

# Multiple testing when many $p$ -values are uniformly conservative, with application to testing qualitative interaction in educational interventions

Qingyuan Zhao<sup>1</sup>, Dylan S. Small, Weijie Su

University of Pennsylvania, Philadelphia

**Abstract.** In the evaluation of treatment effects, it is of major policy interest to know if the treatment is beneficial for some and harmful for others, a phenomenon known as qualitative interaction. We formulate this question as a multiple testing problem with many conservative null  $p$ -values, in which the classical multiple testing methods may lose power substantially. We propose a simple technique—conditioning—to improve the power. A crucial assumption we need is uniform conservativeness, meaning for any conservative  $p$ -value  $p$ , the conditional distribution  $(p/\tau) | p \leq \tau$  is stochastically larger than the uniform distribution on  $(0, 1)$  for any  $\tau$ . We show this property holds for one-sided tests in a one-dimensional exponential family (e.g. testing for qualitative interaction) as well as testing  $|\mu| \leq \eta$  using a statistic  $X \sim N(\mu, 1)$  (e.g. testing for practical importance with threshold  $\eta$ ). We propose an adaptive method to select the threshold  $\tau$ . Our theoretical and simulation results suggest the proposed tests gain significant power when many  $p$ -values are uniformly conservative and lose little power when no  $p$ -value is uniformly conservative. We apply our method to two educational intervention datasets.

**Keywords:** Global null; Meta-analysis; Multisite study; Selective inference; Treatment effect heterogeneity; Uniform conditional stochastic order.

---

<sup>1</sup>*Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: qyzhao@wharton.upenn.edu. 23 August 2017.

## 1 Introduction

To make the most informed policy decisions from a randomized experiment or an observational study, it is often important to understand variation in treatment effects [Wang et al., 2007, Schochet et al., 2014]. For example, an influential framework in education called Aptitude-Treatment Interaction [Cronbach and Snow, 1977] is based on the claim that different learners may benefit from different styles of instruction. There are also numerous examples in medical studies. Pizzocaro et al. [2001] found that a drug class called interferon increases the survival probability of patients in the late stage of renal cell carcinoma, but is harmful for patients in the early stage. In the evaluation of a randomized trial [CRASH-2-Collaborators, 2011], the collaboration team found that tranexamic acid, the drug under study, reduces the risk of bleeding to death when used within 3 hours from trauma injury, but increases the risk when used after 3 hours.

Typically, the heterogeneous nature of treatment effect is examined by subgroup analysis, where study participants are grouped by some baseline covariates measured prior to the treatment. We will use the terms “treatment effect heterogeneity” and “treatment interaction” interchangeably because both of them means there is a noticeable interaction between the treatment and the subgroup indicator when modeling the outcome. There are two types of treatment effect heterogeneity: the *qualitative* or *disordinal* interaction, where there exists one subgroup whom the treatment benefits and another subgroup whom the treatment harms, and the *quantitative* or *ordinal* interaction, where the subgroup treatment effects may have different magnitude but the same sign [Gail and Simon, 1985]. A good example of qualitative interaction is the study of interferon described in the last paragraph. There is also interest in understanding treatment effects variation across sites in multisite studies or across studies in meta-analysis [Bloom et al., 2017].

### 1.1 Motivating applications

In this paper we will consider the problem of testing qualitative interaction. Since qualitative interactions imply the optimal treatment rule must be personalized, they usually have much more policy or clinical significance than non-qualitative interactions. To motivate this, we first introduce two examples from education. In the first example, we consider the effect of modified school calendars. Instead of following the more traditional school calendar with a long summer break (in addition to a short winter and spring break), some schools have switched to a modified school calendar comprising more frequent but shorter

intermittent breaks (e.g., 9 weeks of school followed by 3 weeks off), while keeping the total number of days at school approximately the same. Cooper et al. [2003] investigated the effect of modified school calendars on student achievement using studies of 55 schools in 11 districts. Using the published dataset in Konstantopoulos [2011], we summarize their main results in Figure 1a.

In the second example, we consider the effectiveness of writing-to-learn interventions, in which students receive instruction with increased emphasis on writing tasks compared to conventional instruction. The outcome of interest is the academic achievement of the students (for example some exam score). Figure 1b summarizes the results of a meta-analysis of 48 studies by Bangert-Drowns et al. [2004].

In both examples, the treatment has an overall significantly positive effect when fitting a random effect model (the “RE Model” row in Figure 1). However, for the modified school calendar, a multi-level analysis in Konstantopoulos [2011] showed significant heterogeneity in the treatment effect; see also Section 6. Furthermore, in both examples there is at least one significantly negative finding (without correcting for multiple testing) in the forest plot. Therefore, it is interesting to know if qualitative interaction exists, that is, if for some cohorts the treatment effect is indeed negative.

## 1.2 Formulation of testing qualitative interaction as a multiple testing problem

Next we formulate this question as a multiple testing problem. Suppose we want to analyze a randomized experiment or an observational study in which there are  $n$  subgroups or independent studies and the observed treatment effect in study  $i$  is distributed as  $N(\mu_i, \sigma_i^2)$ . Then the null hypothesis of no qualitative interaction is the union of two hypotheses,  $H_0 = H_0^+ \cup H_0^-$ , where  $H_0^+ = \{\mu_j \geq 0, \forall i\}$  and  $H_0^- = \{\mu_i \leq 0, \forall i\}$ . The one-sided hypothesis  $H_0^+$  is the intersection of  $n$  individual hypotheses,  $H_0^+ = \cap_{i=1}^n H_{0i}^+$  where  $H_{0i}^+ = \{\mu_i \geq 0\}$ , and is called a *global null hypothesis* because  $H_0^+$  is correct if and only if all the individual hypotheses are correct. To test the null hypothesis  $H_0$  at level  $\alpha$ , one can simply test both  $H_0^+$  and  $H_0^-$  and reject  $H_0$  if both  $H_0^+$  and  $H_0^-$  are rejected at level  $\alpha$  (because  $H_0$  is the *union* hypothesis, we do not need to correct for multiplicity here).

Our statistical methodology only requires  $p$ -values for the individual hypotheses,  $p_i$ ,  $i = 1, \dots, n$ . In the classical setting of a multiple testing problem, the  $p$ -values are assumed to either follow the uniform distribution or be stochastically larger than the uniform distribution over  $(0, 1)$  under the null, or stochastically smaller than the uniform distri-

bution under the alternative. Many effective procedures have been proposed and some have certain theoretical optimality guarantees especially if the null  $p$ -values are exactly uniformly distributed. Some distinguished examples of testing the global null hypothesis include Bonferroni’s correction, Fisher’s combination, and Tukey’s higher criticism, which are reviewed in Section 2.1. When the global null is rejected, usually we are further interested in knowing which individual hypotheses (namely  $H_{0i}^+$  and  $H_{0i}^-$ ) are false. Subsequent methods such as Holm [1979]’s procedure or Benjamini and Hochberg [1995]’s procedure can be used to control the family-wise error rate or the false discovery rate.

However, in the motivating examples above as well as many other applications, it is common that the majority of the null  $p$ -values may be very conservative (stochastically larger than the uniform distribution). For example, for testing  $H_{0i}^+$ , if the individual effect  $\mu_i$  is positive, then the corresponding  $p$ -value will be conservative. In both motivating examples, many of the estimated individual effects are positive as shown in Figure 1, and it seems plausible that many of the effects are truly positive which would result in conservative  $p$ -values for testing  $H_0^+$ . Ideally, if we knew in priori that some studies have positive effects, we would like to exclude them when testing  $H_0^+$ .

### 1.3 Overview of our approach

In this paper we propose a simple technique—conditioning—that can be used in conjunction with all the existing multiple testing procedures to improve power when many null  $p$ -values are conservative. This avoids paying an unnecessary price of adjusting for multiplicity for the conservative tests (see Section 2.2 for some heuristics). The proposed method can be applied to any global and simultaneous testing problem if the following assumptions are satisfied: 1. the tests are independent; 2. the  $p$ -values are *uniformly valid* (meaning the null  $p$ -values are still valid after conditioning). When these two assumptions hold, we show, by providing theoretical results and extensive simulation studies, that the conditional tests reduce little power in the classical non-conservative scenario and greatly increase power in the conservative scenario (including testing qualitative interaction). Detailed proofs of the claims in this paper can be found in the appendix.

### 1.4 Uniform validity and its applications

Among the two key assumptions, the first independence assumption can be relaxed (Section 4.2), while we show the second assumption (uniform validity) holds for one-sided tests

in a family of distributions that has monotone likelihood ratio (MLR). This includes one-sided tests in any one-dimensional exponential family and hence includes the motivating problem of testing qualitative interactions. The location family of folded normal distribution also has MLR (see Appendix A.2). This leads to another application of the proposed conditional test—testing for practical importance. In many problems, rejecting small deviations from no effect is often practically inconsequential and instead we would like to test whether there is a practically meaningful difference from no effect. Sun and McLain [2012] formulated this problem as a two-sided normal means problem. Let  $X_i \sim N(\mu_i, \sigma_i^2)$  with known  $\sigma_i^2$ . The  $i$ -th null hypothesis is  $H_{0i} : |\mu_i| \leq \eta$  where  $\eta$  is the practical importance threshold. This can be viewed as a one-sided testing problem in the folded normal distribution (if  $X \sim N(\mu, \sigma^2)$ , then  $|X|$  is said to have a folded normal distribution with parameters  $\mu$  and  $\sigma^2$ ). Since the location family of folded normal distributions (fixed  $\sigma^2$ ) has MLR. Therefore, our conditional test is also valid for the two-sided normal means problem.

