

1 Introduction

1.1 Overview of multiple testing

Hypothesis testing is concerned with using observed data to test hypotheses, i.e., make decisions, regarding properties of the unknown data generating distribution. Below, we discuss in turn the main ingredients of a multiple testing problem, namely: data, null and alternative hypotheses, test statistics, a multiple testing procedure (MTP) to define rejection regions for the test statistics, and Type I and Type II errors.

Data. Let X_1, \dots, X_n be a *random sample* of n independent and identically distributed (i.i.d.) random variables, $X \sim P \in \mathcal{M}$, where the *data generating distribution* P is known to be an element of a particular *statistical model* \mathcal{M} (i.e., a set of possibly nonparametric distributions).

Null and alternative hypotheses. In order to cover a broad class of testing problems, define M null hypotheses in terms of a collection of *submodels*, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \dots, M$, for the data generating distribution P . The M *null hypotheses* are defined as $H_0(m) \equiv \mathbb{I}(P \in \mathcal{M}(m))$ and the corresponding *alternative hypotheses* as $H_1(m) \equiv \mathbb{I}(P \notin \mathcal{M}(m))$.

In many testing problems, the submodels concern *parameters*, i.e., functions of the data generating distribution P , $\Psi(P) = \psi = (\psi(m) : m = 1, \dots, M)$, such as means, differences in means, correlations, and parameters in linear models, generalized linear models, survival models, time-series models, dose-response models, etc. One distinguishes between two types of testing problems: *one-sided tests*, where $H_0(m) = \mathbb{I}(\psi(m) \leq \psi_0(m))$, and *two-sided tests*, where $H_0(m) = \mathbb{I}(\psi(m) = \psi_0(m))$. The hypothesized *null values*, $\psi_0(m)$, are frequently zero.

Let $\mathcal{H}_0 = \mathcal{H}_0(P) \equiv \{m : H_0(m) = 1\} = \{m : P \in \mathcal{M}(m)\}$ be the set of $h_0 \equiv |\mathcal{H}_0|$ true null hypotheses, and note that \mathcal{H}_0 depends on the data generating distribution P . Let $\mathcal{H}_1 = \mathcal{H}_1(P) \equiv \mathcal{H}_0^c(P) = \{m : H_1(m) = 1\} = \{m : P \notin \mathcal{M}(m)\}$ be the set of $h_1 \equiv |\mathcal{H}_1| = M - h_0$ false null hypotheses, i.e., true positives. The goal of a multiple testing procedure is to accurately estimate the set \mathcal{H}_0 , while controlling probabilistically the number of false positives at a user-supplied level α .

Test statistics. A testing procedure is a data-driven rule for deciding whether or not to *reject* each of the M null hypotheses $H_0(m)$, i.e., declare that $H_0(m)$ is false (zero) and hence $P \notin \mathcal{M}(m)$. The decisions to reject the null hypotheses are based on an M -vector of *test statistics*, $T_n = (T_n(m) : m = 1, \dots, M)$, that are functions of the data, X_1, \dots, X_n . Denote the typically unknown (finite sample) *joint distribution* of the test statistics T_n by $Q_n = Q_n(P)$.

Single-parameter null hypotheses are commonly tested using *t-statistics*, i.e., standardized differences,

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null value}}{\text{Standard error}} = \sqrt{n} \frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}. \quad (1)$$

In general, the M -vector $\psi_n = (\psi_n(m) : m = 1, \dots, M)$ denotes an asymptotically linear *estimator* of the parameter M -vector $\psi = (\psi(m) : m = 1, \dots, M)$ and $(\sigma_n(m) : m = 1, \dots, M)$ denote consistent estimators of the asymptotic variances of $(\sqrt{n}(\psi_n(m) - \psi(m)) : m = 1, \dots, M)$. For tests of means, one recovers the usual one-sample and two-sample t -statistics, where the $\psi_n(m)$ and $\sigma_n(m)$ are based on sample means and variances, respectively. In some settings, it may be appropriate to use (unstandardized) *difference statistics*, $T_n(m) \equiv \sqrt{n}(\psi_n(m) - \psi_0(m))$. Test statistics for other types of null hypotheses include F -statistics, χ^2 -statistics, and likelihood ratio statistics.

Multiple testing procedure. A *multiple testing procedure* (MTP) provides *rejection regions*, $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the null hypothesis $H_0(m)$. In other words, a MTP produces a random (i.e., data-dependent) subset \mathcal{R}_n of rejected hypotheses that estimates \mathcal{H}_1 , the set of true positives,

$$\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \equiv \{m : H_0(m) \text{ is rejected}\} = \{m : T_n(m) \in \mathcal{C}_n(m)\}, \quad (2)$$

where $\mathcal{C}_n(m) = \mathcal{C}(T_n, Q_{0n}, \alpha)(m)$, $m = 1, \dots, M$, denote possibly random rejection regions. The long notation $\mathcal{R}(T_n, Q_{0n}, \alpha)$ and $\mathcal{C}(T_n, Q_{0n}, \alpha)(m)$ emphasizes that the MTP depends on: (i) the *data*, X_1, \dots, X_n , through the M -vector of *test statistics*, $T_n = (T_n(m) : m = 1, \dots, M)$; (ii) a test statistic *null distribution*, and (iii) the *nominal level* α of the MTP, i.e., the desired upper bound for a suitably defined false positive rate.

Unless specified otherwise, it is assumed that large values of the test statistic $T_n(m)$ provide evidence against the corresponding null hypothesis $H_0(m)$, that is, we consider rejection regions of the form $\mathcal{C}_n(m) = (c_n(m), \infty)$, where $c_n(m)$ are to-be-determined *cut-offs*, or *critical values*.

