Multiple Testing of One-Sided Hypotheses: Combining Bonferroni and the Bootstrap

Joseph P. Romano[1] and Michael Wolf[2(✉)]

[1] Departments of Economics and Statistics, Stanford University, Stanford, USA
romano@stanford.edu
[2] Department of Economics, University of Zurich, Zurich, Switzerland
michael.wolf@econ.uzh.ch

Abstract. In many multiple testing problems, the individual null hypotheses (i) concern univariate parameters and (ii) are one-sided. In such problems, power gains can be obtained for bootstrap multiple testing procedures in scenarios where some of the parameters are 'deep in the null' by making certain adjustments to the null distribution under which to resample. In this paper, we compare a Bonferroni adjustment that is based on finite-sample considerations with certain 'asymptotic' adjustments previously suggested in the literature.

1 Introduction

Multiple testing refers to any situation that involves the simultaneous testing of several hypotheses. This scenario is quite common in empirical research in just about any field, including economics and finance. Some examples are: one fits a multiple regression model and wishes to decide which coefficients are different from zero; one compares several forecasting strategies to a benchmark and wishes to decide which strategies are outperforming the benchmark; and one evaluates a policy with respect to multiple outcomes and wishes to decide for which outcomes the policy yields significant effects.

If one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses will be rejected by chance alone is generally unduly large. Take the case of $S = 100$ hypotheses being tested at the same time, all of them being true, with the size and level of each test exactly equal to α. For $\alpha = 0.05$, one then expects five true hypotheses to be rejected. Furthermore, if all test statistics are mutually independent, then the probability that at least one true null hypothesis will be rejected is given by $1 - 0.95^{100} \approx 0.994$.

The most common solution to multiple testing problems is to control the *familywise error rate* (FWE), which is defined as the probability of rejecting at least one of the true null hypotheses. In other words, one uses a *global* error rate that combines all tests under consideration instead of an *individual* error rate that only considers one test at a time.

Controlling the FWE at a pre-specified level α corresponds to controlling the probability of a Type I error when carrying out a single test. But this is only one side of the testing problem — and it can be achieved trivially by rejecting

V. Kreinovich et al. (eds.), *Predictive Econometrics and Big Data*, Studies in Computational Intelligence 753, https://doi.org/10.1007/978-3-319-70942-0_4

a particular hypothesis under test with probability α without even looking at data. The other side of the testing problem is 'power', that is, the ability to reject a false null hypothesis.

In this paper, we shall study certain adjustments to 'null sampling distributions' with the hope of power gains in the setting where the individual null hypotheses (i) concern univariate parameters and (ii) are one-sided.

2 Testing Problem

Suppose data X are generated from some unknown probability mechanism $\mathbb{P}$. A model assumes that $\mathbb{P}$ belongs to a certain family of probability distributions, though we make no rigid requirements for this family; it may be a parametric, semiparametric, or nonparametric model.

We consider the following generic multiple testing problem:

$$H_s : \theta_s \leq 0 \quad \text{vs.} \quad H'_s : \theta_s > 0 \qquad \text{for} \quad s = 1, \ldots, S, \tag{1}$$

where the $\theta_s := \theta_s(\mathbb{P})$ are real-valued, univariate parameters and the values under the null hypotheses are always zero without loss of generality. We also denote $\theta := (\theta_1, \ldots, \theta_S)'$.

Remark 1 (Arbitrary Null Parameters). Of course, in practice the values of the parameters under the null hypotheses ("null parameters") may not always be zero. But this situation can easily be handled by our framework as well. To see how, denote the 'original' parameters of interest by γ_s and consider the multiple testing problem

$$H_s : \gamma_s \leq \gamma_{0,s} \quad \text{vs.} \quad H'_s : \gamma_s > \gamma_{0,s} \qquad \text{for} \quad s = 1, \ldots, S, \tag{2}$$

where the null parameters $\gamma_{0,s}$ can take on any value. In such a case, simply define $\theta_s := \gamma_s - \gamma_{0,s}$, for $s = 1, \ldots, S$. □

The familywise error rate (FWE) is defined as

$$\text{FWE}_{\mathbb{P}} := \mathbb{P}\{\text{Reject at least one hypothesis } H_s : \theta_s \leq 0\}.$$

The goal is to control the FWE rate at a pre-specified level α while at the same time to achieve large 'power', which is loosely defined as the ability to reject false null hypotheses, that is, the ability to reject null hypotheses H_s for which $\theta_s > 0$. For example, particular notions of 'power' can be the following:

- The probability of rejecting at least one of the false null hypotheses
- The probability of rejecting a particular false null hypothesis
- The expected number of the false null hypotheses that will be rejected
- The probability of rejecting all false null hypotheses

Control of the FWE means that, for a given significance level α,

$$\text{FWE}_{\mathbb{P}} \leq \alpha \quad \text{for any } \mathbb{P}. \tag{3}$$

Control of the FWE allows one to be $1-\alpha$ confident that there are no false discoveries among the rejected hypotheses.

Control of the FWE is generally equated with 'finite-sample' control: (3) is required to hold for any given sample size n. However, such a requirement can often only be achieved under strict parametric assumptions or for special permutation set-ups. Instead, we settle for *asymptotic* control of the FWE:

$$\limsup_{n\to\infty} \text{FWE}_{\mathbb{P}} \leq \alpha \quad \text{for any } \mathbb{P}. \tag{4}$$

Note here that the statement "for any $\mathbb{P}$" is meant to mean any $\mathbb{P}$ in the underlying assumed model for the family of distributions generating the data; for example, often one would assume the existence of some moments.