Although uniform validity hold for many hypothesis testing problems as illustrated above, it does not hold in all circumstances. We refer the reader to Section 7.2 for an example and alternative approaches.

## 2 Multiple testing in presence of conservative tests

Suppose we have  $n$   $p$ -values,  $p_i, i = 1, \dots, n$  for  $n$  hypotheses,  $H_{0i}, i = 1, \dots, n$ . We assume that every null  $p$ -value  $p_i$  is individually *valid*, meaning  $P(p_i \leq q) \leq q$  for all  $0 \leq q \leq 1$  if  $H_{0i}$  is true. Furthermore, we call  $p_i$  *exact* if  $P(p_i \leq q) = q$  for all  $0 \leq q \leq 1$ , and *conservative* if it is valid and  $P(p_i \leq q) < q$  for some  $0 < q < 1$ . Geometrically if we plot the cumulative distribution function (CDF) of a  $p$ -value, an exact  $p$ -value has a CDF that is exactly the 45 degree diagonal line (the CDF of the uniform distribution) and a conservative  $p$ -value CDF is below the 45 degree line in at least one point.

### 2.1 Previous methods for testing the global null

We will start with testing the global null hypothesis  $H_0 = \bigcap_{i=1}^n H_{0i}$  that all the individual hypotheses are true before moving into other objectives (controlling family-wise error rate or false discovery rate). Let's first review some classical methods to test the global null. Given a significance level  $0 < \alpha < 1$ , one of the simplest and most widely used methods is Bonferroni's correction, which rejects  $H_0$  if any of the  $p_i$  is less than  $\alpha/n$ . When testing the “needle in a haystack problem” (only one non-null), this method is asymptotically optimal

[Arias-Castro et al., 2011]. Simes [1986] proposed an improved Bonferroni procedure that rejects  $H_0$  if  $p_{(i)} \leq (i/n)\alpha$  for some  $1 \leq i \leq n$ , where  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$  are the ordered  $p$ -values. Notice that Bonferroni’s and Simes’ procedures control type I error for testing  $H_0$  even if the  $p$ -values are not independent.

When the  $p$ -values are independent, another commonly used method is Fisher’s combination probability test, which combines the  $p$ -values into one test statistic  $T = -2 \sum_{i=1}^n \log p_i$ . Fisher [1925] showed that  $T$  has a  $\chi^2$  distribution with  $2n$  degrees of freedom when all the  $p$ -values are uniformly distributed. When some hypotheses are false the corresponding  $p_i$  tend to be small, so the test statistic will be large. A similar method is the truncated product of Zaykin et al. [2002], whose combined test statistic is  $T' = -2 \sum_{i=1}^n (\log p_i) 1_{\{p_i \leq \tau\}}$ , where  $\tau$  is a truncation threshold between 0 and 1.

A third type of method compares the empirical distribution of  $p_1, \dots, p_n$  with the uniform distribution. An interesting representative is Tukey’s higher criticism or second-level significance test, which examines if there is an excessive number of significant tests (e.g., tests with  $p$ -values less than 0.05). Donoho and Jin [2004] considered a modified statistic:

$$HC^* = \max_{0 < q \leq \tau} \frac{n^{-1} \sum_{i=1}^n 1_{\{p_i \leq q\}} - q}{\sqrt{q(1-q)/n}}.$$

Here  $n^{-1} \sum_{i=1}^n 1_{\{p_i \leq q\}}$  and  $q$  are the observed and expected fractions of tests significant at level  $q$ , and  $\sqrt{q(1-q)/n}$  is the variance of the observed fraction under  $H_0$ . Donoho and Jin [2004] showed that this test is very effective at solving the “sparse normal means” problem in certain asymptotic regime.

## 2.2 Problem of conservative tests

This paper considers the situation that the  $p$ -values are independent but many of them are conservative. As an example, consider the one-sided testing problem of normal means. Let  $Y_i, i = 1, \dots, n$  be independent normal variables with mean  $\mu_i$  and variance 1. The null hypothesis  $H_{0i}$  is  $\mu_i \leq 0$ , so the global null hypothesis is  $H_0 : \mu_i \leq 0, \forall i$ . The individual  $p$ -values are given by  $p_i = 1 - \Phi(Y_i)$  where  $\Phi$  is the CDF of the standard normal distribution. The  $p$ -value  $p_i$  has a uniform distribution if  $\mu_i = 0$  and  $p_i$  is conservative if  $\mu_i < 0$ .

When some  $p$ -values are conservative, the classical methods in Section 2.1 usually lose power. Consider a rather extreme example in which the  $p$ -values are generated from the one-sided normal means problem described in the last paragraph with  $n = 100, \mu_1 = \mu_2 = 3$

	Bonferroni	Fisher	Truncated Product
Classical	0.1	1	0.999
Conditional ( $\tau = 0.5$ )	0.004	$5.4 \times 10^{-5}$	$4.72 \times 10^{-5}$

Table 1: Combined  $p$ -values in the hypothetical example in Section 2.2 that  $(p_1, p_2, p_3 \dots, p_{100}) = (0.001, 0.001, 1, \dots, 1)$ .

and  $\mu_3 = \dots = \mu_{100} = -10$ . For simplicity, let's say the observed statistics are just  $Y_i = \mu_i$ , so the  $p$ -values are  $(p_1, p_2, p_3 \dots, p_{100}) \approx (0.001, 0.001, 1, \dots, 1)$ . Although the first two  $p$ -values are highly significant, the classical methods do not find the whole set of  $p$ -values providing enough evidence to reject the global null hypothesis, because they “over-correct” for multiplicity. This is demonstrated in the first row of Table 1. In contrast, using the conditional  $p$ -values proposed in Section 3, the same tests all reject the global null hypothesis.

Intuitively, if we do observe  $(p_1, p_2, p_3 \dots, p_{100}) = (0.001, 0.001, 1, \dots, 1)$ , the first thing to be noticed is there are exceptionally many large  $p$ -values. This indicates many conservative tests. Naturally, we would like to “ignore” these large  $p$ -values and only use the two smaller ones, with which we can easily reject the global null. However, we cannot simply remove the large  $p$ -values because this would be data snooping and make the subsequent inference invalid.

The truncated product method of Zaykin et al. [2002] attempts to address this problem by only multiplying the  $p$ -values below some threshold  $\tau$ . This improves Fisher's combination test when some  $p$ -values are conservative. However, it does not completely resolve the problem, because the null distribution is still computed assuming all  $p$ -values are uniformly distributed even though we have overwhelming evidence that many of them are conservative.

### 3 Conditional test

#### 3.1 Testing the global null

In light of the discussion above, we propose a simple conditional test. Given independent  $p$ -values  $p_1, p_2, \dots, p_n$  and a fixed threshold parameter  $0 < \tau \leq 1$ , let  $\mathcal{S}_\tau = \{i, p_i \leq \tau\}$  be the indices where the  $p$ -values are less than  $\tau$ . When  $p_i$  is exact, it is uniformly distributed on  $[0, \tau]$  given  $p_i \leq \tau$ . In other words, the conditional distribution  $p_i/\tau | \mathcal{S}_\tau \stackrel{i.i.d.}{\sim} \text{Unif}[0, 1]$  if

$p_i$  is exact.

Our proposal is to use any of the global testing methods in Section 2.1 on the set of  $p$ -values  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$  (we assume the combined  $p$ -value is 1 if  $\mathcal{S}_\tau$  is empty). Intuitively, this screens out the very large  $p$ -values in the example in the last section. For example, the conditional Bonferroni test rejects the global null  $H_0$  if the cardinality of  $\mathcal{S}_\tau$  is positive,  $|\mathcal{S}_\tau| > 0$ , and

$$p^{\text{CB}}(p_1, \dots, p_n; \tau) = \left( \min_{1 \leq i \leq n} p_i/\tau \right) \cdot |\mathcal{S}_\tau| \leq \alpha. \quad (1)$$

Notice that  $\tau = 1$  reduces to the original Bonferroni correction since  $|\mathcal{S}_1| = n$ . The conditional Fisher test uses the statistic

$$T^{\text{CF}}(p_1, \dots, p_n; \tau) = -2 \sum_{i=1}^n [\log(p_i/\tau)] \cdot 1_{\{p_i \leq \tau\}}$$

and compares it with the  $\chi^2$  distribution with  $|\mathcal{S}_\tau|$  degrees of freedom (if  $|\mathcal{S}_\tau| > 0$ ); denote the combined  $p$ -value by  $p^{\text{CF}}(p_1, \dots, p_n; \tau)$ . Notice that this test statistic is very similar to the truncated product of Zaykin et al. [2002] except we also divide the  $p$ -values by the threshold  $\tau$ . This allows us to work with  $|\mathcal{S}_\tau|$  instead of  $n$  many  $p$ -values. When many  $p$ -values are conservative,  $|\mathcal{S}_\tau|$  can be substantially smaller than  $n\tau$ , the expected value of  $|\mathcal{S}_\tau|$  when no  $p$ -value is conservative. The simulation and real data examples in Sections 5 and 6 show the conditional tests can be much more powerful than the unconditional tests ( $\tau = 1$ ).