Type I and Type II errors. In any testing situation, two types of errors can be committed: a *false positive*, or *Type I error*, is committed by rejecting a true null hypothesis, and a *false negative*, or *Type II error*, is committed when the test procedure fails to reject a false null hypothesis. The number of Type I errors is $V_n \equiv \sum_{m \in \mathcal{H}_0} \mathbb{I}(T_n(m) \in \mathcal{C}_n(m)) = |\mathcal{R}_n \cap \mathcal{H}_0|$ and the number of Type II errors is $U_n \equiv \sum_{m \in \mathcal{H}_1} \mathbb{I}(T_n(m) \notin \mathcal{C}_n(m)) = |\mathcal{R}_n^c \cap \mathcal{H}_1|$. Note that both U_n and V_n depend on the unknown data generating distribution P through the unknown set of true null hypotheses $\mathcal{H}_0 = \mathcal{H}_0(P)$. The numbers $h_0 = |\mathcal{H}_0|$ and $h_1 = |\mathcal{H}_1| = M - h_0$ of true and false null hypotheses are *unknown parameters*, the number of rejected hypotheses $R_n \equiv \sum_{m=1}^M \mathbb{I}(T_n(m) \in \mathcal{C}_n(m)) = |\mathcal{R}_n|$ is an *observable random variable*, and the entries in the body of the table, U_n , $h_1 - U_n$, V_n , and $h_0 - V_n$, are *unobservable random variables* (depending on P , through $\mathcal{H}_0(P)$).

Ideally, one would like to simultaneously minimize both the chances of committing a Type I error and a Type II error. Unfortunately, this is not feasible and one seeks a *trade-off* between the two types of errors. A standard approach is to specify an acceptable level α for the Type I error rate and derive testing procedures, i.e., rejection regions, that aim to minimize the Type II error rate, i.e., maximize *power*, within the

Table 1: Type I and Type II errors in multiple hypothesis testing.

		Null hypotheses		
		not rejected	rejected	
Null hypotheses	true	$ \mathcal{R}_n^c \cap \mathcal{H}_0 $	$V_n = \mathcal{R}_n \cap \mathcal{H}_0 $ (Type I errors)	$h_0 = \mathcal{H}_0 $
	false	$U_n = \mathcal{R}_n^c \cap \mathcal{H}_1 $ (Type II errors)	$ \mathcal{R}_n \cap \mathcal{H}_1 $	$h_1 = \mathcal{H}_1 $
		$M - R_n$	$R_n = \mathcal{R}_n $	M

class of tests with Type I error rate at most α .

Type I error rates. When testing multiple hypotheses, there are many possible definitions for the Type I error rate (and power). Accordingly, we adopt a general definition of Type I error rates, as parameters, $\theta_n = \theta(F_{V_n, R_n})$, of the joint distribution F_{V_n, R_n} of the numbers of Type I errors V_n and rejected hypotheses R_n . Such a general representation covers the following commonly-used Type I error rates.

1. *Generalized family-wise error rate* (gFWER), or probability of at least $(k + 1)$ Type I errors, $k = 0, \dots, (h_0 - 1)$,

$$gFWER(k) \equiv Pr(V_n > k) = 1 - F_{V_n}(k). \quad (3)$$

When $k = 0$, the gFWER is the usual *family-wise error rate*, FWER, controlled by the classical Bonferroni procedure.

2. *Per-comparison error rate* (PCER), or expected proportion of Type I errors among the M tests,

$$PCER \equiv \frac{1}{M} E[V_n] = \frac{1}{M} \int v dF_{V_n}(v). \quad (4)$$

3. *Tail probabilities for the proportion of false positives* (TPFP) among the rejected hypotheses,

$$TPFP(q) \equiv Pr(V_n/R_n > q) = 1 - F_{V_n/R_n}(q), \quad q \in (0, 1), \quad (5)$$

with the convention that $V_n/R_n \equiv 0$, if $R_n = 0$.

4. *False discovery rate* (FDR), or expected value of the proportion of false positives among the rejected hypotheses,

$$FDR \equiv E[V_n/R_n] = \int q dF_{V_n/R_n}(q), \quad (6)$$

again with the convention that $V_n/R_n \equiv 0$, if $R_n = 0$.

Note that while the gFWER is a parameter of only the *marginal* distribution F_{V_n} for the number of Type I errors V_n (tail probability, or survivor function, for V_n), the TPPFP is a parameter of the *joint* distribution of (V_n, R_n) (tail probability, or survivor function, for V_n/R_n). Error rates based on the *proportion* of false positives (e.g., TPPFP and FDR) are especially appealing for the large-scale testing problems encountered in genomics, compared to error rates based on the *number* of false positives (e.g., gFWER), as they do not increase exponentially with the number of hypotheses.

1.2 A multiple testing procedure which is optimal at a simple alternative

In this paper we define an optimal multiple testing procedure as a function of the true data generating distribution, where optimality is with respect to power defined as the expected number of true rejections. Subsequently, we solve the constrained (i.e. controlling the Type-I error) maximization problem for control of the per-family error rate. This is a natural generalization of the well-known Neyman-Pearson Lemma for single tests.

Our procedure is geared for testing against simple alternatives. There are many situations where this could be practical. Consider the case where someone has already done an experiment and formulated point estimates of $(\Psi_1(P), \dots, \Psi_M(P))$, for P the data generating distribution. When carrying out a similar experiment with new data, and performing M tests for $(\Psi_1(P) > \Psi_{10}(P), \dots, \Psi_M(P) > \Psi_{M0}(P))$, we could use the previous point estimates to specify an alternative distribution of interest. If that alternative distribution were true, our method would then have more power (greater expected number of true rejections) than any other procedure controlling the per-family error rate, while it would remain a valid multiple testing procedure at any data generating distribution.

We consider the following class of estimates of the true subset indexed by an M -dimensional vector c of cutoff values:

$$\mathcal{H}_1(c | P_n) \equiv (I(T_1(P_n) > c_1), \dots, I(T_M(P_n) > c_M)). \quad (7)$$

This random subset rejects each null hypothesis with a test statistic exceeding the cutoff value specified by the corresponding component of c . We refer to this random subset as the set of *positives*. For symmetrically distributed test statistics T_m one-sided testing also covers two-sided testing by working with the absolute values of T_m .

