.3	-0.007	0.006	0.031	0.001	-0.007	0.002	0.201	0.075
0	50	100	0.5	0.003	0.011	0.003	0.000	-0.011	0.003	0.191	0.084
0	50	100	0.7	-0.010	0.013	-0.006	0.000	-0.026	0.005	0.079	0.054
0	50	100	0.9	-0.003	0.017	-0.028	0.001	-0.028	0.005	0.002	0.026
0	100	30	0.1	0.001	0.001	0.123	0.015	0.008	0.000	0.120	0.023
0	100	30	0.3	-0.005	0.002	0.093	0.009	0.004	0.001	0.094	0.034
0	100	30	0.5	-0.001	0.005	0.023	0.001	-0.005	0.001	0.096	0.046
0	100	30	0.7	0.005	0.006	0.007	0.000	0.000	0.002	0.082	0.034
0	100	30	0.9	-0.003	0.010	-0.051	0.003	-0.018	0.003	-0.009	0.019
0	100	100	0.1	0.002	0.001	0.040	0.002	0.001	0.000	0.128	0.023
0	100	100	0.3	0.001	0.003	0.030	0.001	0.002	0.001	0.123	0.045
0	100	100	0.5	0.002	0.005	-0.004	0.000	-0.002	0.002	0.113	0.054
0	100	100	0.7	-0.003	0.008	-0.011	0.000	-0.008	0.002	0.070	0.040
0	100	100	0.9	-0.022	0.009	-0.034	0.001	-0.024	0.003	0.004	0.019
0.3	50	30	0.1	0.006	0.002	0.084	0.007	0.007	0.001	0.216	0.058
0.3	50	30	0.3	-0.023	0.010	0.125	0.016	-0.012	0.004	0.175	0.072
0.3	50	30	0.5	-0.013	0.020	0.059	0.004	-0.010	0.009	0.200	0.098
0.3	50	30	0.7	0.016	0.034	-0.008	0.001	-0.008	0.012	0.141	0.061
0.3	50	30	0.9	0.005	0.045	-0.050	0.004	-0.039	0.016	-0.025	0.036
0.3	50	100	0.1	-0.003	0.003	0.022	0.001	-0.004	0.001	0.207	0.057
0.3	50	100	0.3	0.006	0.008	0.049	0.003	0.002	0.003	0.252	0.112
0.3	50	100	0.5	0.012	0.013	0.020	0.001	0.007	0.005	0.210	0.098
0.3	50	100	0.7	0.018	0.028	-0.013	0.001	-0.002	0.010	0.115	0.054
0.3	50	100	0.9	-0.030	0.048	-0.038	0.002	-0.041	0.020	-0.045	0.050

Table 8: b = bias, M = MSE, SS = Schweder-Spjøtvoll, BSS = bootstrap Schweder-Spjøtvoll, P = Phipson, ST = Storey-Tibshirani. The (mean) bias and MSE are calculated over 100 repetitions and the bootstrap repetitions are set to $B = 100$.