However, the conditional tests are not valid without making further assumptions. Heuristically, we need the transformed  $p$ -values  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$  to be valid. Although the transformed  $p$ -values are exact if the original  $p$ -values are exact, the same conclusion does not in general hold for conservative  $p$ -values. Next we introduce a stronger notion of conservativeness:

**Definition 1.** A valid  $p$ -value  $p_i$  is called *uniformly valid* if for all  $0 < \tau < 1$  such that  $P(p_i \leq \tau) > 0$ ,  $p_i/\tau$  given  $p_i \leq \tau$  is valid. A  $p$ -value is called *uniformly conservative* if it is conservative and uniformly valid.

**Proposition 1.** *The conditional test with any fixed  $0 < \tau \leq 1$  (any global test on  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$ ) controls type I error at the nominal level if  $p_1, p_2, \dots, p_n$  are independent and uniformly valid.*

The proof of Proposition 1 immediately follows from Definition 1 and the validity of

the global test. Next we examine which tests are uniformly valid/conservative. Let  $F_i$  be the CDF of  $p_i$ ,  $F_i(x) = \mathbb{P}(p_i \leq x)$ . By Definition 1, the  $p_i$  is uniformly conservative if and only if

$$F_i(\tau x) \leq xF_i(\tau), \quad \forall 0 \leq x, \tau \leq 1.$$

Geometrically, this means that the function  $F_i(x)$  is always below the segment from  $(0, 0 = F_i(0))$  to  $(\tau, F_i(\tau))$  if  $0 \leq x \leq \tau$ . Therefore, a sufficient condition for uniform conservativeness is convexity of the CDF, since by convexity,

$$F_i(\tau x) = F_i((1-x) \cdot 0 + x \cdot \tau) \leq (1-x)F_i(0) + xF_i(\tau) = xF_i(\tau).$$

However, as pointed out by an anonymous reviewer, convexity is not necessary for uniform conservativeness. The geometric interpretation of uniform conservativeness and an example of nonconvex but uniformly conservative CDF are illustrated in Figure 2.

When the CDF  $F(x)$  is differentiable, convexity of  $F(x)$  is equivalent to the density  $f(x)$  being monotonically increasing. This situation arises when we are performing one-sided tests in a one-dimensional exponential family,  $\{g_\theta, -\infty < \theta < \infty\}$ . Suppose the sufficient statistic is  $T$  (without loss of generality we assume the mean of  $T$  is increasing in  $\theta$ ) and the null hypothesis is  $H_0 : \theta \leq \theta_0$  vs.  $H_1 : \theta > \theta_0$ . The classical Karlin-Rubin theorem states that the uniformly most powerful (UMP) test rejects  $H_0$  when  $T$  is large, and a  $p$ -value can be computed using the right-tail of  $g_{\theta_0}$  (let  $G_\theta$  be the CDF of  $g_\theta$ ) by  $p = 1 - G_{\theta_0}(T)$ . Therefore, if  $T \sim g_\theta$ ,

$$F(x) = \mathbb{P}(p \leq x) = \mathbb{P}(T \geq G_{\theta_0}^{-1}(1-x)) = 1 - G_\theta(G_{\theta_0}^{-1}(1-x)).$$

By the inverse function theorem, this implies that

$$f(x) = \frac{g_\theta(G_{\theta_0}^{-1}(1-x))}{g_{\theta_0}(G_{\theta_0}^{-1}(1-x))}.$$

It is well known that the exponential family has monotone likelihood ratio (MLR), therefore  $f(x)$  is increasing in  $x$  if  $\theta < \theta_0$  (since  $G_{\theta_0}^{-1}(1-x)$  is decreasing in  $x$ ).

In summary, we have proved that

**Proposition 2.** *When the true  $\theta < \theta_0$ , the UMP one-sided test of  $H_0 : \theta \leq \theta_0$  vs.  $H_1 : \theta > \theta_0$  in the one-dimensional exponential family is uniformly conservative.*

Our Proposition 2 can be viewed as a special case of Whitt [1980, Theorem 1.1] who introduced a more general concept called uniform conditional stochastic order (UCSO). When the sample space is totally ordered, Whitt [1980] showed that MLR implies UCSO (uniform conservativeness). We refer the reader to Whitt [1980] for the more general result.

### 3.2 Testing qualitative interaction

The qualitative interaction problem formulated in Section 1.3 is a special case of one-sided testing in an exponential family. The  $i$ -th study ( $1 \leq i \leq n$ ) provides an effect estimate  $X_i \sim N(\mu_i, \sigma_i^2)$  with known  $\sigma_i$  and the null hypothesis of no qualitative interaction can be separated into two global testing problems: all the means are non-negative ( $H_0^+$ ) and all the means are non-positive ( $H_0^-$ ). Consider the first global null hypothesis  $H_0^+ = \cap_{1 \leq i \leq n} H_{0i}^+$ ,  $H_{0i}^+ : \mu_i \geq 0$ . Since the variance  $\sigma_i^2$  is known,  $H_{0i}^+$  is a one-sided problem in the normal location family. By Proposition 1 and Proposition 2, the conditional test of  $H_0^+$  is valid with any  $0 < \tau \leq 1$ .

Since  $H_0$  is the union of  $H_0^+$  and  $H_0^-$ , we can reject  $H_0$  at level  $\alpha$  if both  $H_0^+$  and  $H_0^-$  are rejected at level  $\alpha$ , because if  $H_0^+$  is true,

$$P(H_0 \text{ is rejected}) = P(H_0^+ \text{ and } H_0^- \text{ are rejected}) \leq P(H_0^+ \text{ is rejected}) \leq \alpha.$$

Similarly, the type I error is also controlled if  $H_0^-$  is true. This translates into the following testing procedure: for  $H_0^+$  and  $H_0^-$ , we can compute a combined  $p$ -value using the global tests in Section 2 (conditionally or unconditionally). Then we report a single  $p$ -value for  $H_0$  using the larger of the two combined  $p$ -values and reject  $H_0$  if it is less than the significance level  $\alpha$ .

In Section 5.3, we will compare the performance of our method with two existing tests of qualitative interaction that are widely used in practice. The first method is the likelihood ratio test (LRT) of Gail and Simon [1985]. The second method is the interval based graphical approach (IBGA) of Pan and Wolfe [1997], which is equivalent to the procedure described above using Šidák [1967]'s correction as the global test (applied to the unconditional  $p$ -values).

### 3.3 Adaptively selecting the threshold $\tau$

A remaining practical issue is how to choose the threshold  $\tau$ . Here we provide an adaptive strategy that attempts to select  $\tau$  without sacrificing the validity of the test. Our strategy is based on the following observation

**Proposition 3.** *If  $\tau$  is a backward stopping time in the sense that  $\{\tau \geq x\} \perp \{p_i, i \in \mathcal{S}_x\}$  for any  $0 \leq x \leq 1$ , then Proposition 1 still holds.*

This Proposition is true because our conditional test is based on  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$ , which is independent of how  $\tau$  is selected if  $\tau$  is a backward stopping time.

Proposition 3 suggests an interactive strategy to choose  $\tau$ :

1. The data analyst chooses a sequence of decreasing cutoffs,  $\tau_1, \tau_2, \tau_3, \dots, \tau_K$  (for example, 0.9, 0.85, 0.8,  $\dots$ , 0.1).
2. At step  $k \geq 1$ , the data analyst decides if she wants to continue based on  $\{p_i, i \notin \mathcal{S}_{\tau_k}\} = \{p_i, p_i > \tau_k\}$ . Denote the  $\tau_k$  she stops at as  $\tau$ .
3. Apply a global test on  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$ .

In principle, the starting cutoff  $\tau_1$  should not be too close to 1, otherwise there is little information for the data analyst to decide if she wants to move on. The ending cutoff  $\tau_K$  should not be too close to 0, so not too many signals are excluded.