We will now define cutoff values $c = c(P)$ with optimal power as a parameter of the true data generating distribution P , but first we must mention what we mean by power. In univariate hypothesis testing, the Type-I error is the probability of a false rejection, while the power is the probability of a true rejection. In multiple hypothesis testing, there are many possible definitions of Type-I and Type-II error rates. In this paper, consider the Type-I error rate to be the per-family error rate (PFER), the expected number of false positives. Among the many ways of defining the type-II error, we could use the probability of rejecting all true positives, or the probability of rejecting at least

one true positive. In this paper however, we define power as the expected number of true positives. In what follows, we show how to construct cutoffs that optimize this power, among all cutoffs controlling the (PFER) at a given Type-I error rate. To formalize this last paragraph, we define

$$\begin{aligned}
R(c) = R(c | P_n) &= \sum_{m=1}^M I(T_m(P_n) > c_m) = | \mathcal{H}_1(c | P_n) | \\
V(c) = V(c | P_n, \mathcal{H}_0) &= \sum_{m=1}^M I(T_m(P_n) > c_m, \mathcal{H}_0(m) = 1) \\
S(c) = S(c | P_n, \mathcal{H}_1) &= \sum_{m=1}^M I(T_m(P_n) > c_m, \mathcal{H}_1(m) = 1).
\end{aligned}$$

Thus R denotes the number of rejected hypotheses, while V and S are the number of false and true positives, respectively. Note that $R = V + S$ is the number of false positives and true positives in $\mathcal{H}_1(c | P_n)$,

Let P_0 be a probability distribution of X satisfying $\Psi_m(P_0) = \Psi_{m0}$ for all $m = 1, \dots, M$, which satisfies the overall null hypothesis. We will study the following optimal choice of $c(P)$ in this paper: the cutoffs that maximize the expected number of true positives (under P), while controlling the expected number of false positives (under P_0) at a user-supplied level q :

$$c(P | q) \equiv \max_{\{c: E_{P_0} R(c | P_n) \leq q\}}^{-1} E_P S(c | P_n, \mathcal{H}_1). \quad (8)$$

Note that $E_P S(c | P_n, \mathcal{H}_1) = \sum_{m \in \mathcal{H}_1} \Phi_m(c_m)$ and $E_{P_0} R(c | P_n) = \sum_m \Phi_m^0(c_m)$, where

$$\begin{aligned}
\Phi_m(c) &\equiv P(T_m(P_n) > c) \\
\Phi_m^0(c) &\equiv P_0(T_m(P_n) > c).
\end{aligned}$$

Thus, with our definition of Type-I error control and power, the choice P_0 only needs to specify the marginal survival function $\Phi_m^0(c)$ of the test statistics $T_m(P_n)$ under P_0 . In the notation above, dependence on n is suppressed for $c(P | q)$, $\Phi_m(c)$, and $\Phi_m^0(c)$.

As an example, consider the important situation in which $T_m(P_n) = (\Psi_{mn} - \Psi_{m0})/(\sigma_{mn}/\sqrt{n})$, where Ψ_{mn} is an estimator of the hypothesised parameter Ψ_m and σ_{mn}/\sqrt{n} is an estimate of the standard error of Ψ_{mn} , $m = 1, \dots, M$. In this case, the Central Limit Theorem teaches us that, for n large enough, $\Phi_m^0(t) \approx \Phi(t)$, and $\Phi_m(t) \approx \Phi(t - d_m)$, where Φ is the survivor function of a standard normal distribution and $d_m = \sqrt{n}(\Psi_m - \Psi_{m0})/\sigma_m$ is a shift parameter, $m = 1, \dots, M$. Let $d(P) = (d_1(P), \dots, d_M(P))$ be the vector of shifts defining the alternative. In this case $c(P) = c(d(P))$ only depends on the true data generating distribution through the shift vector $d(P)$. In Section 2, we present a closed form expression for $c(P)$.

We will assume that the test statistics obey the following natural requirement (i.e. the test statistics are stochastically larger under the null distribution P_0 than under the true distribution P), called the *null domination condition*:

$$\Phi_m(\cdot) \leq \Phi_m^0(\cdot) \text{ for all } m \in \mathcal{H}_0(P). \quad (9)$$

Note that (9) implies that the expected number of false positives $E_P V(c | P_n, \mathcal{H}_0(P))$ in $\mathcal{H}_1(c | P_n)$ is bounded by the controlled quantity $E_{P_0} R(c | P_n)$ for all c . In other words, for any c with $E_{P_0} R(c | P_n) \leq q$ we have,

$$E_P V(c | P_n, \mathcal{H}(P)) \leq q.$$

Thus, $c(P)$ controls the expected number of false positives in our set $\mathcal{H}_1(c(P) | P_n)$ of positives.

1.3 Organization

In Section 2, we present a theorem which provides an analytical simple characterization of the optimal cutoffs $c(P)$. Subsequently, we apply this theorem to obtain closed form expressions for $c(P)$ in the classical contexts of shift alternatives for 1) standard Normal distribution, 2) t-distribution, 3) logistic distribution, and 4) Chi-square distribution, respectively. Results for these four distributions are presented and proved in the Appendix. In Section 3, we discuss Bayesian procedures, where a testing method is provided when there is a prior distribution on the true alternatives. In Section 4, we numerically investigate to what degree guessing the true alternative parameters results in a more powerful test than the common threshold test. In Section 5, we present a numerical study investigating the difference in power of the optimal test and the common threshold test for normally distributed test statistics and shift-alternatives. We end this paper with a discussion about the possible practical use of these optimal testing procedures.

2 A multiple testing procedure which is optimal at a simple alternative

Below we give the main theorem of the paper, which provides the optimal multiple testing procedure (that controls the per-family error rate) as a function of the true data generating distribution. Here optimality is with respect to power defined as the expected number of true rejections.

2.1 The main theorem.

Theorem 2.1. *Let $q < M$. Let $\Phi_m^0(t) = P_0(T_m > t)$ be the survival function of T_m under the null distribution P_0 and $\Phi_m(t) = P(T_m > t)$ be the survival function of T_m under the true data generating distribution P . Assume $\Phi_m(t) \geq \Phi_m^0(t)$, $m \in \mathcal{H}_1$, and the null domination condition $\Phi_m(t) \leq \Phi_m^0(t)$, $m \in \mathcal{H}_0$. Let*

$$\begin{aligned} c_m(\lambda) &= \infty \text{ if } m \in \mathcal{H}_0(P). \\ c_m(\lambda) &= \max_x^{-1} \Phi_m(x) - \lambda \Phi_m^0(x) \text{ if } m \in \mathcal{H}_1(P). \end{aligned}$$

For each $m \in \mathcal{H}_1(P)$, if $x = c_m(\lambda)$ is finite and Φ_m^0, Φ_m are twice differentiable with density $f_m(t) \equiv -\frac{d}{dt}\Phi_m(t)$ and $f_m^0(t) \equiv -\frac{d}{dt}\Phi_m^0(t)$, then:

$$-f_m(x) + \lambda f_m^0(x) = 0 \quad (10)$$

$$f_m'(x) - \lambda f_m^{0'}(x) < 0, \quad (11)$$

where $f_m', f_m^{0'}$ are the derivatives of f_m, f_m^0 .