3 Multiple Testing Procedures

We assume that individual test statistics are available of the form

$$T_{n,s} := \frac{\hat{\theta}_{n,s}}{\hat{\sigma}_{n,s}},$$

where $\hat{\theta}_{n,s}$ is an estimator of θ_s based on a sample of size n and $\hat{\sigma}_{n,s}$ is a corresponding standard error.[1] We also denote $\hat{\theta}_n := (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,s})'$. We further assume that these test statistics are 'proper' t-statistics in the sense that $T_{n,s}$ converges in distribution to the standard normal distribution under $\theta_s = 0$, for $s = 1, \ldots, S$.

There exists by now a sizeable number of multiple testing procedures (MTPs) designed to control the FWE, at least asymptotically. The oldest and best-known such procedure is the Bonferroni procedure that rejects hypothesis H_s if $\hat{p}_{n,s} \leq \alpha/S$, where $\hat{p}_{n,s}$ is a p-value for H_s. Such a p-value can be obtained via asymptotic approximations or alternatively via resampling methods; for example, an 'asymptotic' p-value is obtained as $\hat{p}_{n,s} := 1-\Phi(T_{n,s})$, where $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution. Although the Bonferroni procedure controls the FWE asymptotically under weak regularity conditions, it is generally suboptimal in terms of 'power'.

There are two main avenues of increasing 'power' while maintaining (asymptotic) control of the FWE. The first avenue, dating back to [Hol79], is to use stepwise procedures where the threshold for rejecting hypotheses becomes less lenient in subsequent steps in case some hypotheses have been rejected in a first step. The second avenue, dating back to [Whi00], at least in nonparametric settings, is to take the dependence structure of the individual test statistics $T_{n,s}$

[1] This means that $\hat{\sigma}_{n,s}$ is an estimator of the standard deviation of $\hat{\theta}_{n,s}$.

into account rather than assuming a 'worst-case' dependence structure as the Bonferroni procedure does; taking this true dependence structure into account — in the absence of strict assumptions — requires the use of resampling methods, such as the bootstrap, subsampling, and permutation methods. [RW05] suggest to combine both avenues, resulting in resampling-based stepwise MTPs.

We start by discussing a bootstrap-based single-step method. An idealized method would reject all H_s for which $T_{n,s} \geq d_1$ where d_1 is the $1-\alpha$ quantile under the true probability mechanism $\mathbb{P}$ of the random variable $\max_s(\hat{\theta}_{n,s} - \theta_s)/\hat{\sigma}_{n,s}$. Naturally, the quantile d_1 not only depends on the marginal distributions of the centered statistics $(\hat{\theta}_{n,s} - \theta_s)/\hat{\sigma}_{n,s}$ but, crucially, also on their dependence structure.

Since the true probability mechanism $\mathbb{P}$ is unknown, the idealized critical value d_1 is not available. But it can be estimated consistently under weak regularity conditions as follows. Take $\hat{d}_1$ as the $1-\alpha$ quantile under $\hat{\mathbb{P}}_n$ of $\max_s(\hat{\theta}^*_{n,s} - \hat{\theta}_{n,s})/\hat{\sigma}^*_{n,s}$. Here, $\hat{\mathbb{P}}_n$ is an *unrestricted* estimate of $\mathbb{P}$. For example, if $X = (X_1, \ldots, X_n)$ with $X_i \overset{\text{iid}}{\sim} \mathbb{P}$, then $\hat{\mathbb{P}}_n$ is typically the empirical distribution of the X_i. Furthermore, $\hat{\theta}^*_{n,s}$ is the estimator of $\hat{\theta}_s$ and $\hat{\sigma}^*_{n,s}$ is the corresponding standard error, both computed from X^* where $X^* \sim \hat{\mathbb{P}}_n$. In other words, we use the bootstrap to estimate d_1. The particular choice of $\hat{\mathbb{P}}_n$ depends on the situation. In particular, if the data are collected over time a suitable time series bootstrap needs to be employed; for example, see [DH97, Lah03].

We have thus described a single-step MTP. However, a stepwise improvement is possible.[2] In any given step j, one simply discards the hypotheses that have been rejected so far and applies the single-step MTP to the remaining universe of non-rejected hypotheses. The resulting critical value $\hat{d}_j$ necessarily satisfies $\hat{d}_j \leq \hat{d}_{j-1}$, and typically satisfies $\hat{d}_j < \hat{d}_{j-1}$, so that new rejections may result; otherwise the method stops with no further rejections.

This bootstrap stepwise MTP provides asymptotic control of the FWE under remarkably weak regularity conditions. Mainly, it is sufficient that (i) $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to a (multivariate) continuous limit distribution and that the bootstrap consistently estimates this limit distribution; and that (ii) the 'scaled' standard errors $\sqrt{n}\hat{\sigma}_{n,s}$ and $\sqrt{n}\hat{\sigma}^*_{n,s}$ converge to the same, non-zero limiting values in probability, both in the 'real world' and in the 'bootstrap world'. Under even weaker regularity conditions, a subsampling approach could be used instead; see [RW05]. Furthermore, when a randomization setup applies, randomization methods can be used as an alternative; see [RW05] again.

4 Adjustments for Power Gains

As stated before, the bootstrap stepwise MTP of the previous section provides asymptotic control of the FWE under weak regularity conditions. But in the one-sided setting (1) considered in this paper, it might be possible to obtain

[2] More precisely, the improvement is a stepdown method.

further power gains by making adjustments for null hypotheses that are 'deep in the null', an idea going back to [Han05].