Finally, we describe when the data analyst may want to stop. Consider the conditional Bonferroni test defined in (1). Since the minimum  $p$ -value does not depend on the threshold  $\tau$  (unless  $\tau$  is very small so  $\mathcal{S}_\tau$  is empty), it is reasonable to devise an adaptive strategy to minimize  $|\mathcal{S}_\tau|/\tau$ . Let  $\mathcal{F}$  be the distribution of the  $p$ -values:  $F(x) = (1/n) \sum_{i=1}^n F_i(x)$ , then  $|\mathcal{S}_\tau|/\tau \approx [nF(\tau)]/\tau$ . Notice that

$$\frac{d}{d\tau} \frac{F(\tau)}{\tau} = \frac{f(\tau)\tau - F(\tau)}{\tau^2}.$$

Therefore, a sensible criterion is to stop at step  $k$  if there is no strong evidence that  $f(\tau_k)\tau_k - F(\tau_k) > 0$ . More specifically, let  $0 < w \leq 1 - \tau_1$  be some prespecified window size (for example, 0.1). We can estimate  $F$  and  $f$  by

$$\hat{F}(\tau) = \frac{|\mathcal{S}_\tau|}{n}, \quad \hat{f}(\tau) = \frac{|\{i, \tau \leq p_i \leq \tau + w\}|}{nw},$$

and stop if we fail to reject  $nw\hat{f}(\tau_k) \sim \text{Binomial}(n, qw)$  with  $q < \hat{F}(\tau_k)/\tau_k$  at some pre-specified significance level (for example, 1%). We implement this heuristic strategy in the numerical studies in Sections 5 and 6 and find it generally improves the power of the conditional tests with fixed  $\tau$ . It also works well with other global tests too.

### 3.4 Beyond global testing

So far we have focused on testing the global null hypothesis that all the individual hypotheses are true. When the global null is rejected, it is often interesting to know which individual hypotheses are false. In this case, it is often desirable to control some multiple testing criterion such as the family-wise error rate (FWER) and the false discovery rate (FDR).

The conditional tests proposed above can be easily extended to general multiple testing problems. In fact, the conditional tests are closely related to the selective inference framework of Fithian et al. [2014] by viewing  $\mathcal{S}_\tau$  as model selection. Fithian et al. [2014] argued that the statistical inference should be performed conditioning on the selection event  $\mathcal{S}_\tau$ . See Benjamini [2010] for a discussion on the difference between simultaneous and selective inference. Notice that the conditional  $p$ -values  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$  can be viewed as usual  $p$ -values. We can apply, for example, Hochberg's step-up procedure to control the FWER, or the Benjamini-Hochberg procedure to control the FDR. In general, we expect the procedures using conditional  $p$ -values will be more powerful than their unconditional versions when many tests are conservative.

## 4 Some theoretical results

### 4.1 Power of the conditional test

**Theorem 1.** *Suppose all the  $p$ -values are independent and uniformly valid, and the cutoff  $0 < \tau < 1$  is a fixed constant. Let  $F_i$  be the CDF of the  $i$ -th  $p$ -value. Then the Bonferroni-adjusted conditional  $p$ -value  $p^{\text{CB}} = (\min_{1 \leq i \leq n} p_i/\tau) |\mathcal{S}_\tau|$  and the Bonferroni-adjusted unconditional  $p$ -value  $p^{\text{B}} = n \min_{1 \leq i \leq n} p_i$  satisfy*

$$\liminf_{n \rightarrow \infty} \frac{1}{\tau n} \sum_{i=1}^n F_i(\tau) + o_p(1) \leq \liminf_{n \rightarrow \infty} \frac{p^{\text{CB}}}{p^{\text{B}}} \leq \limsup_{n \rightarrow \infty} \frac{p^{\text{CB}}}{p^{\text{B}}} \leq \limsup_{n \rightarrow \infty} \frac{1}{\tau n} \sum_{i=1}^n F_i(\tau) + o_p(1). \quad (2)$$

Therefore, if the right hand side of (2) is less than 1, the conditional Bonferroni test is asymptotically more powerful than the conventional Bonferroni test.

Consider the case that the number of non-null  $p$ -values is a vanishing fraction of  $n$  (the situation in which the Bonferroni test is desirable; see, for example, Arias-Castro et al. [2011]). Recall that a  $p$ -value  $p_i$  is valid if  $F_i(\tau) \leq \tau$  for all  $\tau$  and is conservative if it is valid and the inequality  $F_i(\tau) \leq \tau$  is strict for some  $\tau$ . Therefore, the right hand side of (2) is always not greater than 1. Furthermore, when the non-nulls are sparse and the fraction of nulls that are conservative at  $\tau$ ,  $|\{1 \leq i \leq n : F_i(\tau) < \tau\}|/n$ , is non-negligible, the right hand side of (2) is less than 1 and thus the conditional Bonferroni test is more powerful than the unconditional test (consider the example in Section 2.2).

## 4.2 Validity of the conditional test under dependence

Next, we show that the conditional test is asymptotically valid when the test statistics exhibit exchangeable correlations. Suppose  $(Y_1, \dots, Y_n)$  follow the multivariate normal distribution with zero mean, unit variance and equal covariance  $\rho$ . The correlation  $\rho$  can arbitrarily vary as  $n$  increases, with the exception that  $\rho$  is bounded away from 1. For example,  $\rho$  can be any number no greater than 0.99. Note that, although  $\rho$  can be negative, it obeys  $\rho \geq -\frac{1}{n-1}$  in order to keep the covariance of  $Y_1, \dots, Y_n$  positive semidefinite. Finally, let  $p_i = 1 - \Phi(Y_i)$  be the one-sided  $p$ -value.

The next Theorem states that the conditional Bonferroni test still controls the type I error asymptotically when the test statistics  $Y_i$  are not independent but equally correlated.

**Theorem 2.** *In the above setting, we have  $P((\min_{1 \leq i \leq n} p_i/\tau) | \mathcal{S}_\tau| \leq \alpha) \leq (1 + o(1))\alpha$ .*

## 5 Simulation

### 5.1 Power of the global test

To assess the performance of the proposed procedures in Section 3, we implement a simulation study with  $n = 100$  one-sided tests of normal means  $H_{0j} : \mu_j \leq 0$  vs.  $H_{1j} : \mu_j > 0$  in the following five settings ( $\mu_{i:j}$  stands for the vector  $(\mu_i, \dots, \mu_j)$ ):

1. All null:  $\mu_{1:100} = 0$ ;
2. 1 strong 99 null:  $\mu_1 = 4, \mu_{2:100} = 0$ ;

3. 1 strong 99 conservative:  $\mu_1 = 4, \mu_{2:100} = -1$ ;
4. 20 weak 80 null:  $\mu_{1:20} = 1, \mu_{21:100} = 0$ ;
5. 20 weak 80 conservative:  $\mu_{1:20} = 1, \mu_{21:100} = -1$ .

The test statistics are generated by  $Y_i \sim N(\mu_i, 1)$  and the  $p$ -values are computed by  $p_i = 1 - \Phi(Y_i)$  where  $\Phi$  is the CDF of standard normal. Then we apply four combination tests, Bonferroni, Fisher, Tukey and truncated product, in two forms—the original unconditional test in Section 2.1 and the conditional ( $\tau = 0.5$ ). The null distribution of Tukey’s test is approximated by 10000 samples from the global null.

Table 2 reports the power of these tests in 10000 simulations when the significance level is  $\alpha = 0.05$ . When all the null hypotheses are true, all the tests have power (a.k.a. type I error) about 5%. When there is only one strong signal, Bonferroni performs the best, but notice that conditioning does not make the power deteriorate substantially when no test is conservative (in fact it improves the power of Fisher, Tukey and truncated product) and substantially improves the power when many tests are conservative. The same thing is true when we have many weak signals, except that Fisher’s combination and the truncated product perform the best in this regime.

## 5.2 Power of signal detection

Next we investigate the empirical performance of four Bonferroni procedures that control the family-wise error rate (FWER): the original Bonferroni correction and the conditional Bonferroni test with  $\tau = 0.5$  and  $\tau = 0.8$ . We simulate  $n = 1000$  one-sided tests of normal means in the following two settings:

1. No conservative:  $\mu_{1:20} = 4, \mu_{21:1000} = 0$ ;
2. Conservative:  $\mu_{1:20} = 4, \mu_{21:1000} = -1$ .

The test statistics are generated by  $Y_i \sim N(\mu_i, 1)$  and the  $p$ -values are computed by  $p_i = 1 - \Phi(Y_i)$ .

Table 3 compares the power of four different Bonferroni procedures. The numbers in Table 3 are summary statistics of the number of correct rejections in 1000 simulations. For example, when no test is conservative, the original Bonferroni procedure rejects 10.91 tests on average, while the conditional Bonferroni procedures reject 10.87, 10.90 and 10.90

Setting	Method	Uncond.	Cond.	Adaptive
1. All null	Bonferroni	4.9	4.9	4.9
	Fisher	5.1	4.8	5.1
	Tukey	5.3	5.0	5.4
	TruncatedP	5.0	4.9	5.0
2. 1 strong 99 null	Bonferroni	<b>78.0</b>	<b>78.0</b>	<b>78.0</b>
	Fisher	25.9	34.7	27.2
	Tukey	7.0	6.9	7.5
	TruncatedP	23.4	31.2	24.5
3. 1 strong 99 conservative	Bonferroni	<b>76.2</b>	<b>85.1</b>	<b>88.7</b>
	Fisher	0.0	20.3	84.7
	Tukey	0.0	0.1	4.2
	TruncatedP	0.0	21.0	84.4
4. 20 weak 80 null	Bonferroni	22.8	20.5	22.3
	Fisher	<b>73.9</b>	<b>57.2</b>	<b>71.4</b>
	Tukey	57.8	40.4	54.4
	TruncatedP	70.2	53.9	67.2
5. 20 weak 80 conservative	Bonferroni	<b>20.0</b>	28.2	28.1
	Fisher	0.0	48.7	<b>52.3</b>
	Tukey	0.5	25.9	30.8
	TruncatedP	0.3	<b>51.0</b>	<b>52.9</b>

Table 2: Power (in %) in the 5 simulation settings in which the global tests (Bonferroni, Fisher, Tukey, and TruncatedP) are applied to the unconditional and conditional  $p$ -values. The truncation threshold  $\tau$  is 0.5 or chosen adaptively as described in Section 3.3.

