University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 172

Resampling Based Multiple Testing Procedure Controlling Tail Probability of the Proportion of False Positives

Mark J. van der Laan* Merrill D. Birkner† Alan E. Hubbard‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, mbirkner@berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

Resampling Based Multiple Testing Procedure Controlling Tail Probability of the Proportion of False Positives

Mark J. van der Laan, Merrill D. Birkner, and Alan E. Hubbard

Abstract

Simultaneously testing a collection of null hypotheses about a data generating distribution based on a sample of independent and identically distributed observations is a fundamental and important statistical problem involving many applications. In this article we propose a new resampling based multiple testing procedure asymptotically controlling the probability that the proportion of false positives among the set of rejections exceeds q at level alpha, where q and alpha are user supplied numbers. The procedure involves 1) specifying a conditional distribution for a guessed set of true null hypotheses, given the data, which asymptotically is degenerate at the true set of null hypotheses, and 2) specifying a generally valid null distribution for the vector of test-statistics proposed in Pollard and van der Laan (2003), and generalized in our subsequent articles Dudoit et al. (2004), van der Laan et al. (2004a) and van der Laan et al. (2004b). We establish the finite sample rational behind our proposal, and prove that this new multiple testing procedure asymptotically controls the wished tail probability for the proportion of false positives under general data generating distributions. In addition, we provide simulation studies establishing that this method is generally more powerful in finite samples than our previously proposed augmentation multiple testing procedure (van der Laan et al. (2004b)) and competing procedures from the literature. Finally, we illustrate our methodology with a data analysis.

1 Introduction

Recent technological developments in biological research, for instance genomics and proteomics, have created new statistical challenges by providing simultaneously thousands of biological measurements (e.g., gene expressions) on the same experimental unit. Typically, the collection of these measurements is made to determine, for example, which genes of the thousands of candidates are associated with some other, often phenotypic, characteristic (e.g., disease status). This has lead to the problem of properly accounting for simultaneously testing a large number of null hypotheses when making inferences about the tests for which the null is rejected. Multiple testing is a subfield of statistics concerned with proposing decision procedures involving a rejection or acceptance decision for each null hypothesis. Multiple testing procedures are used to control various parameters of either the distribution of the number of false rejections or the proportion of false rejections, and these are often referred to as different varieties of Type-I error rates. In addition, among such procedures controlling a particular Type-I error rate, one aims to find a procedure which has maximal power in the sense that it finds more of the true positives than competing procedures.

One such Type-I error rate is the probability of the proportion of false positives *among the rejections* exceeding a user supplied q (e.g., 0.05). We will refer to this Type-I error as TPPFP(q) which stands for Tail Probability of the Proportion of False Positives at a user defined level q. For example, one might wish to use a multiple testing procedure which satisfies that the proportion of false positives among the rejections is larger than 0.05 with probability $\alpha = 0.05$ (in this case, $q = \alpha = 0.05$). A popular error rate to control in large multiple testing problems is the false discovery rate (FDR) by using, for instance, the Benjamini-Hochberg method. The FDR is defined as the expectation of the proportion of false positives among the rejections. Contrary to a multiple testing procedure controlling the TPPFP(q), a procedure controlling the FDR provides no probabilistic bound that the proportion of false positives is smaller than some cut-off (e.g., 0.05). In this paper, we propose a new method for estimating the TPPFP for specific decisions rules that is asymptotically sharp, but also behaves better and less conservatively than existing methods in finite samples.

Existing TPPFP multiple testing procedures include marginal step-down procedures of Lehmann and Romano (2003), the inversion method of Genovese and Wasserman (2003a,b) for independent test statistics and its conservative version for general dependence structures. These multiple testing procedures are based only on marginal p-values and thereby either rely on 1) assumptions concerning the joint distribution of the test statistics, such as, independence, specific dependence structure (e.g., positive regression dependence, ergodic dependence), and 2) err on the conservative side by using a Bonferroni- type of adjustment. In previous work (van der Laan et al. (2004b), we showed that any single-step or stepwise procedure (asymptotically) controlling the family wise error can be straightforwardly augmented to (asymptotically) control the TPPFP, for general data generating distributions, and hence, arbitrary dependence structures among the test statistics. Specifically, given an initial set of rejections of size r_0 corresponding with a multiple testing procedure controlling the family wise error rate, FWER (FWER is the probability of at least one Type-I error), at level α, one simply adds the next $\lceil \frac{q}{1-q} r_0 \rceil$ most significant tests to the rejection set to control TPPFP(q) at level α. This corresponds to adding rejections to r_0, which are counted as false positives, until the ratio of false positives to total rejections is equal to q. In Dudoit et al. (2004a) we review the above mentioned procedures and compare our augmentation method with the Lehmann and Romano (2003) marginal p-value methods in an extensive simulation study.

In van der Laan et al. (2004b) it is shown that this simple augmentation method controls the TPPFP(q), and, if the FWER-procedure is also asymptotically sharp, then this augmentation procedure is also asymptotically sharp at fixed alternatives. That is, in the latter case it asymptotically controls the proportion of false positives exactly at q with probability exactly equal to α. The main problem occurs in finite samples where this procedure can be too conservative by counting every addition to the FWER-procedure as a false positive. Though, the augmentation procedure compared favorably to the marginal p-value methods referenced above in our finite sample simulations, and theoretically outperforms these methods asymptotically under dependence, our simulations clearly suggested that all methods are conservative in finite samples. Specifically, we found that the augmentation method becomes more conservative as the number of tests increases, which is particularly important in large genomic datasets where there are small numbers of biological replicates but thousands of genes and thus thousands of tests. In this paper, we propose a new multiple testing method controlling TPPFP(q), still asymptotically valid for general data generating distributions (as the augmentation method), but less conservative in finite samples. Our new proposal involves specifying 1) a conditional distribution for a guessed set of true null hypotheses, given the data, which asymptotically is degenerate at the true set of null hypotheses, and 2) a generally valid null distribution for the vector of test-statistics proposed in Pollard and van der Laan (2003), and generalized in our subsequent article Dudoit et al. (2004b); van der Laan et al. (2004a,b).

Regarding 1), we provide an explicit proposal of a distribution of a guessed sets of null hypotheses based on Bernoulli draws with probability being the posterior probability of a null hypothesis being true, given the value of its test-statistic, which is based on the model assuming that the test-statistics are i.i.d. from a mixture of a null density and an alternative density (as in Efron et al. (2001a,b)). Regarding 2), a generally valid null distribution, avoiding the need for the subset-pivotality condition, was originally proposed in Pollard and van der Laan (2003) for tests concerning (general) real valued parameters, and generalized to general hypotheses in our subsequent articles Dudoit et al. (2004b), van der Laan et al. (2004a), van der Laan et al. (2004b). That is, we choose as null distribution, the null-value shifted true distribution of the test-statistics (e.g., centered t-statistic), which conserves the covariance structure of the test-statistics, and thereby guarantees that the number of false rejections under the true distribution is dominated by the number of false rejections under our null distribution. The latter null distribution is naturally estimated with the model based or nonparametric bootstrap. Given a draw of the set of null hypotheses, we draw a new vector of test-statistics by replacing the subvector of test-statistics corresponding with the null hypotheses by a draw of the null distribution, but leaving the remaining test-statistics identical to the observed test-statistics. For each cut-off level, we can now evaluate the proportion of false positives among the set of rejections for this given guessed set of null hypotheses. By randomly sampling sets of null hypotheses and test-statistics from the null distribution, we obtain a distribution of proportion of false positives at any cut-off level. Finally, we fine-tune the cut-off level so that the exceedance probability at q equals α.