To motivate such an idea, it is helpful to first point out that for many parameters of interest, θ, there is a one-to-one relation between the bootstrap stepwise MTP of the previous section, which is based on an *unrestricted* estimate $\hat{\mathbb{P}}_n$ of $\mathbb{P}$, and a bootstrap stepwise MTP that is based on a *restricted* estimate $\hat{\mathbb{P}}_{0,n}$ of $\mathbb{P}$, satisfying the constraints of the S null hypotheses. In the latter approach the critical value $\hat{d}_1$ in the first step is obtained as the $1-\alpha$ quantile under $\hat{\mathbb{P}}_{0,n}$ of $\max_s \hat{\theta}^*_{n,s}/\hat{\sigma}^*_{n,s}$. Here, $\hat{\theta}^*_{n,s}$ is the estimator of θ_s and $\hat{\sigma}^*_{n,s}$ is the corresponding standard error, both computed from X^* where $X^* \sim \hat{\mathbb{P}}_{0,n}$. Note that in this latter approach, there is no (explicit) centering in the numerator of the bootstrap test statistics, since the centering already takes place implicitly in the restricted estimator $\hat{\mathbb{P}}_{0,n}$ by incorporating the constraints of the null hypotheses.

For many parameters of interest, the unrestricted bootstrap stepwise MTP of the previous section is equivalent to the restricted bootstrap stepwise MTP of the previous paragraph based on an estimator $\hat{\mathbb{P}}_{0,n}$ that satisfies $\theta_s(\hat{\mathbb{P}}_{0,n}) = 0$ for $s = 1, \ldots, S$. In statistical lingo, such a null parameter $\theta(\hat{\mathbb{P}}_{0,n})$ corresponds to a *least favorable configuration* (LFC), since all the components $\theta_s(\hat{\mathbb{P}}_{0,n})$ lie on the boundary of the respective null hypotheses H_s.

Remark 2 (Example: Testing Means). To provide a specific example of a null-restricted estimator $\hat{\mathbb{P}}_{0,n}$, consider the setting where $X = (X_1, \ldots, X_n)$ with $X_i \overset{\text{iid}}{\sim} \mathbb{P}$, $X_i \in \mathbb{R}^S$, and $(\theta_1, \ldots, \theta_S)' = \theta := \mathbb{E}(X_i)$. Then an unrestricted estimator $\hat{\mathbb{P}}_n$ is given by the empirical distribution of the X_i whereas a null-restricted estimator $\hat{\mathbb{P}}_{0,n}$ is given by the empirical distribution of the $X_i - \hat{\theta}_n$, where $\hat{\theta}_n$ is the sample average of the X_i. In other words, $\hat{\mathbb{P}}_{0,n}$ is obtained by suitably shifting $\hat{\mathbb{P}}_n$ to achieve mean zero for all components. □

[Han05] argues that such an approach is overly conservative when some of the θ_s lie 'deep in the null', that is, for $\theta_s \ll 0$. Indeed, it can easily be shown that asymptotic control of the FWE based on the restricted bootstrap stepwise MTP could be achieved based on an infeasible 'estimator' $\hat{\mathbb{P}}_{0,n}$ that satisfies

$$\theta_s(\hat{\mathbb{P}}_{0,n}) = \min\{\theta_s, 0\}.$$

(We use the term 'estimator' here, since such an $\hat{\mathbb{P}}_{0,n}$ is infeasible in practice because one does not know the true values θ_s.) Clearly, when some of the θ_s are smaller than zero, one would obtain smaller critical values $\hat{d}_j$ in this way compared to using the LFC.

The idea then is to adjust $\hat{\mathbb{P}}_{0,n}$ in a *feasible*, data-dependent fashion such that $\theta_s(\hat{\mathbb{P}}_{0,n}) < 0$ for all θ_s 'deep in the null'.

4.1 Asymptotic Adjustments

Based on the law of iterated logarithm, [Han05] proposes an adjustment $\hat{\mathbb{P}}^{\mathrm{A}}_{0,n}$ that satisfies

$$\theta_s(\hat{\mathbb{P}}^{\mathrm{A}}_{0,n}) := \hat{\theta}_{n,s}\mathbb{1}_{\{T_{n,s}<-\sqrt{2\log\log n}\}}, \tag{5}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function of a set. Therefore, if the t-statistic $T_{n,s}$ is sufficiently small, the parameter of the restricted bootstrap distribution is adjusted to the sample-based estimator $\hat{\theta}_s$, and otherwise it is left unchanged at zero. How one can construct such an estimator $\hat{\mathbb{P}}^{\mathrm{A}}_{0,n}$ depends on the particular application. In the example of Remark 2, say, $\hat{\mathbb{P}}^{\mathrm{A}}_{0,n}$ can be constructed by suitably shifting the empirical distribution $\hat{\mathbb{P}}_n$.

[Han05] only considers a bootstrap single-step MTP. [HHK10] propose the same adjustment (5) in the context of a bootstrap stepwise MTP in the spirit of [RW05].

The adjustment (5) is of asymptotic nature, since one does not have to pay any 'penalty' in the proposals of [Han05, HHK10]. In other words, the MTP procedure proceeds as if $\theta_s(\hat{\mathbb{P}}^{\mathrm{A}}_{0,n}) = \theta_s$ in case $\theta_s(\hat{\mathbb{P}}^{\mathrm{A}}_{0,n})$ has been adjusted to $\hat{\theta}_{n,s} < 0$. The point here is that in finite samples, it may happen that $T_{n,s} < -\sqrt{2\log\log n}$ even though $\theta_s \geq 0$ in reality; in such cases, the null distribution $\hat{\mathbb{P}}_{0,n}$ is generally too 'optimistic' and results in critical values $\hat{d}_j$ that are too small. As a consequence, control of the FWE in finite samples will be negatively affected.

Also note that the cutoff $-\sqrt{2\log\log n}$ is actually quite arbitrary and could be replaced by any multiple of it, however big or small, without affecting the asymptotic validity of the method.