Setting	Method	1st Quart.	Med.	Mean	3rd Quart.
No conservative	Bonferroni	9	11	10.91	12
	Cond. Bonf. ( $\tau = 0.5$ )	9	11	10.87	12
	Cond. Bonf. ( $\tau = 0.8$ )	9	11	10.90	12
	Cond. Bonf. (adaptive $\tau$ )	10	11	10.90	12
Conservative	Bonferroni	9	11	10.78	12
	Cond. Bonf. ( $\tau = 0.5$ )	11	13	12.78	14
	Cond. Bonf. ( $\tau = 0.8$ )	10	12	11.88	13
	Cond. Bonf. (adaptive $\tau$ )	12	13	13.12	15

Table 3: Quartiles and means of the number of correct rejections in 1000 simulations.

tests on average depending on which threshold  $\tau$  is used. In contrast, in the conservative scenario the conditional Bonferroni with  $\tau = 0.5$  rejects about 2 more tests than the original Bonferroni procedure on average. The conditional Bonferroni's method with adaptively selected  $\tau$  makes even more discoveries. In conclusion, the proposed conditional tests are slightly less powerful when no test is conservative, but are much more powerful when many tests are conservative.

### 5.3 Power of testing qualitative interaction

Finally, we study the power of the tests of qualitative interaction that are described in Section 5.3. We simulate  $n = 100$  normal variables  $Y_i \sim N(\mu_i, 1)$  in the following six settings:

1. 1 positive 99 null:  $\mu_1 = 4, \mu_{2:100} = 0$ .
2. 1 positive 1 negative:  $\mu_1 = 4, \mu_2 = -4, \mu_{3:100} = 0$ ;
3. 1 positive 99 negative:  $\mu_1 = 4, \mu_{2:100} = -1$ ;
4. 20 positive 80 negative:  $\mu_{1:20} = 1, \mu_{21:100} = -1$ ;
5. 50 positive 50 negative:  $\mu_{1:50} = 1, \mu_{51:100} = -1$ ;
6. Gradual (1st setting):  $\mu_{1:100}$  are equally spaced between  $-1.5$  and  $2$ ;
7. Gradual (2nd setting):  $\mu_{1:100}$  are equally spaced between  $-1.5$  and  $4$ .

Apart from the first setting, the null hypothesis of no qualitative interaction is false. Table 4 compares the power of the proposed tests with two existing methods, the likelihood ratio test (LRT) of Gail and Simon [1985] and the interval based graphical approach (IBGA) of Pan and Wolfe [1997] as implemented in the R package `QualInt` [Yu et al., 2014]. The performance of IBGA is very similar to the unconditional Bonferroni test, since IBGA is equivalent to applying Šidák's correction to the original  $p$ -values in our framework and it is well known that Šidák's correction is only slightly more powerful than Bonferroni's correction.

Across all settings and all global tests, conditioning (whether using  $\tau = 0.5$  or  $\tau$  adaptively chosen) improves the power of detecting qualitative interaction. Apart from the second setting with only 1 positive and 1 negative signal where the Bonferroni/Šidák's

Setting	Method	Uncond.	Cond.	Adaptive
1 positive 99 null	Bonferroni	3.6	3.6	3.6
	Fisher	0.1	1.7	1.0
	Tukey	0.0	0.3	0.3
	TruncatedP	0.5	1.6	1.4
	IBGA	3.7	3.7	3.7
	LRT	1.2	1.2	1.2
1 positive 1 negative	Bonferroni	59.9	<b>59.9</b>	<b>60.0</b>
	Fisher	1.0	11.6	6.1
	Tukey	0.0	0.4	0.3
	TruncatedP	2.9	9.7	6.5
	IBGA	<b>60.4</b>		
	LRT	12.8		
1 positive 99 negative	Bonferroni	50.9	<b>45.4</b>	57.6
	Fisher	0.0	19.6	<b>84.9</b>
	Tukey	0.0	0.2	4.2
	TruncatedP	0.0	20.7	<b>84.5</b>
	IBGA	<b>51.9</b>		
	LRT	0.0		
20 positive 80 negative	Bonferroni	<b>11.7</b>	14.4	16.7
	Fisher	0.0	49.6	<b>51.8</b>
	Tukey	0.6	27.4	30.0
	TruncatedP	0.3	<b>51.4</b>	<b>52.5</b>
	IBGA	<b>12.1</b>		
	LRT	3.0		
50 positive 50 negative	Bonferroni	18.6	18.7	21.7
	Fisher	71.5	<b>97.1</b>	<b>98.3</b>
	Tukey	73.7	90.7	94.0
	TruncatedP	92.5	94.9	<b>98.1</b>
	IBGA	19.5		
	LRT	<b>93.8</b>		
Gradual (from $-1.5$ to $2$ )	Bonferroni	26.5	28.0	30.2
	Fisher	18.3	<b>86.8</b>	<b>87.9</b>
	Tukey	29.6	71.2	72.4
	TruncatedP	53.7	85.0	<b>88.2</b>
	IBGA	27.4		
	LRT	<b>67.5</b>		
Gradual (from $-1.5$ to $4$ )	Bonferroni	<b>24.8</b>	35.4	36.4
	Fisher	0.0	<b>72.9</b>	<b>73.7</b>
	Tukey	1.0	51.9	48.9
	TruncatedP	1.1	70.6	<b>73.6</b>
	IBGA	<b>25.3</b>		
	LRT	7.7		

Table 4: Power (in %) of testing qualitative interaction in the 7 simulation settings. The proposed methods—global tests (Bonferroni, Fisher, Tukey, and TruncatedP) applied to the unconditional and conditional  $p$ -values—are compared with two existing methods, the interval based graphical approach (IBGA) [Pan and Wolfe, 1997] and the likelihood ratio test (LRT) [Gail and Simon, 1985]. For<sup>17</sup>the conditional global tests, the truncation threshold  $\tau$  is 0.5 or chosen adaptively as described in Section 3.3.

tests have the most power, in all the other settings the conditional Fisher’s test (and its variant, truncated product) with adaptively chosen  $\tau$  is always the most powerful method. In practice, if there is qualitative interaction, it is rare that there is only one subgroup with strong signal of the opposite sign (in other words, the last four settings are more plausible than the second and third settings). Therefore, we expect the conditional Fisher’s test with adaptively chosen  $\tau$  to perform the best in practice among the tests considered in this paper.

## 6 Applications to educational interventions

### 6.1 Random effects model only tests heterogeneity of treatment effect

In Section 1 we introduced two motivating applications in evaluating educational interventions. The standard practice to analyze such datasets is the linear fixed/random/mixed effects model. For example, for the modified school calendar intervention, a typical random effects model is

$$Y_i = \mu_i + \epsilon_i = (\mu + \alpha_{D_i} + \beta_{S_i}) + \epsilon_i, \quad i = 1, \dots, n, \quad (3)$$

where  $Y_i$  is the observed average treatment effect in study  $S_i$  nested in district  $D_i$ ,  $\mu$  is the overall treatment effect,  $\alpha_{D_i} \sim N(0, \sigma_D^2)$  is the random district effect,  $\beta_{S_i} \sim N(0, \sigma_S^2)$  is the random study effect (nested in the district), and  $\epsilon_i \sim N(0, \sigma_i^2)$  is the noise with known sampling variance  $\sigma_i^2$  because the per-study effect  $Y_i$  is aggregated over many individuals.

To test heterogeneity of treatment effect, we use the function `rma.mv` in the R package `metafor` [Viechtbauer, 2010] to fit a multi-level random effects model using restricted maximum likelihood. For  $\sigma_D^2$ , the point estimate is 0.0651 and the 95% confidence interval is (0.0222, 0.2072). For  $\sigma_S^2$ , the point estimate is 0.0327 and the 95% confidence interval is (0.0163, 0.0628). Therefore, there is strong evidence that the treatment effect of modified school calendar varies across districts and schools. This is consistent with the conclusions of Konstantopoulos [2011].