In the next section we will describe our method in detail, provide its finite sample rational, and establish the wished formal asymptotic result. In Section 3 we carry out simulation studies comparing this new method to our existing augmentation method based on augmenting a resampling based multiple testing procedure controlling the family wise error rate (FWER), where both methods rely on the same null distribution of the test-statistics (making them nicely comparable). In Section 4 we present a data analysis, and we conclude with a summary.

2 Rational and Method

Throughout this section we will let $T_n = (T_n(1), \ldots, T_n(m))$ be a vector of test-statistics with unknown distribution Q_n corresponding with a set of null hypotheses $H_{01}, \ldots, H_{0m}$ such that large values of $T_n(j)$ provide statistical evidence that the null hypothesis H_{0j} is false, and n indicates the sample size. Here T_n is a test-statistic vector based on a sample of n i.i.d. $X_1, \ldots, X_n$ with a common distribution P so that the distribution $Q_n = Q_n(P)$ of T_n is identified by the data generating distribution P. In addition, $H_{0j} : P \in \mathcal{M}_j$ states that P is an element of a set of probability distributions $\mathcal{M}_j$ for a certain hypothesized subset $\mathcal{M}_j$ of data generating distributions. We will also let $\mathcal{S}_0 \equiv \{j : H_{0j} \text{ is true}\}$ be the set of true null hypotheses.

It will be assumed that there exists a vector of null-values $(\theta_0(j) : j = 1, \ldots, m)$ such that $\limsup_{n \to \infty} ET_n(j) \leq \theta_0(j)$ for $j \in \mathcal{S}_0$. This allows us to specify the generally asymptotically valid null distribution $(T_n(j) - ET_n(j) + \theta_0(j) : j = 1, \ldots, m)$ for the vector of test-statistics, proposed in Pollard and van der Laan (2003), and generalized in Dudoit et al. (2004b). As detailed in these articles, this distribution can be naturally estimated with the bootstrap. This null-value shifted null distribution is an asymptotically valid null distribution in the sense that the distribution of the subvector $(T_n(j) : j \in \mathcal{S}_0)$ is asymptotically dominated by the distribution of the null-value shifted $(T_n(j) - ET_n(j) + \theta_0(j) : j \in \mathcal{S}_0)$ so that probabilistic control of the number of rejections under this null distribution implies the wished asymptotic probabilistic control of the number of false rejections under the true data generating distribution. The null distribution should also be scaled at a null-value (upper bound under the null hypothesis) for the variance under the null hypotheses, in the case that the variance of the null-valued centered test-statistics converges to infinity (Dudoit et al., 2004b).

A possibly data dependent cut-off vector $c_n = (c_n(1), \ldots, c_n(m))$, specifies a multiple testing procedure (i.e., a set of rejections) given by

$$\mathcal{S}_n \equiv \{j : T_n(j) > c_n(j)\} \subset \{1, \ldots, m\}.$$

For simplicity, we will focus on common cut-off vectors, which are appropriate if the test-statistics $T_n(j)$ have a common marginal distribution, $j = 1, \ldots, m$, or at least a common marginal variance. Given user supplied numbers $q, \alpha \in (0,1)$, our goal is to construct a multiple testing procedure such that

$$Pr\left(\frac{\sum_{j=1}^m I(T_n(j) > c_n(j), j \in \mathcal{S}_0)}{\sum_{j=1}^m I(T_n(j) > c_n(j))} > q\right) \leq \alpha. \tag{1}$$

We make the convention that $0/0 = 0$.

That is, we are interested in controlling the probability that the proportion of false positives (Type I errors) to total rejections is greater than a level q, at a level α. In order to explicitly understand the challenge, we consider the common cut-off:

$$c(Q_n, \mathcal{S}_0 \mid q, \alpha) \equiv \inf\{c : \bar{F}_{V_n(c)/R_n(c)}(q) \leq \alpha\},$$

where

$$\begin{aligned} V_n(c) &= V_n(c \mid \mathcal{S}_0) = \sum_{j=1}^m I(T_n(j) > c, j \in \mathcal{S}_0), \\ R_n(c) &\equiv \sum_{j=1}^m I(T_n(j) > c), \end{aligned}$$

are the number of false rejections and number of rejections, respectively. Given a random variable X, $\bar{F}_X(x) \equiv P(X > x)$ denotes the survivor function of the random variable X. Clearly, the multiple testing procedure corresponding with cut-off $c(Q_n, \mathcal{S}_0 \mid q, \alpha)$ satisfies (1).

This representation $c(Q_n, \mathcal{S}_0 \mid q, \alpha)$ as the optimal cut-off in terms of the unknown distribution of T_n and the set of true null hypotheses inspires our approach proposed in this article. In the next two subsections we present this approach, and present the corresponding finite sample rational, respectively.

2.1 The Proposed Multiple Testing Procedure

Before presenting the finite and asymptotic rational of our procedure, we will outline the actual steps of the proposed technique. Recall that the observed data is n i.i.d. copies $X_1, ..., X_n$ of a random variable X, and $T_n = (T_n(1), \ldots, T_n(m))$ denotes the vector of test-statistics corresponding with m null hypotheses.

Our method for choosing c involves controlling the tail probability of a random variable $\tilde{r}_n(c)$ defined as

$$\tilde{r}_n(c) = \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n})}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n}) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_{0n})}.$$

This random variable represents a guessed proportion of false positives among rejections, defined by drawing a random set $\mathcal{S}_{0n}$ which represents a guess of the set of true null hypotheses $\mathcal{S}_0$ and a draw $\tilde{T}_n$ from a null distribution for the test-statistic vector. The distribution of $\mathcal{S}_{0n}$ and null distribution of $\tilde{T}_n$ are chosen so that $\tilde{r}_n(c)$ asymptotically dominates in distribution the true proportion of false positives, $\frac{V_n(c)}{V_n(c)+S_n(c)}$. By selecting a conservative finite sample distribution of $\mathcal{S}_{0n}$, it is expected to also dominate this true proportion of false positives in finite samples. We expand on this in the next subsection.

Firstly, we describe the null distribution of $\tilde{T}_n$. $\tilde{T}_n$ is computed by drawing a bootstrap sample $X_1^{\#}, \ldots, X_n^{\#}$ from the empirical distribution P_n the original sample $X_1, ..., X_n$, or from a model based estimate $\tilde{P}_n$ of P, and subsequently calculating the test statistics based on this bootstrap sample. This will be repeated B^* times and will result in an $m \times B^*$ matrix of test-statistic vectors, representing a draw from the test-statistic vector under the empirical distribution P_n (or the model based estimate $\tilde{P}_n$). Subsequently, we compute the row means $E[T_n^{\#}(j)]$ (conditional on P_n) of the matrix, and the matrix is shifted (centered) by the respective means so that the row means after this shift are equal to the null-value $\theta_0(j)$. This matrix represents a sample of B^* draws from a null distribution $Q_{0,n}$ (Pollard and van der Laan, 2003; Dudoit et al., 2004b). Each row of this matrix will specify a draw of $\tilde{T}_n = (\tilde{T}_n(j) : j = 1, \ldots, m)$. One can also scale the columns so that the row-variances equal a null value.

Secondly, we will define the distribution of our guessed set of null hypotheses $\mathcal{S}_{0n}$, and describe how this random set is drawn. This random set is defined by drawing a null or alternative status for each of the test statistics. The working model for defining the distribution of the guessed set $\tilde{\mathcal{S}}_{0n}$ will assume $T_n(j) \sim p_0 f_0 + (1 - p_0) f_1$, a mixture of a null density f_0 and alternative density f_1. Let $B(j)$ represent the underlying Bernoulli random variable, such that $f_0 \sim (T_n(j)|B(j) = 0)$, is the density of $T_n(j)$ if $H_0(j)$ is true, and $f_1 \sim (T_n(j)|B(j) = 1)$ is the density of $T_n(j)$ if $H_0(j)$ is false.