If $\lambda > 0$ solves $E_P[R(c(\lambda))] - q = 0$, then $E_P S(c(P) | q) = E_P S(c(\lambda))$. Thus, if $c(P)$ is unique, then $c(P) = c(\lambda)$.

Proof. The proof of this theorem requires a generalization of the Lagrange multiplier method to handle situations in which the constrained maximization problem is solved by points on the edge of the parameter space (i.e. ∞ or $-\infty$) so that the derivative cannot be set equal to zero.

Define the function $g : \mathbb{R}^{M+1} \rightarrow \mathbb{R}$

$$\begin{aligned} g(c, \lambda) &= E_P S(c) - \lambda(E_{P_0} R(c) - q) \\ &= \sum_{m \in \mathcal{H}_1} \Phi_m(c_m) - \lambda \left(\sum_{m=1}^M \Phi_m^0(c_m) - q \right). \end{aligned}$$

By the fact that $g(\cdot, \cdot)$ is an additive function in functions of c_m it follows that $c(\lambda) = \max_c^{-1} g(c, \lambda)$, where $c_m(\lambda) = \max_x^{-1} \Phi_m(x) - \lambda \Phi_m^0(x)$ for $m \in \mathcal{H}_1$ and $c_m(\lambda) = \max_x^{-1} (-\lambda \Phi_m^0(x))$ for $m \in \mathcal{H}_0$.

Since λ solves $\sum_{m=1}^M \Phi_m^0(c_m(\lambda)) - q = 0$, this implies

$$E_P S(c(\lambda)) = g(c(\lambda), \lambda) \geq g(c(P), \lambda) = E_P S(c(P)).$$

By definition of $c(P)$, we also have $E_P S(c(P)) \geq E_P S(c(\lambda))$. This proves $E_P S(c(P)) = E_P S(c(\lambda))$.

If $\lambda < 0$, then $c_m(\lambda) = -\infty$ for $m = 1, \dots, M$ which means that $\sum_{m=1}^M \Phi_m^0(c_m(\lambda)) = M$. Therefore, if $q < M$, then we can exclude $c(\lambda), \lambda < 0$, as possible solutions. If $\lambda > 0$, then $c_m(\lambda) = \infty$ for $m \in \mathcal{H}_0$, as stated in the theorem. Finally, (10) is just applying that a finite maximum of a twice differentiable function satisfies that the derivative at the maximum equals zero and the second derivative at the maximum is negative. \square

2.2 Application to shift alternatives

In this subsection we apply the theorem to the case that the actual density f_m of the test statistic T_m satisfies $f_m(x) = f^0(x - d_m)$ for some shift d_m and common smooth null-density f^0 with survival function Φ^0 , where $\mathcal{H}_0(P) = (I(d_1 \leq 0), \dots, I(d_M \leq 0))$. Theorem 2.1 teaches us that the solutions $c(P)$ can be solely expressed in terms of the solution of one maximization problem:

$$c_m(\lambda) \equiv \max_x^{-1} \Phi^0(x - d) - \lambda \Phi^0(x). \quad (12)$$

Specifically, we have

$$\begin{aligned} c_m(\lambda) &= \infty \text{ if } d_m \leq 0 \\ c_m(\lambda) &= m(d_m, \lambda) \text{ if } d_m > 0 \end{aligned}$$

and λ is obtained by solving

$$0 = \sum_{m=1}^M \Phi^0(c_m(\lambda)) - q.$$

The univariate maximization problem (12) is handled by 1) setting the derivative equal to zero, 2) if a solution exists, then we check if it is a maximum (i.e. second derivative is negative) and 3) if no solution exists, then the derivative is either always positive or always negative.

Theorem 2.2. (*Normal distribution*) Consider the setting of Theorem 2.1 with $\Phi_m(x) = \Phi^0(x - d_m)$ and $f_m(x) = f^0(x - d_m)$, where $f^0(x) = 1/\sqrt{2\pi} \exp(-x^2/2)$ and Φ^0 are the respective density and survival functions of a standard Normal distribution, $m = 1, \dots, M$.

Define

$$g(d, \lambda) = \begin{cases} \frac{\log(\lambda) + 0.5d^2}{d} & \text{if } d > 0 \\ \infty & \text{if } d \leq 0. \end{cases}$$

We have that $E_P S(c(P | q)) = E_P S(c(\lambda))$, where 1) if $d_m \leq 0$, then $c_m = \infty$, 2) if $d_m > 0$, then $c_m(\lambda) = g(d_m, \lambda)$ and 3) $\lambda > 0$ is the unique solution of $\sum_m \Phi^0(c_m(\lambda)) - q = 0$

Proof. We apply Theorem 2.1. Thus $c_m = \infty$ if $d_m \leq 0$. We need to find $c_m(\lambda) = \max_x^{-1} \Phi^0(x - d_m) - \lambda \Phi^0(x)$ for $m \in \mathcal{H}_1$. Setting the derivative equal to zero yields that $c_m(\lambda) = g(d_m, \lambda)$ for $m \in \mathcal{H}_1$. In addition, at this solution we have that the second derivatives $\frac{d^2}{dc_m^2} \Phi^0(c_m - d_m) - \lambda \Phi^0(c_m) = -d_m f^0(c_m - d_m) < 0$ for $m \in \mathcal{H}_1$ are strictly negative if and only if $d_m > 0$. Thus $c_m(\lambda)$ is indeed the wished unique maximum. Now, the application of theorem 2.1 yields the proof of the first result about $c(\lambda)$. \square

We note from this proof that the common cutoff vector, where each of the M cutoffs has the same value, is optimal against constant alternatives (d, \dots, d) .

3 Bayesian approach

So far our results have only applied to the case where there is a simple known alternative distribution. There is a more practical method of considering prior distributions on alternatives. Because we know how to carry out the optimal test for one alternative, it might make more sense to carry out the optimal tests for many alternatives, and combine the results.