In general, the random effects model is not suitable for testing qualitative interaction. If we take model (3) and its random effects assumptions literally, there is always a positive chance that some  $\mu_i$  is negative if  $\sigma_D^2 > 0$  or  $\sigma_S^2 > 0$ . In other words, the hypothesis of no qualitative interaction is automatically false in a random effects model.

Alternatively, we may treat the district effects  $\alpha_{D_i}$  as fixed and ask if any district experiences a negative treatment effect. Outputs for this mixed effect model is reported in

Table 5: Output table of a mixed effect model for the modified school calendar application. In this model, each district has a fixed effect but the schools have random effects.

	Estimate	Std. Err.	$z$ -value	$p$ -value
District 11	-0.129	0.181	-0.71	0.476
District 12	0.063	0.064	0.98	0.325
District 18	0.347	0.083	4.18	0.000
District 27	0.486	0.040	12.01	0.000
District 56	0.041	0.042	0.98	0.329
District 58	-0.042	0.033	-1.30	0.194
District 71	0.879	0.064	13.75	0.000
District 86	-0.029	0.015	-1.86	0.063
District 91	0.250	0.044	5.68	0.000
District 108	0.015	0.079	0.19	0.853
District 644	0.157	0.137	1.14	0.253

Table 5, where none of the districts shows a significantly negative effect.

## 6.2 Applying the proposed tests for qualitative interaction

We apply the tests for qualitative interaction described in Section 5.3 to the two datasets. The results are reported in Table 6.

For the modified school calendar intervention, none of the individual schools has a strong enough effect after Bonferroni’s correction to reject the hypothesis at significance level 0.01. Conditioning ( $\tau = 0.5$  and  $\tau = 0.8$ ) helps to make the Bonferroni adjusted  $p$ -values smaller, but they are still greater than 0.01. In contrast, by combining the weak evidence from several schools and by reducing the number of conservative  $p$ -values via conditioning, the conditional Fisher’s test with  $\tau = 0.5$  gives a  $p$ -value of 0.0002. The  $p$ -value is still significant when the truncation threshold is set to  $\tau = 0.8$ . Without conditioning, Fisher’s combination test does not have enough power to detect the qualitative interaction in this application.

We can also test if the treatment effect has qualitative interaction among the districts. Using the  $z$ -values in Table 5, we apply the same global tests and obtained 6  $p$ -values in the second row of Table 6. None of them is significant at level 0.05, indicating insufficient evidence of qualitative interaction in the district level.

For the writing-to-learn intervention, all the tests cannot reject the null hypothesis of no qualitative interaction. In the forest plot (Figure 1b), Ayers (1993) study shows

Table 6: Combined  $p$ -values for qualitative interaction in the motivating applications in Section 1.1. Three versions of the Bonferroni’s test and Fisher’s combination test are used: the unconditional test (Unc.), the conditional test with threshold 0.5 and 0.8, and the conditional test with adaptively selected threshold  $\tau$ .

		Unc.	$\tau = 0.5$	$\tau = 0.8$	$\tau$ adaptive
Modified calendar (school)	Bonferroni	0.044	0.031	0.034	0.033
	Fisher	0.224	< 0.001	0.004	0.003
	IBGA	0.042			
	LRT	0.011			
Modified calendar (district)	Bonferroni	0.347	0.189	0.158	0.245
	Fisher	0.788	0.113	0.088	0.374
	IBGA	0.274			
	LRT	0.351			
Writing-to-learn	Bonferroni	0.830	0.381	0.519	0.503
	Fisher	1	0.578	0.917	0.877
	IBGA	0.556			
	LRT	0.985			

a significantly negative effect, but our results suggest that it is plausible that is due to random chance.

## 7 Discussion

### 7.1 Multiparameter hypothesis testing

The global testing problem considered in this paper is also closely related to the multiparameter hypothesis testing problem considered by Lehmann [1952], Berger [1982] that tests  $H_0 : \theta \leq 0$  for a multidimensional parameter  $\theta$ . Lehmann [1952] showed that in general there is no unbiased test for this problem, i.e. apart from the trivial test that has constant power function, any valid test must have power less than  $\alpha$  at some alternative. In our paper, we assume there are independent tests for each individual hypothesis  $\theta_i \leq 0$  and we restrict our attention to the alternative that many  $\theta_i$ s are much smaller than 0, so our results do not contradict the conclusions in Lehmann [1952]. Another distinction is that we allow the dimension of  $\theta$  to go to infinity, while in the classical multiparameter setting the dimension of  $\theta$  is fixed.

## 7.2 Uniform validity/conservativeness

As mentioned in Section 1.4, not all conservative tests are uniformly conservative. One important exception we are aware of is the sensitivity analysis of observational studies [Rosenbaum, 2002, Chapter 4] which places bounds on the  $p$ -value for a specific magnitude of departure from random treatment assignment. When there is no treatment effect, the  $p$ -value under random treatment assignment is uniformly distributed and not conservative, but the  $p$ -value bounds under departure from randomization are inevitably very conservative because many possible departures are considered. Unfortunately, the  $p$ -value bounds are generally not uniformly valid [see e.g., Zhao, 2017]. When uniform conservativeness does not hold, other methods (e.g. sample splitting in Heller et al. [2009]) must be used to reduce the number of hypotheses. However, sample splitting loses some efficiency because it discards some information in the data whereas the conditional test proposed in this paper makes full use of the information.

Another notable exception of uniform validity is when the  $p$ -values are discrete. In this case, the  $p$ -values cannot be strictly uniformly valid. However, for tests with asymptotically normal approximations (such as the bootstrap or Wilcoxon’s rank sum test), we expect the conditional tests in this paper are still asymptotically valid.

## A Proofs

### A.1 Proposition 1

*Proof.* Conditioning on the set  $\mathcal{S}_\tau$ , for any  $i, j \in \mathcal{S}_\tau$ ,  $p_i/\tau$  is a valid  $p$ -value and  $p_i/\tau$  is independent of  $p_j/\tau$ . Therefore, conditioning on the set  $\mathcal{S}_\tau$ , the global test on  $\{p_i/\tau, i \in \mathcal{S}_\tau\}$  controls type I error at the nominal level (on  $\mathcal{S}_\tau$ ). By marginalizing over  $\mathcal{S}_\tau$ , the statement holds unconditionally as well.  $\square$

### A.2 Folded normal distribution

**Proposition 4.** *The family of folded normal distributions with standard deviation  $\sigma = 1$  and varying  $\mu$  has monotone likelihood ratio. More precisely, if  $\mu_1 > \mu_2 \geq 0$ , then*

$$\frac{\partial}{\partial x} \frac{\phi(x - \mu_1) + \phi(x + \mu_1)}{\phi(x - \mu_2) + \phi(x + \mu_2)} > 0, \quad \forall x > 0. \quad (4)$$

*Proof.* We will repeatedly use the fact  $(d/dx)\phi(x) = -x\phi(x)$  in the proof. By evaluating

the differentiation in (4), it suffices to prove

$$g(\mu) = \frac{-(x-\mu)\phi(x-\mu) - (x+\mu)\phi(x+\mu)}{\phi(x-\mu) + \phi(x+\mu)}$$

is an increasing function of  $\mu \geq 0$ . Taking the derivative of  $g(\mu)$ , we have

$$\begin{aligned} & [\phi(x-\mu) + \phi(x+\mu)]^2 \cdot \frac{d}{d\mu} g(\mu) \\ &= \{[-(x-\mu)^2 + 1]\phi(x-\mu) + [(x+\mu)^2 - 1]\phi(x+\mu)\} [\phi(x-\mu) + \phi(x+\mu)] \\ &\quad - [-(x-\mu)\phi(x-\mu) - (x+\mu)\phi(x+\mu)] [(x-\mu)\phi(x-\mu) - (x+\mu)\phi(x+\mu)] \\ &= \phi(x-\mu)^2 - \phi(x+\mu)^2 + 4\mu x \cdot \phi(x-\mu)\phi(x+\mu) > 0. \end{aligned}$$

□

### A.3 Theorem 1

*Proof.* Denote

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[|S_\tau|]}{\tau n} = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n F_i(\tau)/\tau = c.$$

By applying the Chebyshev inequality and making use of the above display, we get, for any  $\epsilon > 0$

$$\begin{aligned} \limsup \mathbb{P} \left( \frac{|S_\tau|}{\tau} > (c + \epsilon)n \right) &= \limsup \mathbb{P} \left( \frac{|S_\tau|}{\tau} - \frac{\mathbb{E}[|S_\tau|]}{\tau} > (c + \epsilon)n - \frac{\mathbb{E}[|S_\tau|]}{\tau} \right) \\ &\leq \limsup \frac{\mathbb{E} \left[ \frac{|S_\tau|}{\tau} - \frac{\mathbb{E}[|S_\tau|]}{\tau} \right]^2}{\left[ (c + \epsilon)n - \frac{\mathbb{E}[|S_\tau|]}{\tau} \right]^2} \\ &\leq \limsup \frac{n/4}{[(\epsilon + o(1))n]^2} \\ &= 0. \end{aligned}$$

This implies that

$$p^{\text{CB}} = \frac{|S_\tau|}{\tau} \cdot \min_{1 \leq i \leq n} p_i \leq (c + o_p(1)) n \min_{1 \leq i \leq n} p_i = (c + o_p(1)) p^{\text{B}}.$$

The other side of the inequality can be proven similarly.