Under this working model, the posterior probability defined as the probability that $T_n(j)$ came from a true H_{0j}, given its observed value $T_n(j)$, can now be calculated:

$$P(B(j) = 0|T_n(j)) = p_0 \frac{f_0(T_n(j))}{f(T_n(j))}$$

We will use this posterior probability as the Bernoulli probability on H_{0j} being true, given the test statistic, where we have to specify or estimate p_0, f_0 and f. Since f_0 plays the roll of the density of test-statistics under the null hypothesis, in some situations f_0 is simply known: e.g., $f_0 \sim N(0,1)$. However, in cases where the marginal distribution of $T_n(j)$ is not known if H_{0j} is true, one can use a kernel density (**density()** in R with a given kernel and bandwidth) on the mean centered elements in the matrix representing B draws of $\tilde{T}_n$. The elements from this matrix are pooled into a vector of length $m * B^*$ in the kernel density function. In order to estimate the density f, we can again apply a kernel smoother on the bootstrapped test statistics, before they are mean centered. Again, the elements of the matrix are pooled into a vector of length $m * B^*$ in the kernel density function.

Finally, p_0 represents the proportion of null hypotheses $\mid \mathcal{S}_0 \mid /m$ and typically the user might use a conservative p_0^* for this true proportion of null hypotheses. We use the most conservative prior, $p_0^* = 1$, throughout this paper. Now, given T_n, we can define the random set

$$\mathcal{S}_{0n} = \{j : C(j) = 1\}, C(j) \sim Bernoulli\left(\min\left(1, p_0^* \frac{f_0(T_n(j))}{f(T_n(j))}\right)\right).$$

Given the data $X_1, \ldots, X_n$ (i.e., P_n), $\mathcal{S}_{0n}$ and $\tilde{T}_n$ are drawn independently.

We will now draw $(\mathcal{S}_{0n}, (\tilde{T}_n(j))$ B^* times, and each time calculate the corresponding realization of $\tilde{r}_n(c)$, where T_n is fixed at the true original test statistics (at each realization of $\mathcal{S}_{0n}$, in order to calculate $\tilde{r}_n(c)$, we need $\sum_{j \notin \tilde{\mathcal{S}}_{0n}} I(T_n(j) > c)$). This provides us with a sample of B^* realizations of $(\tilde{r}_n^b(c) : c \geq 0)$, $b = 1, \ldots, B^*$, conditional on the data P_n (and thus, conditional on T_n as well).

The cut-off c is set so that the tail probability, at a user supplied level q, of the random variable, $\tilde{r}_n(c)$, equals α. To do so, we will then choose c such that average over B^* draws of both $\tilde{T}_n(j)$ and $\mathcal{S}_{0n}(j)$ equals α.

Specifically, we set

$$c_n = \inf\left\{c : \frac{1}{B^*} \sum_{b=1}^{B^*} I(\tilde{r}_n^b(c) > q) \leq \alpha\right\}.$$

This finishes the description of our procedure. Finally, at a fixed data generating distribution, typically the distribution of $\mathcal{S}_{0n}$ converges to the constant set $\mathcal{S}_0$ for n converging to infinity. Given $p_0^* = 1$, the estimated posterior probability as $p_n(j) \equiv min\left(\frac{f_0(T_n(j))}{f_n(T_n(j))}, 1\right)$. Two conditions guarantee this convergence.

1. Given $T_n(j)$ is distributed as f_0 or is dominated by f_0, if $j \in \mathcal{S}_0$ implies that $f_{1n}(T_n(j))/f_0(T_n(j)) \to_P 0$ as $n \to \infty$ (which one typically expects, since the alternative density f_{1n} will be shifted towards $+\infty$), then

$$p_n(j) = min\left(\frac{f_0(T_n(j))}{p_0 f_0(T_n(j)) + (1-p_0) f_{1n}(T_n(j))}, 1\right) \to_P min\left(\frac{1}{p_0}, 1\right) = 1$$

   as $n \to \infty$.

2. If $j \notin \mathcal{S}_0$ implies that $f_0(T_n(j))/f_1(T_n(j)) \to_P 0$ as $n \to \infty$, then

$$p_n(j) = min\left(\frac{f_0(T_n(j))}{p_0 f_0(T_n(j)) + (1-p_0) f_{1n}(T_n(j))}, 1\right) \to_P 0.$$

   as $n \to \infty$.

2.2 Finite sample rational of our proposal.

In this section we provide a semi-formal finite sample rational of our proposal, and in the next section we will prove the asymptotic validity of our method. Firstly, we will point out that if one is able to provide a conservative guess for the set of true null hypotheses (that is, this guessed set contains the set of true null hypotheses), then it follows that one can simply choose the cut-off so that the corresponding guessed actual proportion of false positives equals q. However, this method will be extremely sensitive to the set of guessed null hypotheses not containing any true positives. To reduce this sensitivity, our method replaces the test-statistics corresponding with the guessed null hypotheses by a random draw of test-statistics from a null distribution with the correct covariance structure (which is the same as the true covariance structure), and replaces the single guess of the set of true null hypotheses by a random guess from a distribution which is asymptotically degenerate at the set of true null hypotheses. This yields a random guessed proportion of false positives, and we in turn choose the cut-off so that it's survivor function at q, conditional on the data, equals α.

Given a vector of test-statistics T_n, the guessed proportion of false positives corresponding with a guessed set $\tilde{s}_0 \subset \{1, \ldots, m\}$ of true null hypotheses and cut-off c is given by

$$\frac{\sum_j I(T_n(j) > c, j \in \tilde{s}_0)}{\sum_j I(T_n(j) > c, j \in \tilde{s}_0) + \sum_j I(T_n(j) > c, j \notin \tilde{s}_0)}.$$

Since the function $x \rightarrow \frac{x}{x+c}$ is monotone increasing (and convex), it follows that, if our set of guessed true null hypotheses contains the set of true null hypotheses, i.e., $\tilde{s}_0 \supset \mathcal{S}_0$, then

$$\begin{aligned} &\frac{\sum_j I(T_n(j) > c, j \in \tilde{s}_0)}{\sum_j I(T_n(j) > c, j \in \tilde{s}_0) + \sum_j I(T_n(j) > c, j \notin \tilde{s}_0)} \\ &\geq \frac{\sum_j I(T_n(j) > c, j \in \mathcal{S}_0)}{\sum_j I(T_n(j) > c, j \in \mathcal{S}_0) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_0)}. \end{aligned}$$

That is, if $\tilde{s}_0 \supset \mathcal{S}_0$, and we simply choose the cut-off such that the proportion of test-statistics $T_n(j)$ with $j \in \tilde{s}_0$ among the rejections equals q, then the proportion of actual false positives among the rejections is smaller or equal than q. We do not recommend this approach since it will be extremely sensitive to $\tilde{s}_0$ containing all of the true null hypotheses $\mathcal{S}_0$, due to the fact that if $j \in \tilde{s}_0$ while $j \notin \mathcal{S}_0$, the cut-off chosen will be too large. Thus, our proposal involves 1) replacing the observed test-statistics $(T_n(j) : j \in \tilde{s}_0)$ by a random draw $(\tilde{T}_n(j) : j \in \tilde{s}_0)$ from our null distribution, which has asymptotically the same distribution up until a simple shift, 2) replacing the fixed $\tilde{s}_0$ by a random draw $\mathcal{S}_{0n}$ independent of $\tilde{T}_n$, given the data P_n, from a conservatively chosen distribution which is asymptotically degenerate at $\mathcal{S}_0$, and 3) controlling the tail probability at q over the distribution of $(\mathcal{S}_{0n}, \tilde{T}_n)$, conditional on P_n.