Remark 3 (Related Problem: Testing Moment Inequalities). The literature on moment inequalities is concerned with the related testing problem

$$H : \theta_s \leq 0 \quad \text{for all } s \quad \text{vs.} \quad H' : \theta_s > 0 \quad \text{for at least one } s. \tag{6}$$

This is not a multiple testing problem but the multivariate hypothesis H, which is a single hypothesis, also involves an S-dimensional parameter θ and is one-sided in nature. For this testing problem, [AS10] suggest an adjustment to $\hat{\mathbb{P}}_{0,n}$ that is of asymptotic nature and corresponds to the adjustment of [Han05] for testing problem (1). But then, in a follow-up paper, [AB12] propose an alternative method based on finite-sample considerations that incorporates an explicit 'penalty' for making adjustments to the LFC. The proposal of [AB12] is computationally quite complex and also lacks a rigorous proof of validity. [RSW14] suggest a Bonferroni adjustment as an alternative, which is simpler to implement and also comes with a rigorous proof of validity. □

4.2 Bonferroni Adjustments

We now 'translate' the Bonferroni adjustment of [RSW14] for testing problem (6) to the multiple testing problem (1).

In the first step, we adjust $\hat{\mathbb{P}}_{0,n}$ based on a nominal $1-\beta$ upper rectangular joint confidence region for θ of the form

$$(-\infty, \hat{\theta}_{n,1} + \hat{c}\,\hat{\sigma}_{n,1}] \times \cdots \times (-\infty, \theta_{n,S} + \hat{c}\,\hat{\sigma}_{n,S}]. \tag{7}$$

Here, $0 < \beta < \alpha$ and $\hat{c}$ is a bootstrap-based estimator of the $1-\beta$ quantile of the sampling distribution of the statistic

$$\max_s \frac{\theta_s - \hat{\theta}_{n,s}}{\hat{\sigma}_{n,s}}.$$

For notational compactness, denote the upper end of a generic joint confidence interval in (7) by

$$\hat{u}_{n,s} := \hat{\theta}_{n,s} + \hat{c}\,\hat{\sigma}_{n,s}. \tag{8}$$

Then we propose an adjustment $\hat{\mathbb{P}}^{\mathrm{B}}_{0,n}$ that satisfies

$$\theta_s(\hat{\mathbb{P}}^{\mathrm{B}}_{0,n}) := \min\{\hat{u}_{n,s}, 0\}. \tag{9}$$

How one can construct such an estimator $\hat{\mathbb{P}}^{\mathrm{B}}_{0,n}$ depends on the particular application. In the example of Remark 2, say, $\hat{\mathbb{P}}^{\mathrm{B}}_{0,n}$ can be constructed by suitably shifting the empirical distribution $\hat{\mathbb{P}}_n$.

In the second step, the restricted bootstrap stepwise MTP (i) uses $\theta(\hat{\mathbb{P}}^{\mathrm{B}}_{0,n})$ defined by (9) and (ii) is carried out at nominal level $\alpha - \beta$ as opposed to at nominal level α. Feature (ii) is a finite-sample 'penalty' that accounts for the fact that with probability β, the true θ will not be contained in the joint confidence region (7) in the first step and, consequently, the adjustment in (i) will be overly optimistic.

As reasonable 'generic' choice for β is $\beta := \alpha/10$, as per the suggestion of [RSW14].

It is clear that the Bonferroni adjustment is necessarily less powerful compared to the asymptotic adjustment for two reasons. First, typically $\theta_s(\hat{\mathbb{P}}^{\mathrm{A}}_{0,n}) \leq \theta_s(\hat{\mathbb{P}}^{\mathrm{B}}_{0,n})$ for all $s = 1, \ldots, S$. Second, the asymptotic adjustment uses the full nominal level α in the stepwise MTP whereas the Bonferroni adjustment only uses the reduced level $\alpha - \beta$. On the other hand, it can be expected that the asymptotic adjustments will be liberal in terms of the finite-sample control of the FWE in some scenarios.

4.3 Adjustments for Unrestricted Bootstrap MTPs

We have detailed the asymptotic and Bonferroni adjustments in the context of the restricted bootstrap stepwise MTPs, since they are conceptually somewhat easier to understand.

But needless to say, these adjustments carry over one-to-one to the unrestricted bootstrap stepwise MTPs of [RW05].

Focusing on the first step to be specific, the asymptotic adjustment takes $\hat{d}_1$ as the $1-\alpha$ quantile under $\hat{\mathbb{P}}_n$ of $\max_s(\hat{\theta}^*_{n,s} - \hat{\theta}^{\mathrm{A}}_{n,s})/\hat{\sigma}^*_{n,s}$. Here, $\hat{\mathbb{P}}_n$ is an unrestricted estimator of $\mathbb{P}$ and

$$\hat{\theta}^{\mathrm{A}}_{n,s} := \begin{cases} \hat{\theta}_{n,s} & \text{if } T_{n,s} < -\sqrt{2\log\log n} \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, $\hat{\theta}^*_{n,s}$ and $\hat{\sigma}^*_{n,s}$ are the estimator of θ_s and the corresponding standard error, respectively, computed from X^*, where $X^* \sim \hat{\mathbb{P}}_n$.

On the other hand, the Bonferroni adjustment takes $\hat{d}_1$ as the $1-\alpha+\beta$ quantile under $\hat{\mathbb{P}}_n$ of $\max_s(\hat{\theta}^*_{n,s} - \hat{\theta}^{\mathrm{B}}_{n,s})/\hat{\sigma}^*_{n,s}$. Here, $\hat{\mathbb{P}}_n$ is an unrestricted estimator of $\mathbb{P}$ and

$$\hat{\theta}^{\mathrm{B}}_{n,s} := \hat{\theta}_{n,s} - \min\{\hat{u}_{n,s}, 0\}, \tag{10}$$

with $\hat{u}_{n,s}$ defined as in (8). Furthermore, $\hat{\theta}^*_{n,s}$ and $\hat{\sigma}^*_{n,s}$ are the estimator of θ_s and the corresponding standard error, respectively, computed from X^* where $X^* \sim \hat{\mathbb{P}}_n$.