□

#### A.4 Theorem 2

**Lemma 1.** *Let  $\epsilon > 0$  be an arbitrary constant. Then Theorem 2 holds for any correlation sequence  $\{\rho_l\}_{l=1}^{\infty}$  such that  $\rho_l \geq \epsilon$  for all  $l$ .*

**Lemma 2.** *Theorem 2 holds for any correlation sequence  $\{\rho_l\}_{l=1}^{\infty}$  such that  $\rho_l \rightarrow 0$ .*

Taking these two lemmas as given for the moment, a proof of Theorem 2 is readily given below.

*Proof of Theorem 2.* Let  $\hat{n}_\tau = |\mathcal{S}_\tau|/\tau$ . Suppose on the contrary that Theorem is false. Then, we can pick a subsequence  $\{\rho_{s_1}, \rho_{s_2}, \dots\}$  such that, restricted to this subsequence,

$$P(\hat{n}_\tau \cdot p_{\min} \leq \alpha) > (1 + c)\alpha \quad (5)$$

for some constant  $c > 0$ .

Note that the sequence  $\{\rho_{s_1}, \rho_{s_2}, \dots\}$  must further contain a subsequence with each element bounded below by 0 or a subsequence with vanishing elements. In the former case, Lemma 1 contradicts with (5), and in the latter case, a contradiction arises between Lemma 2 and (5). Hence, such subsequence  $\rho_{s_1}, \rho_{s_2}, \dots$  should not exist at all, leading to the correctness of this theorem.

□

*Proof of Lemma 1.* Recognizing that the equi-correlations  $\rho$  are positive, we start with the following representation

$$Y_i \stackrel{d}{=} \sqrt{1 - \rho}X_i + \sqrt{\rho}W,$$

where  $X_1, \dots, X_n, W$  are iid  $\mathcal{N}(0, 1)$ . Write  $X_{\max} = \max\{X_1, \dots, X_n\}$ . Then

$$p_{\min} = \Phi(-\sqrt{1 - \rho}X_{\max} - \sqrt{\rho}W). \quad (6)$$

Making use the fact that  $\Phi(-x) = (1 + o(1))\varphi(x)/x$  for  $x \rightarrow \infty$ , from (6) we get

$$p_{\min} = (1 + o_p(1)) \frac{1}{\sqrt{1 - \rho}X_{\max} + \sqrt{\rho}W} \varphi(\sqrt{1 - \rho}X_{\max} + \sqrt{\rho}W), \quad (7)$$

where the term  $o_p(1)$  results from recognizing  $\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W \rightarrow \infty$  as  $n \rightarrow \infty$  in probability. We proceed to bound  $\varphi(\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W)$ . Note that

$$\begin{aligned}\varphi(\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W) &= \frac{1}{\sqrt{2\pi}} \exp \left[ -(\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W)^2/2 \right] \\ &= \frac{1}{\sqrt{2\pi}} e^{-I_1 - I_2 - I_3},\end{aligned}$$

where  $I_1 = (1-\rho)X_{\max}^2/2$ ,  $I_2 = \rho W^2/2$ ,  $I_3 = \sqrt{\rho(1-\rho)}X_{\max}W$ . Using  $X_{\max} = (1 + o_p(1))\sqrt{2\log n}$ , we see the first term  $I_1$  obeys

$$I_1 = (1-\rho)X_{\max}^2/2 = (1-\rho)(1 + o_p(1)) \left( \sqrt{2\log n} \right)^2 / 2 \leq (1 - \epsilon + o_p(1)) \log n.$$

The second term satisfies  $I_2 = \rho W^2/2 = O_p(1) = o_p(I_1)$ , and the last terms obeys

$$I_3 = \sqrt{\rho(1-\rho)}X_{\max}W = O_p(\sqrt{2\log n}) = o_p(I_1).$$

Taking these results together yields  $I_1 + I_2 + I_3 = (1 + o_p(1))I_1 \leq (1 - \epsilon + o_p(1)) \log n$ . Hence, we obtain

$$\varphi(\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W) \geq \frac{1}{\sqrt{2\pi}} e^{-(1-\epsilon+o_p(1))\log n} = \frac{1}{\sqrt{2\pi}n^{1-\epsilon+o_p(1)}}.$$

Plugging the inequality above into the right-hand side of (7) gives

$$\begin{aligned}p_{\min} &= (1 + o_p(1)) \frac{\varphi(\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W)}{\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W} \\ &\geq (1 + o_p(1)) \frac{1}{\sqrt{2\pi}n^{1-\epsilon+o_p(1)} [\sqrt{1-\rho}X_{\max} + \sqrt{\rho}W]} \\ &= (1 + o_p(1)) \frac{1}{2n^{1-\epsilon+o_p(1)} \sqrt{\pi(1-\rho)} \log n}.\end{aligned}\tag{8}$$

Next, we move on to consider  $\hat{n}_c = n_c/c$ . Each  $p$ -value  $p_i = \Phi(-\sqrt{1-\rho}X_i - \sqrt{\rho}W)$  is below the cutoff  $c$  if and only if

$$X_i \geq -\frac{\Phi^{-1}(c) + \sqrt{\rho}W}{\sqrt{1-\rho}},$$

which asserts

$$\frac{n_c}{n} = (1 + o_p(1))\Phi\left(\frac{\Phi^{-1}(c) + \sqrt{\rho}W}{\sqrt{1-\rho}}\right). \quad (9)$$

Combing (8) and (9) yields

$$\begin{aligned} \hat{n}_c \cdot p_{\min} &= (1 + o_p(1)) \frac{n\Phi\left(\frac{\Phi^{-1}(c) + \sqrt{\rho}W}{\sqrt{1-\rho}}\right)}{c} p_{\min} \\ &\geq (1 + o_p(1)) \frac{n\Phi\left(\frac{\Phi^{-1}(c) + \sqrt{\rho}W}{\sqrt{1-\rho}}\right)}{c} \cdot \frac{1}{2n^{1-\epsilon+o_p(1)}\sqrt{\pi(1-\rho)}\log n} \\ &= (1 + o_p(1)) \frac{\Phi\left(\frac{\Phi^{-1}(c) + \sqrt{\rho}W}{\sqrt{1-\rho}}\right)}{2c\sqrt{\pi(1-\rho)}} \cdot \frac{n^{\epsilon+o_p(1)}}{\sqrt{\log n}}. \end{aligned}$$

Observe that the first term

$$\frac{\Phi\left(\frac{\Phi^{-1}(c) + \sqrt{\rho}W}{\sqrt{1-\rho}}\right)}{2c\sqrt{\pi(1-\rho)}}$$

is a positive random variable bounded away from 0 with high probability (though it depends on  $n$ ), whereas the second term  $n^{\epsilon+o_p(1)}/\sqrt{\log n}$  diverges to  $\infty$  as  $n \rightarrow \infty$ . This immediately implies

$$\mathbb{P}(\hat{n}_c \cdot p_{\min} \leq \alpha) \rightarrow 0,$$

which is stronger than what the lemma claims.  $\square$

*Proof of Lemma 2.* We start by proving the fact that  $\hat{n}_c = (1 + o_p(1))n$ . First, we note that

$$\mathbb{E}[\hat{n}_c] = n. \quad (10)$$

Next, its variance is given as

$$\begin{aligned} \text{Var}(\hat{n}_c) &= \frac{\text{Var}(\sum_{i=1}^n \mathbf{1}(p_i \leq c))}{c^2} \\ &= \frac{n \text{Var}(\mathbf{1}(p_1 \leq c)) + n(n-1) \text{Cov}(\mathbf{1}(p_1 \leq c), \mathbf{1}(p_2 \leq c))}{c^2} \\ &\leq \frac{n/4 + n(n-1) \text{Cov}(\mathbf{1}(p_1 \leq c), \mathbf{1}(p_2 \leq c))}{c^2}. \end{aligned}$$

To proceed, use the fact that  $\text{Cov}(\mathbf{1}(p_1 \leq c), \mathbf{1}(p_2 \leq c)) = O(\rho)$ . Then we get

$$\sqrt{\text{Var}(\hat{n}_c)} = \sqrt{n^2 O\left(\frac{1}{n} + \rho\right)} = o(n), \quad (11)$$

which together with (10) gives

$$\hat{n}_c = (1 + o_p(1))n.$$

Hence, we get

$$\begin{aligned} \mathbb{P}(\hat{n}_c \cdot p_{\min} \leq \alpha) &\leq \sum_{i=1}^n \mathbb{P}(\hat{n}_c \cdot p_i \leq \alpha) \\ &= \sum_{i=1}^n \mathbb{P}((1 + o_p(1))n \cdot p_i \leq \alpha) \\ &= \sum_{i=1}^n (1 + o(1)) \frac{\alpha}{n} \\ &= (1 + o(1))\alpha, \end{aligned}$$

as desired. □

## References

- Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Annals of Statistics*, 39(1):2533–2556, 2011.
- Robert L Bangert-Drowns, Marlene M Hurley, and Barbara Wilkinson. The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, 74(1):29–58, 2004.
- Yoav Benjamini. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