ρ	m	n	π_0	b(SS)	M(SS)	b(BSS)	M(BSS)	b(P)	M(P)	b(ST)	M(ST)
0.3	100	30	0.1	0.004	0.001	0.120	0.015	0.007	0.001	0.119	0.026
0.3	100	30	0.3	-0.006	0.006	0.074	0.006	0.001	0.003	0.134	0.050
0.3	100	30	0.5	-0.019	0.014	0.042	0.002	-0.010	0.006	0.075	0.051
0.3	100	30	0.7	0.022	0.021	0.012	0.000	0.007	0.009	0.076	0.063
0.3	100	30	0.9	0.027	0.034	-0.037	0.002	-0.006	0.011	-0.021	0.035
0.3	100	100	0.1	-0.004	0.001	0.040	0.002	-0.002	0.000	0.110	0.020
0.3	100	100	0.3	-0.012	0.006	0.018	0.000	-0.010	0.003	0.086	0.033
0.3	100	100	0.5	0.011	0.013	0.008	0.000	0.002	0.006	0.114	0.072
0.3	100	100	0.7	-0.015	0.024	-0.010	0.000	-0.019	0.011	0.063	0.054
0.3	100	100	0.9	-0.013	0.042	-0.030	0.001	-0.030	0.016	-0.035	0.039
0.7	50	30	0.1	0.002	0.004	0.096	0.010	0.003	0.001	0.222	0.062
0.7	50	30	0.3	-0.009	0.026	0.080	0.007	-0.003	0.010	0.196	0.101
0.7	50	30	0.5	-0.020	0.063	0.022	0.002	-0.017	0.029	0.168	0.125
0.7	50	30	0.7	-0.048	0.129	-0.014	0.002	-0.056	0.057	0.000	0.096
0.7	50	30	0.9	-0.040	0.246	-0.052	0.007	-0.128	0.090	-0.122	0.102
0.7	50	100	0.1	0.005	0.004	0.026	0.001	0.001	0.002	0.256	0.086
0.7	50	100	0.3	0.006	0.022	0.026	0.001	0.001	0.009	0.256	0.120
0.7	50	100	0.5	-0.008	0.064	-0.002	0.000	-0.013	0.027	0.165	0.112
0.7	50	100	0.7	0.004	0.145	-0.017	0.001	-0.029	0.057	0.046	0.088
0.7	50	100	0.9	0.011	0.234	-0.035	0.003	-0.098	0.086	-0.102	0.096
0.7	100	30	0.1	0.010	0.003	0.122	0.016	0.012	0.001	0.143	0.033
0.7	100	30	0.3	0.027	0.021	0.082	0.007	0.018	0.009	0.140	0.085
0.7	100	30	0.5	0.017	0.065	0.058	0.005	0.012	0.028	0.115	0.129
0.7	100	30	0.7	-0.011	0.125	-0.007	0.001	-0.023	0.050	-0.036	0.116
0.7	100	30	0.9	-0.027	0.214	-0.049	0.004	-0.091	0.062	-0.185	0.151
0.7	100	100	0.1	-0.005	0.003	0.038	0.002	-0.004	0.001	0.113	0.021
0.7	100	100	0.3	0.005	0.023	0.021	0.001	0.003	0.010	0.123	0.065
0.7	100	100	0.5	-0.011	0.074	0.013	0.001	-0.016	0.035	0.091	0.117
0.7	100	100	0.7	-0.102	0.156	-0.026	0.002	-0.091	0.077	-0.093	0.141
0.7	100	100	0.9	-0.011	0.207	-0.035	0.003	-0.094	0.075	-0.183	0.152

Table 9: Estimation of π_0 for different methods. The number of bootstrap repetitions has been set to $B = 100$.

	Original	Bootstrap version
Schweder-Spjøtvoll	0.616	0.819
Phipson	0.612	0.809
Storey-Tibshirani	0.472	0.759
Storey et al.	0.514	0.754
Jiang-Doerge	0.531	0.767
Nettleton et al.	0.572	0.770
Langaas et al.	0.502	0.754
Pounds-Cheng	0.613	0.810
Wang et al.	0.735	0.876
Lu-Perkins	0.651	0.911

lowing the steps of Section 3.6 in Tong and Zhao (2008), we perform for every gene an unbalanced two-sample t-test with equal variances on the normalized dataset. Tong and Zhao (2008) mention that equal variances are only assumed for convenience. The p-values are approximated using the standard normal distribution and the tuning parameter λ of the Schweder-Spjøtvoll estimator and its variants is set to the median of the observed p-values, for the sake of comparability with the initial data analysis by Tong and Zhao (2008), where the Schweder-Spjøtvoll estimate $\hat{\pi}_0^{SS} \approx 0.616$ was used.

Table 9 displays the different estimation results for π_0. The original estimates range from 0.47 to 0.74. The bootstrap estimates are much more conservative in this dataset and range from 0.75 to 0.91. Notice that the range of these values is much smaller compared to the range of the values obtained with the original methods.

3.6 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 (see 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 3.2 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 3.3 in the case of a completely unspecified copula C_X. One could, however, consider special (parametric) model classes for C_X and design whitening procedures which exploit these parametric assumptions regarding the dependencies.

References

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences 96*(12), 6745–6750.

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 Statist. Assoc. Bull. 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. *Annales de l'IHP Probabilités et statistiques 44*(6), 1096–1127.

Hunt, D. L., C. Cheng, and S. Pounds (2009). The beta-binomial distribution for estimating the number of false rejections in microarray gene expression studies. *Comput. Statist. Data Anal. 53*(5), 1688–1700.

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.

Jiang, H. and R. Doerge (2008). Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer informatics 6*, 25.

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.