The computation of the critical constants $\hat{d}_j$ in subsequent steps $j > 1$ is analogous for both adjustments.

Remark 4 (Single Adjustment versus Multiple Adjustments). In principle, the Bonferroni adjustments (10) could be updated in each step of the bootstrap stepwise MTP by updating the joint confidence region for the remaining part of θ in each step, that is, for the elements θ_s of θ for which the corresponding null hypotheses H_s have not been rejected in previous steps. This approach can be expected to lead to small further power gains, though at additional computational (and software coding) costs. □

5 The Gaussian Problem

5.1 Single-Step Method

In this section, we derive an exact finite-sample result for the multivariate Gaussian model, which motivates the method proposed in the paper. Assume that $W := (W_1, \ldots, W_S)' \sim \mathbb{P} \in \mathbf{P} := \{N(\theta, \Sigma) : \mu \in \mathbb{R}^S\}$ for a known covariance matrix Σ. The multiple testing problem consists of S one-sided hypotheses

$$H_s : \theta_s \leq 0 \quad \text{vs.} \quad H'_s : \theta_s > 0 \qquad \text{for} \quad s = 1, \ldots, S. \tag{11}$$

The goal is to control the FWE exactly at nominal level α in this model, for any possible choice of the θ_s, for some pre-specified value of $\alpha \in (0, 1)$. Note further that, because Σ is assumed known, we may assume without loss of generality that its diagonal consists of ones; otherwise, we can simply replace W_s by W_s divided by its standard deviation. This limiting model applies to the nonparametric problem in the large-sample case, since standardized sample

means are asymptotically multivariate Gaussian with a covariance matrix that can be estimated consistently.

First, if instead of the multiple testing problem, we were interested in the single multivariate joint hypothesis that all θ_s satisfy $\theta_s \leq 0$, then we are in the moment inequalities problem; see Remark 3. For such a problem, there are, of course, many ways in which to construct a test that controls size at level α. For instance, given any test statistic $T := T(W_1, \ldots, W_S)$ that is nondecreasing in each of its arguments, we may consider a test that rejects H for large values of T. Note that, for any given fixed critical value c, $\mathbb{P}_\theta\{T(W_1, \ldots, W_S) > c\}$ is a nondecreasing function of each component θ_s in θ. Therefore, if $c := c_{1-\alpha}$ is chosen to satisfy

$$\mathbb{P}_0\{T(W_1, \ldots, W_S) > c_{1-\alpha}\} \leq \alpha,$$

then the test that rejects H_0 when $T > c_{1-\alpha}$ is a level α test. A reasonable choice of test statistic T is the likelihood ratio statistic or the maximum statistic $\max(W_1, \ldots, W_S)$. For this latter choice of test statistic, $c_{1-\alpha}$ may be determined as the $1-\alpha$ quantile of the distribution of $\max(W_1, \ldots, W_S)$ when $(W_1, \ldots, W_S)'$ is multivariate normal with mean 0 and covariance matrix Σ. Unfortunately, as S increases, so does the critical value, which can make it difficult to have any reasonable power against alternatives. The same issue occurs in multiple testing, as described below. The main idea of our procedure is to essentially remove from consideration those θ_s that are 'negative'.[3] If we can eliminate such θ_s from consideration, then we may use a smaller critical value with the hope of increased power against alternatives.

In the multiple testing problem using the max statistic, one could simply reject any θ_s for which $X_s > c_{1-\alpha}$. But as in the single testing problem above, $c_{1-\alpha}$ increases with S and therefore it may be helpful to make certain adjustments if one is fairly confident that a hypothesis H_s satisfies $\theta_s < 0$. Using this reasoning as a motivation, we may use a confidence region to help determine which θ_s are 'negative'. To this end, let $M(1-\beta)$ denote an upper rectangular joint confidence region for θ at level $1-\beta$. Specifically, let

$$\begin{aligned} M(1-\beta) &:= \{\theta \in \mathbb{R}^S : \max_{1 \leq s \leq S}(\theta_s - W_s) \leq K^{-1}(1-\beta)\} \\ &= \{\theta \in \mathbb{R}^S : \theta_s \leq W_s + K^{-1}(1-\beta) \text{ for all } 1 \leq s \leq S\}, \end{aligned} \tag{12}$$

where $K^{-1}(1-\beta)$ is the $1-\beta$ quantile of the distribution (function)

$$K(x) := \mathbb{P}_\theta\{\max_{1 \leq s \leq S}(\theta_s - W_s) \leq x\}.$$

Note that $K(\cdot)$ depends only on the dimension S and the underlying covariance matrix Σ. In particular, it does not depend on the θ_s, so it can be computed under the assumption that all $\theta_s = 0$. By construction, we have for any $\theta \in \mathbb{R}^S$, that

$$\mathbb{P}_\theta\{\theta \in M(1-\beta)\} = 1-\beta.$$

[3] Such a program is carried out in the moment inequality problem by [RSW14].