As discussed above, one can create a random vector $\tilde{T}_n$, representing a draw from the null-value shifted bootstrap distribution of T_n, such that the distribution of $\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)$, given the original sample P_n, asymptotically dominates the distribution of $\sum_j I(T_n(j) > c, j \in \mathcal{S}_0)$ (Dudoit et al., 2004b). Such a result can be derived by establishing the limit distribution of the bootstrap distribution of $\tilde{T}_n$, given P_n, which typically simply corresponds with proving asymptotic validity of the bootstrap. Though such results establish asymptotic domination, in practice these distributions typically also provide finite sample domination, due to the fact that $\theta_0(j)$ provides an upper-bound for the mean of the test-statistics under a true null hypotheses H_{0j}.

Note that such a limit distribution implies that $\tilde{T}_n$ is asymptotically independent of P_n, and thus, $\tilde{T}_n$ is asymptotically independent of T_n. As a consequence, the conditional distribution of $\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)$, given $\sum_j I(T_n(j) > c, j \notin \mathcal{S}_0)$, asymptotically dominates the marginal distribution of $\sum_j I(T_n(j) > c, j \in \mathcal{S}_0)$, even at local alternatives.

Given this substitution of $(\tilde{T}_n(j) : j \in \tilde{s}_0)$ for $(T_n(j) : j \in \tilde{s}_0)$, we obtain the random variable $\frac{\sum_j I(\tilde{T}_n(j)>c, j\in\tilde{s}_0)}{\sum_j I(\tilde{T}_n(j)>c, j\in\tilde{s}_0)+\sum_j I(T_n(j)>c, j\notin\tilde{s}_0)}$. If $\tilde{s}_0 \supset \mathcal{S}_0$, then

$$\begin{aligned}
&\frac{\sum_j I(\tilde{T}_n(j) > c, j \in \tilde{s}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \tilde{s}_0) + \sum_j I(T_n(j) > c, j \notin \tilde{s}_0)} \\
&\geq \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0) + \sum_j I(T_n(j) > c, j \notin \tilde{s}_0)} \\
&\geq \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_0)}
\end{aligned}$$

Recall that our goal is to dominate the latter random variable with $\tilde{T}_n(j)$ replaced by $T_n(j)$. Now, we can use the fact that if a random variable X dominates a random variable Y stochastically, $(X \geq_P Y)$, in the sense that $P(X \leq x) \leq P(Y \leq x)$ for all x, then for a fixed constant a $\frac{X}{X+a}$ dominates the random variable $\frac{Y}{Y+a}$, where a is $S_n(c) = \sum_j I(T_n(j) > c, j \notin \mathcal{S}_0)$, X is $\tilde{V}_n(c) = \sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)$, and Y is the non-conditional number of false positives $V_n^*(c) = \sum_j I(T_n(j) > c, j \in \mathcal{S}_0)$. Here $V_n^*(c)$ is a random variable with the same marginal distribution as $V_n(c)$, but $V_n^*(c)$ is independent of $S_n(c)$.

To summarize: If $\tilde{s}_0 \supset \mathcal{S}_0$, $\tilde{V}_n(c)$ dominates $V_n(c)$ for all c in distribution (marginally), and $\tilde{T}_n$ is independent of T_n, then

$$
\begin{aligned}
&\frac{\sum_j I(\tilde{T}_n(j) > c, j \in \tilde{s}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \tilde{s}_0) + \sum_j I(T_n(j) > c, j \notin \tilde{s}_0)} \\
&\geq \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0) + \sum_j I(T_n(j) > c, j \notin \tilde{s}_0)} \\
&\geq \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_0)} \\
&\geq_P \frac{V_n^*(c)}{V_n * (c) + S_n(c)}, \text{ conditional on } S_n(c)
\end{aligned}
$$

Again, recall that we are aiming to stochastically dominate the random variable $\frac{V_n(c)}{V_n(c)+S_n(c)}$. Thus, if $V_n(c)$ is independent of $S_n(c)$ so that $(V_n^*(c), S_n(c))$ equals in distribution $(V_n(c), S_n(c))$, then we would be dominating the wished $\frac{V_n(c)}{V_n(c)+S_n(c)}$. Thus, in that case, choosing c such that the conditional tail probability of $\frac{\sum_j I(\tilde{T}_n(j)>c, j\in\tilde{s}_0)}{\sum_j I(\tilde{T}_n(j)>c, j\in\tilde{s}_0)+\sum_j I(T_n(j)>c, j\notin\tilde{s}_0)}$, given P_n (i.e., T_n), at q equals α would yield a cut-off larger than or equal to the optimal cut-off $c(Q_n, \mathcal{S}_0 \mid q, \alpha)$, and thereby a multiple testing procedure controlling TPPFP (q) at level α.

The assumption that $V_n(c)$ is independent of $S_n(c)$ is sufficient, but not necessary to obtain the wished stochastic domination. In addition, at a fixed data generating distribution, $S_n(c)$ converges to the constant $\mid \mathcal{S}_0^c \mid$ so that this independence condition is asymptotically empty. It is interesting to note that this independence assumption was also used in the proof of Lehmann and Romano (2003) to establish the wished control of $TPPFP(q)$ for their procedure based on marginal p-values.

Though this multiple testing procedure has a finite sample rational under the assumption that $V_n(c)$ is independent of $S_n(c)$ (for all c), which is asymptotically an empty condition at a fixed data generating distribution, it still relies on a guessed set $\tilde{s}_0$ containing the set of true null hypotheses $\mathcal{S}_0$. Therefore, in our proposed method we simply select c such that the tail probability of

$$\frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n})}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n}) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_{0n})}$$

at q equals α, where $\mathcal{S}_{0n}$ is a random set drawn (independently from $\tilde{T}_n$) from a probability distribution estimated from the data (i.e., P_n) and which is asymptotically degenerate at the true $\mathcal{S}_0$. If $\mathcal{S}_{0n}$ follows a conservatively chosen distribution in the sense that $\mathcal{S}_{0n}$ is typically larger (e.g., its average contains $\mathcal{S}_0$) than $\mathcal{S}_0$ (but still asymptotically consistent for $\mathcal{S}_0$), one would expect that the finite sample rational for a fixed $\tilde{s}_0 \supset \mathcal{S}_0$ above is still approximately true, while our approach will now be more robust (i.e., less variable) in finite samples than an approach based on a single guess $\tilde{s}_0$.

2.3 Formal asymptotic validity.

Though the above rational provides the finite sample heuristic behind our method, the following theorem formally establishes the asymptotic validity of our method at a fixed data generating distribution, under general conditions.

Theorem 1 *Define*

$$\tilde{r}_n(c) \equiv \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n})}{\sum_j I(\tilde{T}_n(j) > c, j \in \tilde{\mathcal{S}}_{0n}) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_{0n})}.$$

Let $\tilde{T}_n$ be independent of $\mathcal{S}_{0n}$, given P_n, and let $\tilde{Q}_n$, G_{0n} denote the conditional distributions of $\tilde{T}_n$ and $\mathcal{S}_{0n}$, given P_n, respectively. Let

$$c_n = c(G_{0n}, \tilde{Q}_n, P_n \mid q, \alpha) \equiv \inf\{c : \bar{F}_{\tilde{r}_n(c)|P_n}(q) \leq \alpha\},$$

where the notation $c(G_{0n}, \tilde{Q}_n, P_n \mid q, \alpha)$ expresses the dependence of this cut-off on the distribution G_{0n} of $\mathcal{S}_{0n}$, given P_n, the distribution $\tilde{Q}_n$ of $\tilde{T}_n$, given P_n, the actual sample identified by P_n (i.e., the values of the test-statistics T_n), and the user supplied (α, q). In addition, $\bar{F}_{X_1|X_2}(q) \equiv P(X_1 > q \mid X_2)$ denotes the conditional survivor function.