Lai, Y. (2007). A moment-based method for estimating the proportion of true null hypotheses and its application to microarray gene expression data. *Biostatistics 8*(4), 744–755.

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.

Lu, X. and D. L. Perkins (2007). Re-sampling strategy to improve the estimation of number of null hypotheses in fdr control under strong correlation structures. *BMC bioinformatics 8*(1), 157.

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.

Nettleton, D., J. G. Hwang, R. A. Caldo, and R. P. Wise (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of agricultural, biological, and environmental statistics 11*(3), 337.

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.

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

Oyeniran, O. and H. Chen (2016). Estimating the proportion of true null hypotheses in multiple testing problems. *J. Probab. Stat. 2016*, Art. ID 3937056, 7.

Phipson, B. (2013). *Empirical Bayes modelling of expression profiles and their associations*. Ph. D. thesis, University of Melbourne, Department of Mathematics and Statistics.

Pounds, S. and C. Cheng (2006). Robust estimation of the false discovery rate. *Bioinformatics 22*(16), 1979–1987.

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.

Singh, A. K., H. Asoh, Y. Takeda, and S. Phillips (2015). Statistical detection of eeg synchrony using empirical bayesian inference. *PloS one 10*(3), e0121795.

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. *AStA Adv. Stat. Anal. 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. *Stat. Probab. Lett. 111*, 32–40.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64*(3), 479–498.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist. 31*(6), 2013–2035.

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.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA 100*(16), 9440–9445.

Tong, T., Z. Feng, J. S. Hilton, and H. Zhao (2013). Estimating the proportion of true null hypotheses using the pattern of observed p-values. *J. Appl. Stat. 40*(9), 1949–1964.

Tong, T. and H. Zhao (2008). Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments. *Stat. Med. 27*(11), 1960–1972.

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.

Beiträge meiner Koautoren

Ich möchte hier gesondert auf die Beiträge meiner Koautoren zu den Manuskripten eingehen. Prof. Dr. Dietmar Pfeifer hat uns freundlicherweise einen Versicherungsdatensatz zum Analysieren in der Section 2.5 zur Verfügung gestellt. Taras Bodnar hat bei der Strukturierung geholfen und allgemein in Diskussionen hilfreiche Ideen und Kritik eingebracht. Zudem hat er die Manuskripte, insbesondere die Beweise und Notationen, Korrektur gelesen. Er verfasste auch in der ursprünglichen Fassung von dem π_0-Manuskript eine Sektion zu elliptischen Copulas. Da sich der Fokus später verändert hat, wurde diese aber wieder entfernt. Mein Betreuer Thorsten Dickhaus hat mir generell sehr geholfen. Konkret hat er die grundlegende Struktur für die Section 2.3, inklusive der Beispiele, verfasst. Weiter hat er zu den Einleitungen und Diskussionen mit Ideen und Referenzen inhaltlich beigetragen. Auch er hat die Manuskripte Korrektur gelesen. Die Beweise, Simulationen und sonstigen Ausarbeitungen stammen von mir.

Danksagung

Ich möchte ganz herzlich Thorsten Dickhaus für die umfangreiche Betreuung in den letzten drei Jahren danken. Für Fragen stand er mit seinem umfangreichem Fachwissen und geeigneten Referenzen jederzeit zur Verfügung. Auch möchte ich meinen Koautoren danken. Insbesondere Taras Bodnar, der sich die Mühe gemacht hat, meine Beweise anzuschauen und sich für Diskussionen die Zeit genommen hat. Weiter danke ich der DFG, die finanziell für meine Stelle aufgekommen ist. In diesem Zusammenhang danke ich auch der von der DFG finanzierten Forschergruppe FOR-1735, deren Mitglieder mich herzlich aufgenommen haben. Abschließend danke ich meinen Kolleginnen und Kollegen, namentlich Natalia Sibirskaya, Rostyslav Bodnar, Jannis Wilken, Jonathan von Schroeder, Magdalena Hernandez, Martina Titze, Werner Brannath, Martin Scharpenberg und meiner Familie, die mich bisher in jeder Lebenssituation tatkräftig unterstützt hat.

Erklärung

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte fremde Hilfe angefertigt. Ich habe keine anderen als die angegebenen Hilfsmittel und Quellen benutzt. Alle aus der Literatur wörtlich oder inhaltlich entnommenen Stellen sind als solche gekennzeichnet.

Bremen, den 22. May 2018 André Neumann