The idea now is that with probability at least $1-\beta$, we may assume that θ will lie in $\Omega_0 \cap M(1-\beta)$ rather than just in Ω_0, where Ω_0 is the 'negative quadrant' given by $\{\theta:\ \theta_s \le 0,\ s=1,\ldots,S\}$. Instead of computing the critical value under $\theta = 0$, the 'largest' value of θ in Ω_0 (or the value under the LFC), we may therefore compute the critical value under $\tilde{\theta}$, the 'largest' value of θ in the (data-dependent) set $\Omega_0 \cap M(1-\beta)$. It is straightforward to determine $\tilde{\theta}$ explicitly because of the simple shape of the joint confidence region for θ. In particular, $\tilde{\theta}$ has sth component equal to

$$\tilde{\theta}_s := \min\{W_s + K^{-1}(1-\beta), 0\}. \tag{13}$$

But, to account for the fact that θ may not lie in $M(1-\beta)$ with probability β, we reject any H_s for which W_s exceeds the $1-\alpha+\beta$ quantile of the distribution of $T := \max(W_1,\ldots,W_S)$ under $\tilde{\theta}$ rather than the $1-\alpha$ quantile of the distribution of T under $\tilde{\theta}$. The following result establishes that this procedure controls the FWE at level α.

Theorem 1. *Let $T := \max(W_1,\ldots,W_S)$. For $\theta \in \mathbb{R}^S$ and $\gamma \in (0,1)$, define*

$$b(\gamma,\theta) := \inf\{x \in \mathbb{R} : \mathbb{P}_\theta\{T(W_1,\ldots,W_k) \le x\} \ge \gamma\},$$

that is, as the γ quantile of the distribution of T under θ. Fix $0 < \beta < \alpha$. The multiple testing procedure that rejects any H_s for which $W_s > b(1-\alpha+\beta,\tilde{\theta})$ controls the FWE at level α.

Remark 5. As emphasized above, an attractive feature of the procedure is that the 'largest' value of θ in $\Omega_0 \cap M(1-\beta)$ may be determined explicitly. This follows from our particular choice of the initial joint confidence region for θ. If, for example, we had instead chosen $M(1-\beta)$ to be the usual Scheffé confidence ellipsoid, then there may not even be a 'largest' value of θ in $\Omega_0 \cap M(1-\beta)$. □

Proof of Theorem 1. First note that $b(\gamma,\theta)$ is nondecreasing in θ, since T is nondecreasing in its arguments. Fix any θ. Let $I_0 := I_0(\theta)$ denote the indices of true null hypotheses, that is,

$$I_0 := \{s:\ \theta_s \le 0\}.$$

Let $\theta_s^* := \min(\theta_s, 0)$ and let E be the event that $\theta \in M(1-\beta)$. Then, the familywise error rate (FWE) satisfies

$$\begin{aligned}\mathbb{P}_\theta\{\text{reject any true } H_s\} &\le \mathbb{P}_\theta\{E^c\} + \mathbb{P}_\theta\{E \cap \{\text{reject any } H_s \text{ with } s \in I_0\}\} \\ &= \beta + \mathbb{P}_\theta\{E \cap \{\text{reject any } H_s \text{ with } s \in I_0\}\}.\end{aligned}$$

But when the event E occurs and some true H_s is rejected — so that $\max_{s\in I_0} W_s > b(1-\alpha+\beta,\tilde{\theta})$ — then the event $\max_{s\in I_0} W_s > b(1-\alpha+\beta,\theta^*)$

must occur, since $b(1-\alpha+\beta,\theta)$ is nondecreasing in θ and $\theta \leq \tilde{\theta}$ when E occurs. Hence, the FWE is bounded above by

$$\beta + \mathbb{P}_{\theta}\{\max_{s\in I_0} W_s > b(1-\alpha+\beta,\theta^*)\} \leq \beta + \mathbb{P}_{\theta^*}\{\max_{s\in I_0} W_s > b(1-\alpha+\beta,\theta^*)\}$$

because the distribution of $\max_{s\in I_0} W_s$ only depends on those θ_s in I_0. Therefore, the last expression is bounded above by

$$\beta + \mathbb{P}_{\theta^*}\{\max_{\text{all } s} W_s > b(1-\alpha+\beta,\theta^*)\} = \beta + 1 - (1-\alpha+\beta) = \beta + (\alpha-\beta) = \alpha.$$

□

5.2 Stepwise Method

One can improve upon the single-step method in Theorem 1 by a stepwise method.[4] More specifically, consider the following method. Begin with the method described above, which rejects any H_s for which $W_s > b(1-\alpha+\beta,\tilde{\theta})$. Basically, one applies the closure method to the above and show that it may be computed in a stepwise fashion. To do this, we first need to describe the situation when testing only a subset of the hypotheses. So, let I denote any subset of $\{1,\ldots,S\}$ and let $b_I(\gamma,\theta)$ denote the γ quantile of the distribution of $\max(T_s : s \in I)$ under θ. Also, let $\tilde{\theta}(I) := \{\tilde{\theta}_s(I) : s \in I\}$ with $\tilde{\theta}_s(I)$ be defined as in (13) except that $K^{-1}(1-\beta)$ is replaced by $K_I^{-1}(1-\beta)$, defined to be the $1-\beta$ quantile of the distribution (function)

$$K_I(x) := \mathbb{P}_{\theta}\{\max_{s\in I}(\theta_s - W_s) \leq x\}.$$

The stepwise method can now be described. Begin by testing all H_s with $s \in \{1,\ldots,S\}$ as described in the single-step method. If there are any rejections, remove the rejected hypotheses from consideration and apply the single-step method to the remaining hypotheses. That is, if I is the set of indices of the remaining hypotheses not previously rejected, then reject any such H_s if $W_s > b_I(1-\alpha+\gamma,\tilde{\theta}(I))$. And so on. (Note that at each step of the procedure, a new joint confidence region is computed to determine $\tilde{\theta}(I)$, but β remains the same in each step.)