We propose to draw P_1, \dots, P_B from Θ , where Θ is a prior distribution on all alternatives of interest. From these B probability distributions, we calculate the optimal cutoff vectors $c(P_1|q), \dots, c(P_B|q)$ and output the proportion of rejected null

hypotheses, $\frac{1}{B} \sum_{b=1}^B I(T_m(P_n) > c_m(P_b|q))$ for $m = 1, \dots, M$. Note that because $\sum_{m=1}^M P_0(T_m(P_n) > c_m(P_b|q)) \leq q$,

$\sum_{m=1}^M \frac{1}{B} \sum_{b=1}^B E_{P_0}[I(T_m(P_n) > c_m(P_b|q))] = \frac{1}{B} \sum_{b=1}^B \sum_{m=1}^M P_0(T_m(P_n) > c_m(P_b|q)) \leq q$. That is, the Type-I error constraint is still satisfied. Although the Monte Carlo step of drawing P_b from Θ injects variability into the procedure, the law of large numbers implies that for fixed X_1, \dots, X_n , our test result converges almost surely to $\sum_{m=1}^M \int I(T_m(P_n) > c_m(P|q)) d\Theta(P)$ as B tends to infinity. For a fixed alternative P_1 , by Fubini's Theorem, this method has power given by:

$$\begin{aligned} & \sum_{m=1}^M I(\Psi_m(P_1) > \Psi_{m0}) E_{P_1}[\int I(T_m(P_n) > c_m(P|q)) d\Theta(P)] \\ &= \sum_{m=1}^M I(\Psi_m(P_1) > \Psi_{m0}) \int P_1(T_m(P_n) > c_m(P|q)) d\Theta(P). \end{aligned}$$

Heuristically, if Θ is close in some sense to the distribution putting a point mass on P_1 and $P \sim \Theta$, then $c_m(P|q)$ should be close to $c_m(P_1|q)$ and this power should be close to the optimal power.

When implementing this procedure, one would draw many alternatives for a given prior distribution, calculate the M test results for each draw, and output the proportion of rejections for each of the M tests over all draws. One difference between this approach and the approach of the previous section is that the outputs are now proportions of rejections, instead of binary outcomes indicating acceptances or rejections.

4 Sample-splitting procedure

In practice the test statistics are often functions of n i.i.d. observations $X_1, \dots, X_n \sim P$. Consider the Normal shift model where X_{1m}, \dots, X_{nm} are i.i.d. $N(n^{-1/2}\Psi_m, 1)$, $m = 1, \dots, M$, and where $T_m \equiv T_m(P_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{im}$. Then $T_m \sim N(\Psi_m, 1)$, and we reject $\Psi_m = 0$ in favor of $\Psi_m > 0$ when T_m is large. If the alternatives Ψ_m are given, we can perform multiple testing as in Section 2. We form two sets of observations, $\{X_{1,m}, \dots, X_{n(1-p),m}\}$ and $\{X_{n(1-p)+1,m}, \dots, X_{n,m}\}$, for $0 < p < 1$, use the first sample to guess the alternatives, and use the second sample for testing.

We call the test statistics from the two samples $T_m^{(1)} = T_m^{(1)}(P_n)$ and $T_m^{(2)} = T_m^{(2)}(P_n)$, with $T^{(i)} = (T_1^{(i)}, \dots, T_M^{(i)})$, $i = 1, 2$, and note that $T_m^{(1)} \sim N(\Psi_m, \frac{1}{1-p})$ while $T_m^{(2)} \sim N(\Psi_m, \frac{1}{p})$. Also note that $T^{(1)}$ and $T^{(2)}$ are independent. For μ an M -component vector and σ^2 a positive scalar, let $c(\mu, \sigma^2)$ denote the mapping into the M optimal cutoffs for normal testing, where the test statistics have variance σ^2 and the alternatives are given by μ . We propose to use $T^{(2)}$ as our test statistics, with cutoffs defined by $c(T^{(1)}, \frac{1}{p})$.

Theorem 4.1. *The above sample-splitting method that rejects the null hypothesis $\Psi_m = 0$ for $T_m^{(2)} > c_m(T^{(1)}, \frac{1}{p})$ has the desired Type-I error control. That is, $E_{P_0} R(c) \leq q$.*

Proof. Let Φ denote the standard Normal cdf.
 $E_{P_0} R(c) = \sum_{m=1}^M P_0(T_m^{(2)} > c_m(T^{(1)}, \frac{1}{p}))$
 $= \sum_{m=1}^M E_0[E_0[I(T_m^{(2)} > c_m(T^{(1)}, \frac{1}{p}) | T^{(1)}]],$ by conditioning on $T^{(1)}$.

Table 2: This table reports the expected number of true positives for the optimal cutoffs, the common cutoffs $c = Z_{1-q/M}$, and the cutoffs based on sample splitting. The results are reported for five different values of the proportion ϵ of false null hypotheses. The sample-splitting method is governed by p (the proportion of the sample used for the testing). Here $M=100$ and $q=5$.

ϵ	p	optimal cutoffs	common cutoffs	sample-splitting cutoffs
1/6	1/6	8.1	2.0	5.5
1/6	2/6			6.6
1/6	3/6			3.3
1/6	4/6			4.6
1/6	5/6			3.6
2/6	1/6	10.3	5.2	6.5
2/6	2/6			6.6
2/6	3/6			7.3
2/6	4/6			8.2
2/6	5/6			7.2
3/6	1/6	10.6	5.7	6.5
3/6	2/6			7.1
3/6	3/6			7.9
3/6	4/6			8.9
3/6	5/6			9.2
4/6	1/6	11.6	8.5	6.6
4/6	2/6			7.4
4/6	3/6			8.4
4/6	4/6			8.9
4/6	5/6			8.7
5/6	1/6	12.2	10.5	6.8
5/6	2/6			7.8
5/6	3/6			8.3
5/6	4/6			9.4
5/6	5/6			9.8

$$= \sum_{m=1}^M E_0[(1 - \Phi(\frac{c_m(T^{(1)}, \frac{1}{p})}{\sqrt{\frac{1}{p}}}))] \text{ as } T^{(1)} \perp T^{(2)} \text{ and } T_m^{(2)} \sim N(0, \frac{1}{p}) \text{ when } \Psi_m = 0.$$

As the constraint is satisfied almost surely inside the expectation by the construction of $c(\cdot, \frac{1}{p})$ in 2.2, this establishes the desired control. \square .