Suppose that

1. *G_{0n} converges to the degenerate distribution which puts probability 1 on the constant set $\mathcal{S}_0$ for n converging to infinity.*

2. *Let*
$$\tilde{c}_n \equiv \inf\{c : \bar{F}_{\tilde{V}_n(c)/(\tilde{V}_n(c)+|S_0^c|)|P_n}(q) \leq \alpha\},$$
where $\tilde{V}_n(c) \equiv \sum_{j=1}^m I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)$. It is assumed that there exists a τ so that $\limsup_{n\to\infty} \tilde{c}_n \leq \tau$, and
$$\sum_{j=1}^m I(T_n(j) > \tau, j \notin \mathcal{S}_0) - \mid \mathcal{S}_0^c \mid \rightarrow 0$$
for n converging to infinity, for almost every $(P_n : n \geq 1)$.

3. *For almost every $(P_n : n \geq 1)$, for each $x \in \{1, \ldots, m\}$, we have*
$$\limsup_{n\to\infty} \sup_{c\in[0,\tau]} \bar{F}_{\tilde{V}_n(c)|P_n}(x) - \bar{F}_{V_n(c)}(x) \leq 0.$$

4. *Given* $(P_n : n \geq 1)$*, if* $\tilde{c}_n$ *is a sequence so that*

$$\limsup_{n\to\infty} \bar{F}_{\tilde{r}_n(\tilde{c}_n)|P_n}(q) \leq \alpha,$$

then $\limsup_{n\to\infty} c_n - \tilde{c}_n \leq 0$.

5. *If* $\tilde{c}_n$ *is a sequence so that for almost every* $(P_n : n \geq 1)$, $\limsup_{n\to\infty} c_n - \tilde{c}_n \leq 0$*, then*

$$\limsup_{n\to\infty} F_{V_n(\tilde{c}_n)/V_n(\tilde{c}_n)+S_n(\tilde{c}_n)}(q) - F_{V_n(c_n)/V_n(c_n)+S_n(c_n)}(q) \geq 0.$$

Then,

$$\limsup_{n\to\infty} \bar{F}_{V_n(c_n)/R_n(c_n)}(q) \leq \alpha, \tag{2}$$

where $V_n(c_n) = \sum_{j=1}^m I(T_n(j) > c_n, j \in \mathcal{S}_0)$*, and* $R_n(c_n) = \sum_{j=1}^m I(T_n(j) > c_n)$*.*

Discussion of conditions. Condition 1) states that our random guess of $\mathcal{S}_0$ should be asymptotically on target, and, as noted above, our actual finite sample distribution of this random guess will be chosen conservatively. Condition 2) naturally holds at a fixed data generating distribution since it states that the test-statistics corresponding with false null hypotheses asymptotically separate from the test-statistics corresponding with the true null hypotheses. Condition 3) states that the number of false rejections under our chosen null distribution asymptotically dominates the number of false rejections under the true distribution. The last two conditions 4) and 5) are very mild regularity conditions avoiding situations in which the tail-probability of the proportion of false positives is not affected by a change in the cut-off.

Proof. Firstly, by condition 1) and 2), it follows that, given almost every $(P_n : n \geq 1)$, $(\tilde{r}_n(c) : c \in [0,\tau])$ equals with probability tending to 1

$$\begin{aligned}(\tilde{r}^*_n(c) : c \in [0,\tau]) &\equiv \left(\frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_0)+ \mid \mathcal{S}_0^c \mid} : c \in [0,\tau] \right) \\ &= \left(\frac{\tilde{V}_n(c)}{\tilde{V}_n(c)+ \mid \mathcal{S}_0^c \mid} : c \in [0,\tau] \right).\end{aligned}$$

As a consequence, the difference between the cumulative survivor function of $\tilde{r}_n(c)$ at q, given P_n, and the cumulative survivor function of $\tilde{r}^*_n(c)$ at q, given P_n, converges to zero uniformly in $c \in [0,\tau]$. Note that, given $(P_n : n \geq 1)$, $\tilde{c}_n$ is a fixed sequence, and, by assumption, there exists a N so that for $n > N$, $\tilde{c}_n \in [0,\tau]$. As a consequence, it follows that the survivor function of $\tilde{r}^*_n(\tilde{c}_n)$, given P_n, at q, which equals a number smaller or equal than α, minus the conditional survivor function of $\tilde{r}_n(\tilde{c}_n)$, given P_n, at q converges to zero. Thus, given almost every $(P_n : n \geq 1)$, the limsup of the conditional survivor function of $\tilde{r}_n(\tilde{c}_n)$, given P_n, at q, converges to a number smaller or equal than α. By assumption 4), this implies, in particular, that, given almost every $(P_n : n \geq 1)$, $\limsup_{n\rightarrow\infty} c_n - \tilde{c}_n \leq 0$, which we will need later.

Now, we note that for all $c \in [0, \tau]$

$$P\left(\frac{\tilde{V}_n(c)}{\tilde{V}_n(c)+ \mid \mathcal{S}_0^c \mid} > q \mid P_n\right) = P\left(\tilde{V}_n(c) > \frac{q \mid \mathcal{S}_0^c \mid}{1-q} \mid P_n\right).$$

By condition 3), the latter conditional probability, given P_n, is asymptotically larger than the marginal probability

$$P\left(V_n(c) > \frac{q \mid \mathcal{S}_0^c \mid}{1-q}\right) = P\left(\frac{V_n(c)}{V_n(c)+ \mid \mathcal{S}_0^c \mid} > q\right),$$

uniformly in $c \in [0, \tau]$. However, by condition 1), the latter probability is asymptotically the same as $P\left(\frac{V_n(c)}{V_n(c)+S_n(c)} > q\right)$, uniformly in $c \in [0, \tau]$. This proves that, for almost every $(P_n : n \geq 1)$,

$$\limsup_{n\rightarrow\infty} \sup_{c\in[0,\tau]} \left\{P\left(\frac{V_n(c)}{V_n(c)+S_n(c)} > q\right) - P(\tilde{r}_n(c) > q \mid P_n)\right\} \leq 0.$$

Since $\tilde{c}_n \in [0, \tau]$ for n large enough, and $P(\tilde{r}_n(\tilde{c}_n) > q \mid P_n) \leq \alpha$ asymptotically, it follows now that, for almost every $(P_n : n \geq 1)$,

$$\limsup_{n\rightarrow\infty} P\left(\frac{V_n(\tilde{c}_n)}{V_n(\tilde{c}_n)+S_n(\tilde{c}_n)} > q\right) \leq \alpha. \tag{3}$$

Finally, since, for almost every $(P_n : n \geq 1)$, $\limsup_{n\rightarrow\infty} c_n - \tilde{c}_n \leq 0$ (shown above), condition 5) teaches us that (3) implies that we also have

$$\limsup_{n\rightarrow\infty} P\left(\frac{V_n(c_n)}{V_n(c_n)+S_n(c_n)} > q\right) \leq \alpha.$$

This completes the proof. □

3 Simulations

The simulation study compares the procedure outlined above with the augmentation procedure of FWER adjusted p-values presented in van der Laan et al.

(2004b). Recall that, given the data P_n, the implementation of our multiple testing procedure involves simulating

$$\tilde{r}_n(c) = \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n})}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n}) + \sum_j I(T_n(j) > c, j \notin \tilde{\mathcal{S}}_{0n})}$$

Recall also that we identify such a random set $\tilde{\mathcal{S}}_{0n}$ with a random vector $(C(1), ..., C(m))$ of Bernoulli indicators $C(j)$ drawn independently from a Bernoulli distribution with probability $1 - \min\left(1, \frac{f_{0n}(T_n(j))}{f_n(T_n(j))}\right)$, where f_{0n} and f_n are kernel density estimators described in Section 2.1. The reader will be referred back to Section 2.1 to show that this posterior probability is asymptotically degenerate at $\mathcal{S}_0$. We will now define several aspects of the simulation of the data.