Theorem 2. *Under the Gaussian setup of Theorem 1, the above stepwise method controls the FWE at level α.*

Proof of Theorem 2. We just need to show that the closure method applied to the above tests results in the stepwise method as described. To do this, it suffices to show that if H_s is rejected by the stepwise method, $s \in I$, and $I \subset J$, then when J is tested (meaning the H_s with $s \in J$ are jointly tested) and the method rejects the joint (intersection) hypothesis, then it also rejects the particular joint (intersection) hypothesis when just I is tested.

[4] More precisely, the improvement is a stepdown method.

First, the distribution of $\max_{\theta_s \in I} W_s$ is stochastically dominated by that of $\max_{\theta_s \in J} W_s$ (since we are just taking the max over a larger set), under any θ and in particular under $\tilde{\theta}(I)$. But the distribution of the maximum statistic $\max_{\theta_s \in J} W_s$ is monotone increasing with respect to θ_s because of the important fact that, component wise,

$$\tilde{\theta}(I) \leq \tilde{\theta}(J).$$

Hence, the distribution of $\max_{\theta_s \in J}$ under $\tilde{\theta}(I)$ is further dominated by the distribution of $\max_{\theta_s \in J} W_s$ under $\tilde{\theta}(J)$. Therefore, the critical values satisfy

$$b_I(1 - \alpha + \beta, \tilde{\theta}(I)) \leq b_J(1 - \alpha + \beta, \tilde{\theta}(J)),$$

which is all we need to show, since then any H_s for which W_s exceeds $b_J(1 - \alpha + \beta, \tilde{\theta}(J))$ will satisfy that W_s also exceeds $b_I(1 - \alpha + \beta, \tilde{\theta}(I))$. □

6 Monte Carlo Simulations

The data are of the form $X := (X_1, X_2, \ldots, X_n)$ with $X_i \overset{\text{iid}}{\sim} N(\theta, \Sigma)$, $\theta \in \mathbb{R}^S$, and $\Sigma \in \mathbb{R}^{S \times S}$. We consider $n = 50, 100$.

For $n = 50$, we consider $S = 25, 50, 100$ and the following mean vectors $\theta = (\theta_1, \ldots, \theta_S)'$:

- All $\theta_s = 0$
- Five of the $\theta_s = 0.4$
- Five of the $\theta_s = 0.4$ and $S/2$ of the $\theta_s = -0.4$
- Five of the $\theta_s = 0.4$ and $S/2$ of the $\theta_s = -0.8$

For $n = 100$, we consider $S = 50, 100, 200$ and the following mean vectors $\theta = (\theta_1, \ldots, \theta_S)'$:

- All $\theta_s = 0$
- Ten of the $\theta_s = 0.3$
- Ten of the $\theta_s = 0.3$ and $S/2$ of the $\theta_s = -0.3$
- Ten of the $\theta_s = 0.3$ and $S/2$ of the $\theta_s = -0.6$

For $S = 50, 100$, the covariance matrix Σ is always a constant-correlation matrix with constant variance one on the diagonal and constant covariance $\rho = 0$, 0.5 on the off-diagonal.

The test statistics $T_{n,s}$ are the usual t-statistics based on the individual sample means and sample standard deviations.

The multiple testing procedure is always the bootstrap stepwise MTP of [RW05] and we consider three variants:

- **LFC:** No adjustment at all
- **Asy:** Asymptotic Adjustment
- **Bon:** Bonferroni Adjustment

Note that for computational simplicity, Bon is based on a single adjustment throughout the stepwise MTP; see Remark 4.

The nominal level for FWE control is $\alpha = 10\%$ and the value of β for the Bonferroni adjustment is chosen as $\beta = 1\%$ following the 'generic' suggestion $\beta := \alpha/10$ of [RSW14].

We consider two performance measures:

- **FWE:** Empirical FWE
- **Power:** Average number of rejected false hypotheses

The number of Monte Carlo repetitions is $B = 50,000$ in each scenario and the bootstrap à la [Efr79] is based on 1,000 resamples always.

The results for $n = 50$ are presented in Sect. A.1 and the results for $n = 100$ are presented in Sect. A.2. They can be summarized as follows.

- As pointed out before, Asy is always more powerful than Bon necessarily.
- There are some scenarios where Asy fails to control the FWE, though the failures are never grave: In the worst case, the empirical FWE is 10.6%.
- Bon can actually be less powerful than LFC (though never by much). This is not surprising: When null parameters are on the boundary or close to the boundary, then the 'minor' adjustment in the first stage of Bon does not offset the reduction in the nominal level (from α to $\alpha - \beta$) in the second stage.
- When null parameters are 'deep in the null', also the power gains of Bon over LFC are noticeable (though never quite as large as the power gains of Asy over LFC). Of course, such power gains would even be greater by increasing the proportion of null parameters 'deep in the null' and/or the distance away from zero of such null parameters.

7 Conclusion

In many multiple testing problems, the individual null hypotheses (i) concern univariate parameters and (ii) are one-sided. In such problems, power gains can be obtained for bootstrap multiple testing procedures in scenarios where some of the parameters are 'deep in the null' by making certain adjustment to the null distribution under which to resample. In this paper we have compared a Bonferroni adjustment that is based on finite-sample considerations to certain 'asymptotic' adjustments previously suggested in the literature. The advantage of the Bonferroni adjustment is that it guarantees better finite-sample control of the familywise error rate. The disadvantage is that it is always somewhat less powerful than the asymptotic adjustments.

A Detailed Monte Carlo Results

A.1 Results for $n = 50$

See Tables 1, 2, 3 and 4.

Table 1. All $\theta_s = 0$: FWE.

S	LFC	Asy	Bon
$\rho = 0$			
25	9.8	10.5	8.8
50	9.7	10.2	8.7
100	9.5	10.1	8.5
$\rho = 0.5$			
25	10.0	10.0	9.0
50	9.8	9.8	8.8
100	9.9	9.9	8.9

Table 2. Five of the $\theta_s = 0.4$: FWE | Power.