We now give an expression for the power of the sample splitting method. Recall that \mathcal{H}_1 is defined as the set of indices for the false null hypotheses.

$$\begin{aligned} E_P S(c) &= \sum_{m \in \mathcal{H}_1} E_P[I(T_m^{(2)} > c_m(T^{(1)}, \frac{1}{p}))] \\ &= \sum_{m \in \mathcal{H}_1} E_P[E_P[I(T_m^{(2)} > c_m(T^{(1)}, \frac{1}{p})|T^{(1)}]], \text{ by conditioning on } T^{(1)}. \\ &= \sum_{m \in \mathcal{H}_1} E_P[(1 - \Phi(\frac{c_m(T^{(1)}, \frac{1}{p}) - \Psi_m}{\sqrt{\frac{1}{p}}}))], \text{ as } T^{(1)} \perp T^{(2)} \text{ and } T_m^{(2)} \sim N(\Psi_m, \frac{1}{p}) \text{ under } P. \end{aligned}$$

This power has an interesting interpretation. Writing the expectation as $E_P f(T^{(1)})$, we might hope that $E_P f(T^{(1)}) \approx f(E_P T^{(1)}) = f(\Psi)$. If this were the case, then the power would be approximately equal to the optimal power against the true alternatives Ψ , for testing with statistics $T^{(2)}$ whose variance is now $\frac{1}{p}$ instead of one. However, the function $f(\cdot)$ is highly nonlinear, so we would only expect the approximation to be close if the variance of $T^{(1)}$ were small. This variance is $\frac{1}{1-p}$, so we would need p to be small, implying a reduced approximate optimal power. Hence, choosing the value of p requires a bias/variance type tradeoff when trying to maximize power.

4.1 Simulation study

In Table 1 we display the results from our simulation, comparing the sample-splitting method to the common-cutoff method for the Normal model of the previous section. We consider the case of $M = 100$ hypothesis with ϵM true positives, with $q = 5$ allowed expected false positives. Here the ϵM true positives have $\Psi_m = 1/2$. The remaining $(1 - \epsilon M)$ parameters have $\Psi_m = -100$, so the null is true. These represent hypotheses where there is no ambiguity as to whether the null is true, so that splitting the sample allows one to determine these nulls as true, and power can be spread out over the other hypotheses. For small ϵ , we see that our sample-splitting method outperforms the common-cutoff rule. We also vary p , the proportion of the sample used for testing in the sample-splitting method. From Table 1, it is clear that the method generally improves as p increases. As ϵ increases toward one, the alternatives became close to a common value, and the common-cutoff rule begins to improve upon the sample-splitting rule.

5 Numerical comparison of power of the common-cutoffs and (misspecified) optimal cutoffs

Consider the context in which the test statistics T_m are Normally distributed with mean d_m and variance 1, $m = 1, \dots, 1000$. In this section we compare the power of the multiple testing procedures corresponding with the common threshold, the optimal threshold at true alternative d and the optimal threshold at a wrong alternative \tilde{d} for the 10 choices of d and \tilde{d} listed below. Here $rep(a, M)$ denotes a vector of length M

Table 3: This table reports the expected number of true positives for the optimal threshold $c_{1,opt}(d, q)$, the misspecified optimal threshold $c_{1,opt}(\tilde{d}, q)$ and the common threshold $c = Z_{1-q/M}$ at the 10 choices (d, \tilde{d}) specified for Type-I error $E_0R(c) \leq q = 0.1$.

$c_{1,opt}(d, q = 0.1)$	$c_{1,opt}(\tilde{d}, q = 0.1)$	$c(q = 0.1)$
16.03	15.90	7.61
37.82	36.71	26.56
15.90	15.45	6.52
448.94	447.99	434.98
147.29	146.83	119.68
16.13	14.68	9.09
142.86	127.14	113.27
49.02	46.90	47.65
883.65	882.13	883.29
149.38	147.77	129.55

Table 4: This table reports the expected number of true positives for the optimal threshold $c_{1,opt}(d, q)$, the misspecified optimal threshold $c_{1,opt}(\tilde{d}, q)$ and the common threshold $c = Z_{1-q/M}$ at the 10 choices (d, \tilde{d}) specified for Type-I error $E_0R(c) \leq q = 0.05$.

$c_{1,opt}(d, q = 0.05)$	$c_{1,opt}(\tilde{d}, q = 0.05)$	$c(q = 0.05)$
13.72	13.71	5.66
27.58	26.23	18.77
13.71	13.63	5.00
433.67	433.09	417.14
118.03	117.84	94.25
13.73	12.39	6.53
130.07	117.43	100.06
34.83	32.34	33.25
849.43	848.29	849.35
119.05	117.88	101.10

Table 5: This table reports the expected number of true positives for the optimal threshold $c_{1,opt}(d, q)$, the misspecified optimal threshold $c_{1,opt}(\tilde{d}, q)$ and the common threshold $c = Z_{1-q/m}$ at the 10 choices (d, \tilde{d}) specified for Type-I error $E_0R(c) \leq q = 2$.

$c_{opt}(d, q = 2)$	$c_{opt}(\tilde{d}, q = 2)$	$c(q = 2)$
40.91	36.99	30.19
138.06	138.07	108.28
27.55	26.58	22.18
499.24	496.78	495.75
318.16	312.88	289.35
45.96	45.25	43.15
186.57	153.56	172.74
199.08	198.54	198.93
975.96	973.94	972.20
347.29	344.67	329.28

Table 6: This table reports the expected number of true positives $ES(c)$ for the optimal threshold $c_{1,opt}(d, q)$, the misspecified optimal threshold $c_{1,opt}(\tilde{d}, q)$ and the common threshold $c = Z_{1-q/M}$ at the 10 choices (d, \tilde{d}) specified for Type-I error $E_0R(c) \leq q = 1$.