3.1 Data

The data are n i.i.d. normally distributed vectors $X_i \sim N(\Psi(P), \Sigma(P))$, $i = 1, \ldots, n$, where $\psi = (\psi(j) : j = 1, \ldots, m) = \Psi(P) = E_P[X]$ and $\sigma = (\sigma(j, j') : j, j' = 1, \ldots, m) = \Sigma(P) = Cov_P[X]$ denote, the m-dimensional mean vector and $m \times m$ covariance matrix.

3.2 Null hypotheses

The null hypotheses of interest concern the m components of the mean vector ψ. That is, we are interested in two-sided tests of the m null hypotheses $H_0(j) = \mathrm{I}\big(\psi(j) = \psi_0(j)\big)$ vs. the alternative hypotheses $H_1(j) = \mathrm{I}\big(\psi(j) \neq \psi_0(j)\big)$, $j = 1, \ldots, m$. We will set the null values equal to zero, i.e., $\psi_0(j) \equiv 0$.

3.3 Test statistics

In the known variance case, one can test the null hypotheses using simple t-statistics. We will rewrite the test-statistics and define the respective shift below:

$$T_n(j) \equiv \sqrt{n}\frac{\psi_n(j) - \psi_0(j)}{\sigma(j)},$$

where $\psi_n(j) = \sum_i \frac{X_i(j)}{n}$ denote the empirical means for the m components of X. For our case, the test statistics $T_n(j)$ can be rewritten in terms of random variables (Z_n) and shift parameters (d_n):

$T_n(j) = \sqrt{n}\frac{\psi_n(j)-\psi(j)}{\sigma(j)} + \sqrt{n}\frac{\psi(j)-\psi_0(j)}{\sigma(j)} = Z_n(j) + d_n(j)$,

where $Z_n \sim N(0, \Sigma^*(P))$ and $\sigma^* = \Sigma^*(P) = Cor[X]$.

Therefore the test statistics T_n have an m–variate Gaussian distribution with mean vector the shift vector d_n and covariance matrix σ^*: $T_n \sim N(d_n, \sigma^*)$. Note that $d_n(j) = 0$ if the null hypothesis $H_0(j)$ is true. Various values of the shift $d_n(j)$ corresponds to different combinations of sample size n, mean $\psi(j)$, and variance $\sigma^2(j)$.

3.4 Simulation parameters

In our simulations we can simulate the test statistics T_n directly from the m–variate Gaussian distribution $T_n \sim N(d_n, \sigma^*)$, where the parameter of interest is now the shift vector d_n, with j^{th} component equal to zero under the corresponding null hypothesis.

The following model parameters where used in the simulation.

- *Number of hypotheses, m*:

  The following two values were considered for the total number of hypotheses, $m = 24$ and $m = 400$.

- *Proportion of true null hypotheses, h_0/m*:

  50% of true null hypotheses ($h_0/m = 0.5$) or 75% of true null hypotheses ($h_0/m = 0.75$).

- *Shift parameters, $d_n(j)$*:

  For the true null hypotheses, i.e., for $j \in S_0$, $d_n(j) = 0$.

  For the false null hypotheses, i.e., $j \notin S_0$, the following (common) shift values were considered: $d_n(j) = 2, 3, 4, [2, 10]$.

  **Note in the case $d_j = [2, 10]$ with m=400, 150 T_n had a shift of 2 and 50 T_n had a shift of 10, thus simulating an actual situation in practice where 50 of the hypotheses are bound to be automatically rejected.

- *Correlation matrix, σ^**:

  The following type of correlation structure was considered:

  Local correlation, where the only non-zero elements of σ^* are the diagonal and first off-diagonal elements, i.e., $\sigma^*(j, j) = 1$, for $j = 1, \ldots, m$, $\sigma^*(j, j -$

$1) = \sigma^*(j-1, j) = 0.5$, for $j = 2, \ldots, m$, and $\sigma^*(j, j') = 0$, for j, $j' = 1, \ldots, m$ and $j' \neq j-1$, j, $j+1$.

- The null distribution, usually obtained from the bootstrap, is generated by creating a $10,000 \times m$ matrix of test statistics null distribution Q_0, $Z \sim N(0, \sigma^*)$. We note that Z represents the limit distribution of the bootstrap null distribution which we actually use in practice.
- The possible cut-off values c are between 2 and 4 by steps of size 0.05.
- The tail probability proportion q and α level are both set to 0.05.
- The number of draws of the Bernoulli-vector $(C(1), \ldots, C(m))$ identifying S_{0n} was equal to 50. Note that in our actual description of the method we are supposed to draw $(\tilde{T}_n, S_{0n})$ repeatedly, while in this simulation we draw more $\tilde{T}_n$ $(10,000)$ than we draw S_{0n}'s (50). However, this was only done for computational reasons. One might expect a minor improvement of our method in the case that both random variables are drawn 10,000 times, as recommended in practice.

Multiple Testing Procedure: TPPFP Augmentation

We have applied the single step maxT Multiple Testing Procedure outlined in Pollard and van der Laan (2003). This procedure is a single-step approach, with common cut-off, which uses a null distribution based on the joint distribution of the test statistics. This null distribution is used to define the rejection regions as well as the adjusted p-values. The null distribution is the $\tilde{T}_n$ matrix (Pollard and van der Laan, 2003). This procedure is based on obtaining a vector of B^* maximum values from the columns of the $\tilde{T}_n$ matrix. The estimated common cut-off value c_o is the $(1-\alpha)$ quantile of the B^*-vector of maximum values, obtained from the estimated bootstrapped distribution. This now defines a Multiple Testing Procedure, which is based on the test statistics, null distribution, and α. We then apply an augmentation defined in van der Laan et al. (2004b) to the FWER adjusted p-values. This is done at a user defined $q = \alpha = 0.05$. As mentioned previously, we will define the initial set of rejections of size r_0 corresponding with a multiple testing procedure controlling FWER at level α. The TPPFP augmentation procedure simply adds the next $\lceil \frac{q}{1-q} r_0 \rceil$ most significant tests to the rejection set to control TPPFP(q) at level α.

Lehmann and Romano TPPFP Procedures:

We also applied the Lehmann and Romano Restricted method to control the tail probability of the proportion of false positives (Lehmann and Romano, 2003). This is a method based on marginal p-values, and the adjusted p-values for such procedures are simple functions of the unadjusted p-values $P_{0n}(j)$ corresponding to each null hypothesis $H_0(j)$: we recall that an adjusted-p-value, given a test-statistic value, is the actual nominal level α one needs to chose to just put the test-statistic in the rejection region. We will denote the adjusted p-values for the MTP by $\widetilde{P}_{0n}(j)$ and the ordered p-values (from smallest to largest) are defined as $O_n(j)$, so that $P_{0n}(O_n(1)) \leq \ldots \leq P_{0n}(O_n(m))$. The Lehmann and Romano Restricted step-down procedure for controlling TPPFP at a user specified level q, is defined as in (Lehmann and Romano, 2003; Dudoit et al., 2004a) in terms of adjusted p-values as follows:

$$\widetilde{P}_{0n}(O_n(j)) = \max_{h=1,\ldots,j} \left\{\min\left(\frac{(m+\lfloor qh\rfloor+1-h)}{(\lfloor qh\rfloor+1)} P_{0n}(O_n(h)), 1\right)\right\}$$

The Lehmann and Romano Restricted procedure is shown to control the TPPFP under either one of two assumptions on the dependence structure of the unadjusted p-values (Theorems 3.1 and 3.2 in Lehmann and Romano (2003)). Lehmann and Romano (2003) have also proposed a General step-down method to control TPPFP, which is outlined in both Lehmann and Romano (2003) and Dudoit et al. (2004a). This method is a very conservative in practice, and controls the TPPFP under arbitrary dependence structures (Theorem 3.3). We will not present results for this Lehmann and Romano General method in this article.