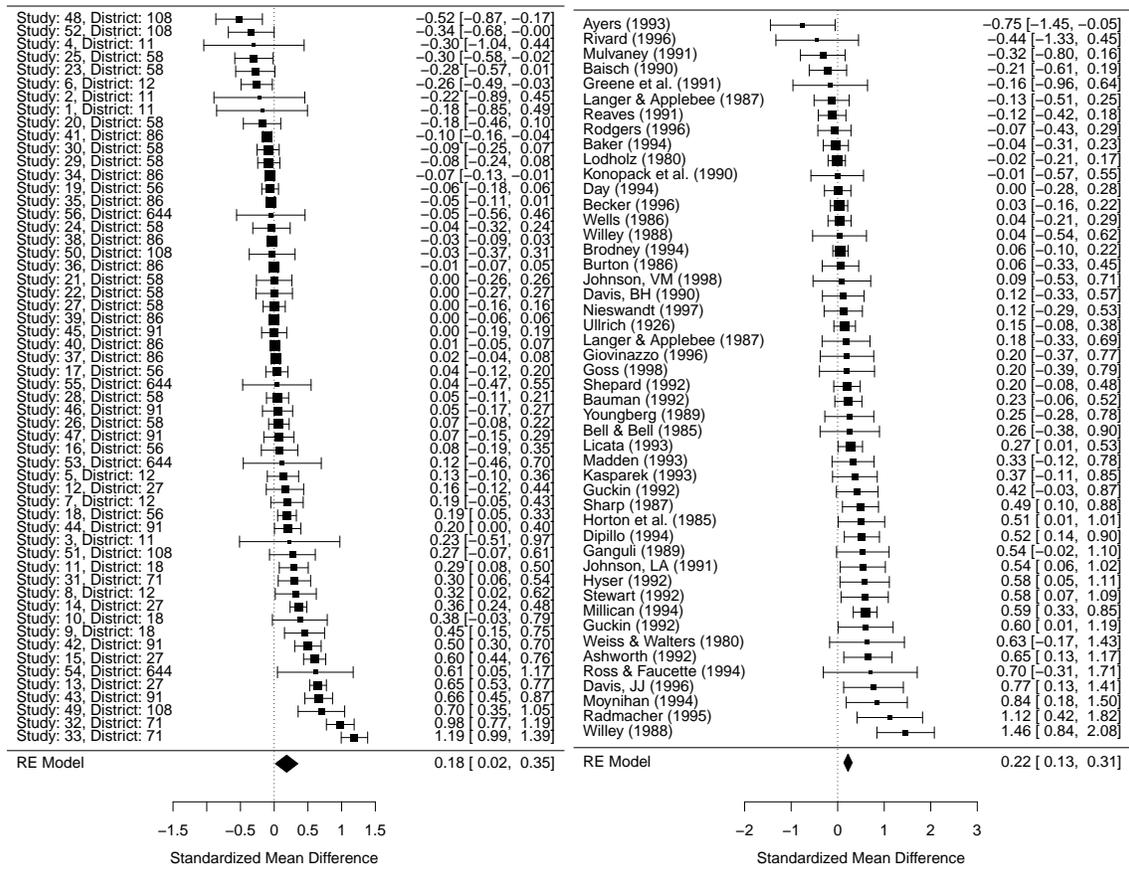
- Roger L Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982.
- Howard S Bloom, Stephen W Raudenbush, Michael J Weiss, and Kristin Porter. Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, to appear, 2017.
- Harris Cooper, Jeffrey C Valentine, Kelly Charlton, and April Melson. The effects of modified school calendars on student achievement and on school and community attitudes. *Review of Educational Research*, 73(1):1–52, 2003.
- CRASH-2-Collaborators. The importance of early treatment with tranexamic acid in bleeding trauma patients: an exploratory analysis of the CRASH-2 randomised controlled trial. *The Lancet*, 377(9771):1096–1101, 2011.
- Lee J Cronbach and Richard E Snow. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. Irvington, 1977.
- David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004.
- Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.
- William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.
- M Gail and R Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372, 1985.
- Ruth Heller, Paul R Rosenbaum, and Dylan S Small. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101, 2009.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- Spyros Konstantopoulos. Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, 2(1):61–76, 2011.

- Erich L Lehmann. Testing multiparameter hypotheses. *Annals of Mathematical Statistics*, 23(4):541–552, 1952.
- Guohua Pan and Douglas A Wolfe. Test for qualitative interaction of clinical significance. *Statistics in Medicine*, 16(14):1645–1652, 1997.
- Giorgio Pizzocaro, Luigi Piva, Maria Colavita, Sonia Ferri, Raffaella Artusi, Patrizia Boracchi, Giorgio Parmiani, and Ettore Marubini. Interferon adjuvant to radical nephrectomy in robson stages ii and iii renal cell carcinoma: a multicentric randomized study. *Journal of Clinical Oncology*, 19(2):425–431, 2001.
- Paul R Rosenbaum. *Observational Studies*. Springer, 2002.
- Peter Z Schochet, Mike Puma, and John Deke. Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods (NCEE 2014-4017). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2014.
- Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.
- R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- Wenguang Sun and Alexander C McLain. Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, 107(498):673–687, 2012.
- Wolfgang Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.
- Rui Wang, Stephen W Lagakos, James H Ware, David J Hunter, and Jeffrey M Drazen. Statistics in medicinereporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194, 2007.
- Ward Whitt. Uniform conditional stochastic order. *Journal of Applied Probability*, 17:112–123, 1980.

Lixi Yu, Eun-Young Suh, and Guohua Pan. *QualInt: Test for Qualitative Interactions*, 2014. R package version 1.0.0.

Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.

Qingyuan Zhao. On sensitivity value of pair-matched observational study. *arXiv preprint arXiv:1702.03442*, 2017.



(a) Example 1: effect of modified school calendar. (b) Example 2: effect of writing-to-learn intervention.

Figure 1: Forest plots of two meta-analyses with potential qualitative interaction.

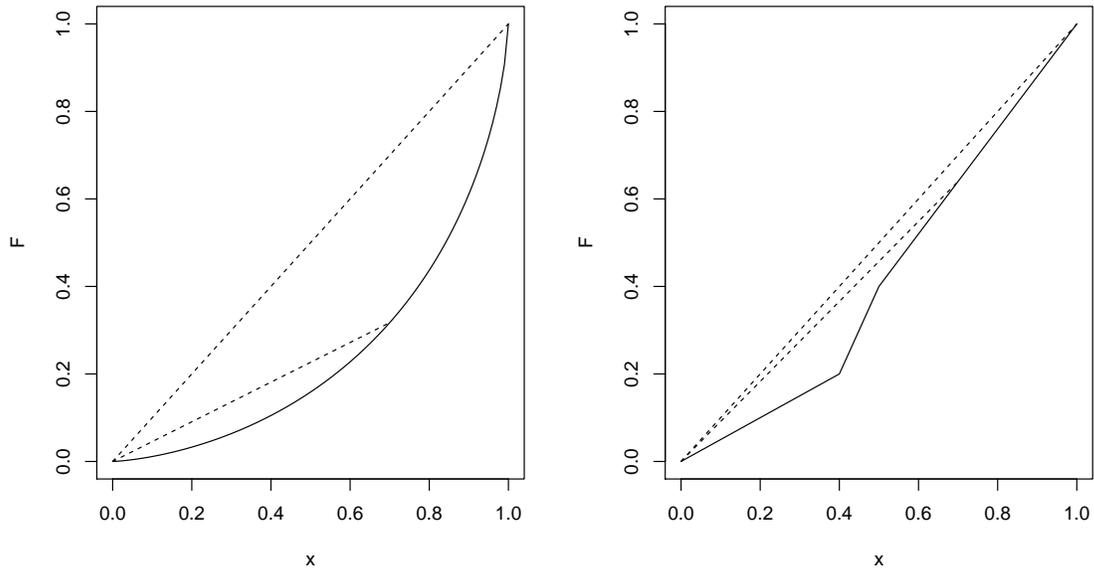


Figure 2: Two examples of uniformly conservative CDFs. The left plot is the distribution of  $\Phi(Y)$  where  $Y \sim N(-1, 1)$ . The right plot corresponds to a piecewise constant density function:  $f(x) = 0.5 \cdot I(0 \leq x \leq 0.4) + 2 \cdot I(0.4 < x \leq 0.5) + 1.2 \cdot I(0.5 < x \leq 1)$ . Both CDFs satisfy the condition  $F(x\tau) \leq xF(\tau)$  for all  $0 \leq x, \tau \leq 1$  so they are uniformly conservative. The geometric interpretation of this condition is illustrated by the two dashed lines corresponding to  $\tau = 0.7$  and 1. The right plot suggests that convexity of CDF is not necessary for uniform conservativeness.