$c_{opt}(d, q = 1)$	$c_{opt}(\tilde{d}, q = 1)$	$c(q = 1)$
30.85	27.73	21.50
104.10	103.95	79.47
23.33	22.14	16.26
486.52	484.58	482.40
274.27	270.12	241.17
33.50	32.56	29.43
178.84	146.77	158.69
146.42	145.51	145.94
962.16	959.95	959.21
291.35	288.69	271.01

with each component equal to a and, given vectors x, y , $c(x, y)$ denotes the concatenated vector (x, y) . We see from the tables that the misspecified optimal procedure almost always outperforms testing with the common threshold, sometimes by a substantial margin, and often comes close to the unattainable (correctly specified) optimal procedure. It should be mentioned that these improvements rely on prior information regarding the relative ordering of the alternatives d_m : that is, if one guesses alternatives having totally different ordering than the true values of d_m , then this procedure will be less powerful than the common threshold procedure.

$$\begin{aligned}
d[1,] &= c(rep(-1, 250), rep(0.5, 250), rep(1, 475), rep(3, 25)) \\
\tilde{d}[1,] &= c(rep(-1, 250), rep(0.5, 250), rep(0.5, 475), rep(3, 25)) \\
d[2,] &= c(rep(0.5, 250), rep(0.5, 250), rep(2, 475), rep(3, 25)) \\
\tilde{d}[2,] &= c(rep(0.05, 250), rep(0.05, 250), rep(3, 475), rep(3, 25)) \\
d[3,] &= c(rep(0.5, 975), rep(3, 25)) \\
\tilde{d}[3,] &= c(rep(1, 975), rep(3, 25)) \\
d[4,] &= c(rep(1, 500), rep(5, 475), rep(3, 25)) \\
\tilde{d}[4,] &= c(rep(1.5, 500), rep(4, 475), rep(3, 25)) \\
d[5,] &= c(rep(1, 500), rep(3, 500)) \\
\tilde{d}[5,] &= c(rep(1.5, 500), rep(2.5, 500)) \\
d[6,] &= c(rep(1, 975), rep(3, 25)) \\
\tilde{d}[6,] &= c(rep(1.5, 975), rep(3, 25)) \\
d[7,] &= c(rep(0.5, 800), rep(4, 175), rep(3, 25)) \\
\tilde{d}[7,] &= c(rep(1, 800), rep(6, 175), rep(3, 25)) \\
d[8,] &= c(rep(2, 975), rep(3, 25)) \\
\tilde{d}[8,] &= c(rep(1.5, 975), rep(3, 25)) \\
d[9,] &= c(rep(5, 975), rep(3, 25)) \\
\tilde{d}[9,] &= c(rep(4, 975), rep(2.5, 25)) \\
d[10,] &= c(rep(1, 250), rep(2, 250), rep(3, 500)) \\
\tilde{d}[10,] &= c(rep(0.5, 250), rep(2.5, 250), rep(3.5, 500)).
\end{aligned}$$

6 Discussion.

The multiple testing procedure defined in theorem 2.1 is optimal against simple known alternatives. Our simulation results show that our procedure works well when one has accurate prior information about the true data generating distribution, but can be improved upon by the common threshold procedure when the prior information is inaccurate. Our Bayesian method allows us to incorporate prior knowledge about the data generating distribution, so that our procedure can be used to test against classes of alternatives. Estimating the alternatives with part of the sample, and performing the

tests on a different part, works well in some cases, but does not significantly outperform testing with common cutoffs.

APPENDIX: Optimal cutoffs for shift alternatives under Student's t-distribution, a Logistic distribution, and a Chi-Square distribution.

Theorem 6.1. (*Student's t-distribution*) Consider the setting of theorem 2.1 and with $\Phi_m(x) = \Phi^0(x - d_m)$, $f_m(x) = f^0(x - d_m)$, $f^0(x) = c(k) \frac{1}{\{1+x^2/(2k-1)\}^k}$, $c = r/2 - 1$, Φ^0 are the density and survivor function of a of a Students' t-distribution with $2k - 1$ degrees of freedom.

We define:

$$\begin{aligned}
e(d) &= \frac{2M + d^2}{M} \\
\lambda_k &\equiv \lambda^{1/k} \\
x_1(d, \lambda) &\equiv \frac{-2\lambda_k d/M - \sqrt{4/M(e(d)\lambda_k - \lambda_k^2 - 1)}}{2(1 - \lambda_k)/M} \\
x_2(d, \lambda) &\equiv \frac{-2\lambda_k d/M + \sqrt{4/M(e(d)\lambda_k - \lambda_k^2 - 1)}}{2(1 - \lambda_k)/M} \\
B_1(d, \lambda) &\equiv I \left(\frac{d - \sqrt{d^2 + 4M}}{2} < x_1(d, \lambda) < \frac{d + \sqrt{d^2 + 4M}}{2} \right) \\
B_2(d, \lambda) &\equiv I \left(\frac{d - \sqrt{d^2 + 4M}}{2} < x_2(d, \lambda) < \frac{d + \sqrt{d^2 + 4M}}{2} \right) \\
B(d, \lambda) &\equiv I \left(\frac{e(d) - \sqrt{e(d)^2 - 4}}{2} < \lambda_k < \frac{e(d) + \sqrt{e(d)^2 - 4}}{2} \right).
\end{aligned}$$

Define

$$m(d, \lambda) \equiv \begin{cases} \infty & \text{if } \lambda_k > \frac{e(d) - \sqrt{e(d)^2 - 4}}{2} \\ -\infty & \text{if } \lambda_k < \frac{e(d) + \sqrt{e(d)^2 - 4}}{2} \\ x_1(d, \lambda) & \text{if } B_1(d, \lambda) = B(d, \lambda) = 1. \\ x_2(d, \lambda) & \text{if } B_2(d, \lambda) = B(d, \lambda) = 1. \end{cases}$$

We have $ES(c(P)) = ES(c(\lambda^*))$, where 1) if $d_m \leq 0$, then $c_m = \infty$, 2) if $d_m > 0$, then $c_m(\lambda) = m(d_m, \lambda)$ and λ^* is the solution of $\sum_m \Phi(c_m(\lambda)) - q = 0$.

Proof. Just solve the equation $-f_m(x_m - d_m) + \lambda f_m(x_m) = 0$. Now, two solutions x_{1m}, x_{2m} exist under a constraint $B_m(\lambda) = 1$. If $B_m(\lambda) = 0$, then either $-f_m(x_m - d_m) + \lambda f_m(x_m) = 0$ is always negative or always positive. In the first case, we know the maximum is $c_m = -\infty$ and in the second case the maximum is $c_m = \infty$. Finally, among the two solutions x_{1m}, x_{2m} we need the solution for which the second derivative is negative which is true for x_{1m} if $B_{1m} = 1$ and for x_{2m} if $B_{2m} = 1$. \square

Theorem 6.2. (Logistic distribution) Consider the setting of theorem 2.1 and with $\Phi_m(x) = \Phi^0(x - d_m)$, $f_m(x) = f^0(x - d_m)$, $f^0(x) = \exp(-x)/(1 + \exp(-x))^2$, Φ^0 are the density and survivor function of a standard logistic distribution. Assume $v(\alpha) \neq 0$.