We will report simulation results for the newly proposed procedure, the TPPFP augmentation method described above, and the Restricted Lehmann and Romano procedure. We note that the Lehmann and Romano method is not directly comparable to the augmentation method based on the single-step maxT method for controlling FWE, since the Lehmann and Romano method is step-down. To make them more comparable, we would have to include the augmentation method based on the step-down method for controlling FWE, as in our simulation studies presented in Dudoit et al. (2004a).

3.5 Type I error rate and power comparisons

Finally, for each data generating distribution, we carry out the multiple testing procedures (newly proposed procedure, augmentation of FWE adjusted p-values procedure, and Lehmann and Romano Restricted procedure) S_n 1000 times. We do this by generating $W = 1000$ $m-$vectors of test statistics $T_n^w \sim N(d_n, \sigma^*)$, $w = 1, \ldots, W$.

For a given nominal level α, we compute the numbers of rejected hypotheses $R_n^w(\alpha) =\mid S_n^w \mid$, Type I errors $V_n^w(\alpha) =\mid S_n^w \cap \mathcal{S}_0 \mid$, and Type II errors $U_n^w(\alpha) =\mid$

$S_n^w \cap \mathcal{S}_0^c \mid$.

Based on this Monte-Carlo sample of $(V_n(\alpha), R_n(\alpha), U_n(\alpha))$ for our multiple testing procedure $S_n(\alpha)$, we can obtain an empirical estimate of the Type-I error and Average Power:

$$\begin{aligned} TPPFP(q;\alpha) &= \frac{1}{W}\sum_{w=1}^{W} \mathrm{I}(V_n^w(\alpha)/R_n^w(\alpha) > q) \\ AvgPwr(\alpha) &= 1 - \frac{1}{h_1}\frac{1}{W}\sum_{w=1}^{W} U_n^w(\alpha). \end{aligned}$$

Similarly, we calculate this Type-I error and Average Power for the augmentation procedure and the Lehmann and Romano procedure.

Table 1: $m = 24$, shift for alternatives $= 2$, and $\frac{h_0}{m} = 0.50$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.033	0.016	0.027
$\alpha = 0.05$ Power	0.1836	0.1374	0.1341
$\alpha = 0.1$ Type I error	0.079	0.035	0.054
$\alpha = 0.1$ Power	0.2816	0.2029	0.192

Table 2: $m = 24$, shift for alternatives $= 3$, and $\frac{h_0}{m} = 0.50$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.037	0.023	0.03
$\alpha = 0.05$ Power	0.5875	0.4813	0.4772
$\alpha = 0.1$ Type I error	0.093	0.053	0.06
$\alpha = 0.1$ Power	0.6764	0.583	0.5775

3.6 Simulation Results

The various simulations indicate that the proposed tail probability of the proportion of false positives (TPPFP) method is more powerful and less conservative as compared to the augmentation method applied to FWER adjusted p-values at nominal α levels of 0.05 and 0.10. The simulations vary several parameters in order to make these comparisons. As mentioned earlier, we were particularly interested in the performance of our new method in situations where the number of tests m increases, therefore in this case $m = 400$, since the augmentation method is known

Table 3: $m = 400$, shift for alternatives $= 2$, and $\frac{h_0}{m} = 0.50$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.037	0.007	0.006
$\alpha = 0.05$ Power	0.1484	0.0555	0.05276
$\alpha = 0.1$ Type I error	0.082	0.016	0.018
$\alpha = 0.1$ Power	0.2135	0.0910	0.0819

Table 4: $m = 400$, shift for alternatives = 3, and $\frac{h_0}{m} = 0.50$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.041	0.009	0.0062
$\alpha = 0.05$ Power	0.549	0.2899	0.3426
$\alpha = 0.1$ Type I error	0.088	0.025	0.017
$\alpha = 0.1$ Power	0.6425	0.3826	0.445

Table 5: $m = 400$, shift for alternatives = 4, and $\frac{h_0}{m} = 0.50$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.037	0.017	0.005
$\alpha = 0.05$ Power	0.89445	0.6869	0.7739
$\alpha = 0.1$ Type I error	0.09	0.037	0.016
$\alpha = 0.1$ Power	0.93075	0.7715	0.8369

Table 6: $m = 400$, shift for alternatives = [2,10], and $\frac{h_0}{m} = 0.50$

α	TPPFP	Augmentation
$\alpha = 0.05$ Type I error	0.036	0.012
$\alpha = 0.05$ Power	0.4009	0.3023
$\alpha = 0.1$ Type I error	0.08	0.021
$\alpha = 0.1$ Power	0.4535	0.3292

Table 7: $m = 400$, shift for alternatives = 2, and $\frac{h_0}{m} = 0.75$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.045	0.011	0.01
$\alpha = 0.05$ Power	0.09624	0.05295	0.0406
$\alpha = 0.1$ Type I error	0.10	0.032	0.024
$\alpha = 0.1$ Power	0.1481	0.0874	0.06469

Table 8: $m = 400$, shift for alternatives = 3, and $\frac{h_0}{m} = 0.75$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.044	0.010	0.01
$\alpha = 0.05$ Power	0.42755	0.2838	0.2679
$\alpha = 0.1$ Type I error	0.094	0.035	0.029
$\alpha = 0.1$ Power	0.5244	0.3771	0.3436

Table 9: $m = 400$, shift for alternatives = 4, and $\frac{h_0}{m} = 0.75$

α	TPPFP	Augmentation	LR Restricted
$\alpha = 0.05$ Type I error	0.043	0.020	0.011
$\alpha = 0.05$ Power	0.8259	0.6823	0.6955
$\alpha = 0.1$ Type I error	0.092	0.049	0.023
$\alpha = 0.1$ Power	0.8822	0.7677	0.7642

to be too conservative in these circumstances. Clearly, as we observed previously, the augmentation method and LR-method are much too conservative in this case, while our new method has an actual TPPFP close to the wished level (e.g., for nominal level $\alpha = 0.1$, we have 0.08 versus 0.018). Thus, we indeed see a greater gain in both the respective power and Type I error rate (closer to the nominal level) as the number of tests increases. In many cases the Type I error rate of the TPPFP method is almost equal to the nominal Type I error rate, which is ideal for a multiple testing procedure.

4 Data Analysis

4.1 Introduction

We applied the proposed TPPFP method to an actual dataset in order to assess the performance by comparing the number of rejections at both $\alpha = 0.05$ and $\alpha = 0.10$ to those produced from the Augmentation method. Before defining the actual analyses, we will briefly describe the background and structure of the data.

4.2 HIV-1 sequence variation and replication capacity

Studying sequence variation for the Human Immunodeficiency Virus Type 1 (HIV-1) genome could potentially give important insight into genotype-phenotype associations for the Acquired Immune Deficiency Syndrome (AIDS).

In this context, the phenotype is the replication capacity (RC) of HIV-1, as it reflects the severity of the disease. A measure of replication capacity may be obtained by monitoring viral replication in an ideal environment, with many cellular targets, no exogenous or endogenous inhibitors, and no immune system responses against the virus (Barbour et al., 2002; Segal et al., 2004).

The genotype of interest correspond to codons in the protease and reverse transcriptase regions of the viral strand. The protease (PR) enzyme affects the reproductive cycle of the virus by breaking protein peptide bonds during viral replication. The reverse transcriptase (RT) enzyme synthesizes double-stranded DNA from the virus' single-stranded RNA genome, thereby facilitating integration into the host's chromosome. Since the PR and RT regions are essential to viral replication, many antiretrovirals (protease inhibitors and reverse transcriptase inhibitors) have been developed to target these specific genomic locations. Studying PR and RT genotypic variation involves sequencing the corresponding HIV-1 genome regions and determining the amino acids encoded by each codon (i.e., each nucleotide triplet).