S	LFC	Asy	Bon	LFC	Asy	Bon
$\rho = 0$						
25	8.9	9.4	7.9	2.7	2.8	2.6
50	9.1	9.6	8.2	2.2	2.2	2.1
100	9.4	10.0	8.4	1.7	1.8	1.6
$\rho = 0.5$						
25	9.9	9.9	8.9	3.2	3.2	3.1
50	9.8	9.8	8.8	2.8	2.8	2.7
100	10.0	10.1	9.2	2.4	2.4	2.3

Table 3. Five of the $\theta_s = 0.4$ and $S/2$ of the $\theta_s = -0.4$: FWE | Power.

S	LFC	Asy	Bon	LFC	Asy	Bon
$\rho = 0$						
25	3.7	7.7	3.6	2.7	3.3	2.7
50	4.2	8.2	4.0	2.2	2.7	2.2
100	4.7	8.8	4.4	1.7	2.1	1.7
$\rho = 0.5$						
25	5.3	7.6	4.7	3.1	3.6	3.2
50	5.9	8.0	5.3	2.8	3.2	2.7
100	6.6	8.4	5.9	2.4	2.8	2.3

Table 4. Five of the $\theta_s = 0.4$ and $S/2$ of the $\theta_s = -0.8$: FWE | Power.

S	LFC	Asy	Bon	LFC	Asy	Bon
$\rho = 0$						
25	3.7	8.8	6.9	2.7	3.4	3.2
50	4.2	9.3	7.1	2.2	2.8	2.6
100	4.7	9.8	7.4	1.7	2.2	2.0
$\rho = 0.5$						
25	5.3	9.8	7.8	3.1	3.7	3.5
50	5.9	9.8	7.7	2.8	3.2	3.1
100	6.6	10.0	7.8	2.4	2.8	2.6

A.2 Results for $n = 100$

See Tables 5, 6, 7 and 8.

Table 5. All $\theta_s = 0$: FWE.

S	LFC	Asy	Bon
$\rho = 0$			
50	9.8	10.2	8.8
100	10.0	10.4	8.9
200	10.0	10.6	8.9
$\rho = 0.5$			
50	10.0	10.0	9.0
100	10.1	10.1	9.0
200	10.1	10.1	9.1

Table 6. Ten of the $\theta_s = 0.3$: FWE | Power.

S	LFC	Asy	Bon	LFC	Asy	Bon
$\rho = 0$						
50	8.8	9.1	7.9	5.4	5.5	5.3
100	9.5	9.8	8.5	4.5	4.5	4.3
200	9.7	10.2	8.7	3.6	3.7	3.5
$\rho = 0.5$						
50	9.9	9.9	8.9	6.5	6.5	6.3
100	10.0	10.0	9.0	5.8	5.8	5.6
200	10.1	10.1	9.1	5.1	5.1	4.9

Table 7. Ten of the $\theta_s = 0.3$ and $S/2$ of the $\theta_s = -0.3$: FWE | Power.

S	LFC	Asy	Bon	LFC	Asy	Bon
$\rho = 0$						
50	3.3	7.3	3.4	5.4	6.4	5.4
100	4.3	8.4	4.1	4.5	5.3	4.4
200	4.7	8.9	4.4	3.6	4.4	3.6
$\rho = 0.5$						
50	5.3	7.7	4.7	6.5	7.3	6.5
100	6.2	8.4	5.6	5.8	6.5	5.7
200	6.9	8.8	6.1	5.1	5.7	5.0

Table 8. Ten of the $\theta_s = 0.3$ and $S/2$ of the $\theta_s = -0.6$: FWE | Power.

S	LFC	Asy	Bon	LFC	Asy	Bon
$\rho = 0$						
50	3.4	8.4	6.9	5.4	6.6	6.4
100	4.3	9.4	7.7	4.5	5.5	5.2
200	4.7	9.9	7.9	3.6	4.5	4.3
$\rho = 0.5$						
50	5.3	9.9	8.3	6.5	7.4	7.1
100	6.2	10.0	8.4	5.8	6.5	6.3
200	6.9	10.2	8.6	5.1	5.9	5.6

References

[AB12] Andrews, D.W.K., Barwick, P.J.: Inference for parameters defined by moment inequalities: a recommended moment selection procedure. Econometrica **80**(6), 2805–2826 (2012)

[AS10] Andrews, D.W.K., Soares, G.: Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica **78**(1), 119–157 (2010)

[DH97] Davison, A.C., Hinkley, D.V.: Bootstrap Methods and their Application. Cambridge University Press, Cambridge (1997)

[Efr79] Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**, 1–26 (1979)

[Han05] Hansen, P.R.: A test for superior predictive ability. J. Bus. Econ. Stat. **23**, 365–380 (2005)

[HHK10] Hsu, P.-H., Hsu, Y.-C., Kuan, C.-M.: Testing the predictive ability of technical analysis using a new stepwise test with data snooping bias. J. Empir. Finance **17**, 471–484 (2010)

[Hol79] Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)

[Lah03] Lahiri, S.N.: Resampling Methods for Dependent Data. Springer, New York (2003)
[RSW14] Romano, J.P., Shaikh, A.M., Wolf, M.: A practical two-step method for testing moment inequalities. Econometrica **82**(5), 1979–2002 (2014)
[RW05] Romano, J.P., Wolf, M.: Exact and approximate stepdown methods for multiple hypothesis testing. J. Am. Stat. Assoc. **100**(469), 94–108 (2005)
[Whi00] White, H.L.: A reality check for data snooping. Econometrica **68**(5), 1097–1126 (2000)