Define

$$B(d, \lambda) \equiv I(\exp(-d) < \lambda < \exp(d))$$

$$B_1(d, \lambda) \equiv I\left(\lambda > \frac{4}{\{\exp(d/2) + \exp(-d/2)\}^2}\right)$$

Define

$$m(d, \lambda) = \begin{cases} \infty & \lambda > \exp(d). \\ -\infty & \lambda < \exp(-d). \\ \log \left\{ \frac{\sqrt{\lambda} \exp(d/2) - 1}{1 - \sqrt{\lambda} \exp(-d/2)} \right\} & \text{if } B(d, \lambda) = B_1(d, \lambda) = 1 \\ \infty & \text{if } B_1(d, \lambda) = 0 \end{cases}$$

We have $ES(c(P)) = ES(c(\lambda^*))$, where 1) if $d_m \leq 0$, then $c_m = \infty$, 2) if $d_m > 0$, then $c_m(\lambda) = m(d_m, \lambda)$ and λ^* is the solution of $\sum_m \Phi(c_m(\lambda)) - q = 0$.

Proof. If $B_m(\lambda) = 1$, then we find a unique solution $c_m(\lambda)$ of $-f_m(c_m - d_m) + \lambda f_m(c_m) = 0$. If $B_m(\lambda) = 0$, then either $-f_m(x_j - d_m) + \lambda f_m(x_m) = 0$ is always negative or always positive. In the first case, we know the maximum is $c_m = -\infty$ and in the second case the maximum is $c_m = \infty$. Finally, the second derivative of $m(c, \lambda)$ at $c_1(\lambda)$ is strictly negative if and only if $B_{1m}(\lambda) = 1$. So we should only accept $c_m(\lambda)$ as a maximum if $B_{1m}(\lambda) = 1$. Otherwise, it is a minimum and the maximum is attained at ∞ . \square

The next theorem provides the closed form formulas for $c(P)$ for shift-alternatives of the Chi-square distribution. It is common that the test statistic is of the form $T_m = \sum_{l=1}^p Z_l^2$, where $Z_m = N(0, 1)$ under P_0 and $Z_m = N(\Psi_m, 1)$ under the true distribution P . Then, we have that T_m equals $\sum_l (Z_l - \Psi_l)^2 + \sum_{l=1}^p \Psi_l^2 + 2\Psi_l \sum_{l=1}^l Z_l$ in distribution. Thus under P_0 T_m follows a Chi-square distribution, but under the alternative P it does not follow a simple shift of a Chi-square distribution. Therefore, to truly obtain the optimal cutoff in this setting one will need to solve the equation $\max_x^{-1} \Phi(x) - \lambda \Phi^0(x)$, where $\Phi(x)$ is the survivor function of the non-central Chi-square distribution of $\sum_l (Z_l - \Psi_l)^2 + \sum_{l=1}^p \Psi_l^2 + 2\Psi_l \sum_{l=1}^l Z_l$. However, simply using the optimal cutoff values specified in the next theorems for the shift-alternatives $d_j = \sum_{l=1}^p \Psi_l^2$ are likely to give significant improvements on standard Chi-square adjustments.

Theorem 6.3. (Chi-Square distribution) Consider the setting of theorem 2.1 with $\Phi_j(x) = \Phi^0(x - d_m)$, $f_m(x) = f^0(x - d_m)$, $f^0(x) = x^c \exp(-x/2)$, $c = r/2 - 1$, Φ^0 are the density and survivor function of a Chi-square distribution with r degrees of freedom, $m = 1, \dots, M$. Assume $v(\alpha) \neq 0$.

Define

$$m(d, \lambda) \equiv \begin{cases} \frac{d}{1 - \lambda^{1/c} \exp(-d/2c)} & \text{if } \lambda < \exp(d/2) \\ \infty & \text{if } \lambda > \exp(d/2) \end{cases}$$

We have that $ES(c(P)) = ES(c(\lambda^*))$, where 1) if $d_m \leq 0$, then $c_m = \infty$, 2) if $d_m > 0$, then $c_m(\lambda) = m(d_m, \lambda)$ and 3) λ^* is the unique solution of $\sum_m \Phi^0(c_m(\lambda)) - q = 0$

Proof. One just imitates the proof of the previous theorem. One has to use that the second derivative of $x \rightarrow \Phi(x - d) - \lambda\Phi(x)$ at the solution $c(d) = \frac{d}{1 - \lambda^{1/c} \exp(-d/2c)}$ of its derivative is given by:

$$-\lambda \frac{cdf(c(d))}{c(d)(c(d) - d)}.$$

Now, note that the latter is negative if and only if $\lambda < \exp(d/2)$ (this is true for positive and negative d). \square

References

Dudoit, S., van der Laan, M.J, and Pollard, K.S. (2004) *Multiple Testing. Part I. Single-Step Procedures for Control of General Type-I Error Rates*, Statistical Applications in Genetics and Molecular Biology: Vol. 3: No. 1, Article 13.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. John Wiley, New York.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. second edition, Springer-Verlag, New York.

Lehmann, E.L., Romano, J.P., and Shaffer J.P. (2004). *On Optimality of Stepdown and Stepup Multiple Test Procedures*. Accepted to the Annals of Statistics.

Pollard, K.S. and van der Laan, M.J. (2003). *Multiple Testing for Gene Expression Data: an Investigation of Null Distributions with Consequences for the Permutation Test*, Proceedings of the 2003 International MultiConference in Computer Science and Engineering, METMBS'03 Conference, pp.3-9.

Shaffer, J.P. (1995), Multiple hypothesis testing, *Annu. Rev. Psychol.*, **46**, 561–584.

van der Laan, M.J., Dudoit, S., and Pollard, K.S. (2004) *Multiple Testing. Part II. Step-Down Procedures for Control of the Family-Wise Error Rate*, Statistical Applications in Genetics and Molecular Biology: Vol. 3: No. 1, Article 14.

van der Laan, M.J., Dudoit, S., and Pollard, K.S. (2004) *Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives*, Statistical Applications in Genetics and Molecular Biology: Vol. 3: No. 1, Article 15.

Westfall, P. and Young, S. (1993) *Resampling-based Multiple Testing*. John Wiley, New York.