4.3 Description of Segal et al. (2004) HIV-1 dataset

The HIV-1 sequence dataset consists of $n = 317$ records, linking viral replication capacity (RC) with protease (PR) and reverse transcriptase (RT) sequence data, from individuals participating in studies at the San Francisco General Hospital and Gladstone Institute of Virology (Segal et al., 2004). Protease codon positions 4 to 99 (i.e., $pr4 - pr99$) and reverse transcriptase codon positions 38 to 223 (i.e., $rt38 - rt223$) of the viral strand are studied in this analysis (Birkner et al., 2005).

The outcome/phenotype of interest is the natural logarithm of a continuous measure of replication capacity, ranging from 0.261 to 151. The M covariates correspond to the $M = 282$ codon positions in the PR and RT regions, with the number of possible codons ranging from one to ten at any given location. A majority of patients typically exhibit one codon at each position. Codons are therefore recoded as binary covariates, with value of **zero** (or "wild-type") corresponding to the most common codon among the $n = 317$ patients and value of **one** (or "mutation") for all other codons. Previous biological research was used to confirm mutations and hence provide accurate PR and RT codon genotypes for each patient (`hivdb.stanford.edu/cgi-bin/RTMut.cgi`) (Wu et al., 2003; Gonzales et al., 2003). The data for each of the $n = 317$ patients therefore consist of a replication capacity outcome/phenotype Y and an M–dimensional covariate vector $X = (X(j) : j = 1, \ldots, m)$ of binary codon genotypes in the PR and RT HIV-1 regions.

4.4 Parameter of Interest

In order to perform multiple testing, one must define the parameter of interest. In this specific case the parameter of interest is the difference $\psi(j)$ in mean replication capacity of viruses with mutant and wild-type codons, that is, $\psi(j) \equiv E[Y|X(j) = 1] - E[Y|X(j) = 0]$, $j = 1, \ldots, m$. To identify codons that are associated with viral replication capacity, one can perform two-sided tests of the null hypotheses $H_0(j) = \mathrm{I}(\psi(j) = 0)$ of no mean difference vs. the alternative hypotheses $H_1(j) = \mathrm{I}(\psi(j) \neq 0)$, using pooled-variance two-sample t-statistics $T_n(j)$. The null hypotheses are rejected, i.e., the corresponding codon positions are declared significantly associated with replication capacity, for large absolute values of the test statistics $T_n(j)$. It is important to note that only 25 of the 282 codon positions have unadjusted p-values less than an $\alpha = 0.05$ and 36 of the 282 codon positions have unadjusted p-values less than an $\alpha = 0.1$

We wish to test for each of the $M = 282$ codon positions whether viral replication capacity Y is associated with the corresponding binary codon genotype, $X(j) \in \{0, 1\}$, $j = 1, \ldots, m$. For the jth codon (i.e., jth hypothesis), the parameter of interest is the difference $\psi(j)$ in mean replication capacity of viruses with mutant and wild-type codons.

We consider two-sided tests of the null hypotheses $H_0(j) = \mathrm{I}(\psi(j) = 0)$ of no mean difference in RC vs. the alternative hypotheses $H_1(j) = \mathrm{I}(\psi(j) \neq 0)$ of different mean RC, based on pooled-variance two-sample t-statistics,

$$\begin{aligned} T_n(j) &\equiv \frac{\bar{Y}_1(j) - \bar{Y}_0(j) - 0}{s_p(j)\sqrt{\frac{1}{n_0(j)} + \frac{1}{n_1(j)}}}, \qquad (4) \\ s_p^2(j) &\equiv \frac{(n_0(j) - 1)s_0^2(j) + (n_1(j) - 1)s_1^2(j)}{n_0(j) + n_1(j) - 2}, \end{aligned}$$

where $n_k(j)$, $\bar{Y}_k(j)$, and $s_k^2(j)$ denote, respectively, the sample sizes, sample means, and sample variances for the RC of patients with codon genotype $X(j) = k \in \{0, 1\}$ at position j. The pooled variance estimator is denoted by $s_p^2(j)$. The null hypotheses are rejected, i.e., the corresponding codons are declared significantly associated with RC, for large absolute values of the test statistics $T_n(j)$. Note that the above two-sample t-statistics correspond to t-statistics for the univariate linear regression of the outcome Y on the binary covariates $X(j)$.

4.5 Methodology

4.5.1 Multiple Testing Procedures

We have applied the multiple testing procedure outlined in Pollard and van der Laan (2003). This procedure is a single-step maxT approach which uses a null distribution based on the joint distribution of the test statistics. This null distribution is used to define the rejection regions as well as the adjusted p-values. The null distribution is the $\tilde{T}_n$ matrix. We then apply the maxT single-step common cutoff procedure to obtain the FWER controlling adjusted p-values (Pollard and van der Laan, 2003). We then apply an augmentation defined in van der Laan et al. (2004b) to the FWER adjusted p-values. This is done at a user defined $q = \alpha = 0.05$.

The FWER method produces 282 adjusted FWER controlling adjusted p-values. Each of these adjusted p-values corresponds to a codon and represents the significance of the association between the codon and replication capacity. The augmentation is applied which results in TPPFP controlling adjusted p-values. We will tabulate the number of codons with adjusted p-values less than an $\alpha = 0.05$ and an $\alpha = 0.1$.

4.5.2 Multiple Testing Procedure: TPPFP

We have applied the presented method to the HIV-1 dataset in order to determine the number of rejected codons at both an $\alpha = 0.05$ and an $\alpha = 0.1$. This procedure was applied as outlined previously in this article. We had to choose a Bernoulli probability from the ratio of the null density f_0 to the empirical density f. We will assume that $f_0 \sim N(0, 1)$. In order to obtain the empirical density we applied a kernel density function (**density()** in R), to 10,000 m bootstrapped test statistics from the dataset. These Bernoulli's were repeated 50 times. The bootstrapped null distribution to which the method was applied was a $10,000 \times m$ matrix and was identical to the null distribution used for the construction of the FWER adjusted p-values in the previous method. We also tried estimating the density f of the bootstrapped test statistics with a normal distribution with the mean and variance equal to the mean and variance of the bootstrapped distribution. The results from this method were equivalent to the results found from using the kernel density method (presented in Section 5.3).

Table 10: HIV-1 Data: Number of Rejected Codons at $\alpha = 0.05, 0.1$

α	Rejections TPPFP	Rejections Augmentation
$\alpha = 0.05$	11	5
$\alpha = 0.1$	13	8

4.6 Results

The results from two methods are presented below. The new method rejects more hypotheses at both an $\alpha = 0.05$ and an $\alpha = 0.1$ as compared to the augmentation. We do observe a greater gain of the new method at the $\alpha = 0.05$ level.

Therefore this method proves to be less conservative as compared to the TPPFP Augmentation, in the sense that it results in more rejections. As shown in the simulation section, the new method appears to be less conservative and more powerful as compared to the augmentation procedure.

It is also important to note that a majority of the the codons which were rejected by the new method, as well as the subset rejected by the augmentation method, are biologically relevant and therefore are associated with an outcome of replication capacity. In particular, protease positions $pr32$, $pr34$, $pr43$, $pr46$, $pr47$, $pr54$, $pr55$, $pr82$, and $pr90$, and reverse transcriptase positions $rt41$, $rt184$, and $rt215$, have been singled out in previous research as related to replication capacity and/or antiretroviral resistance (Birkner et al., 2004; Segal et al., 2004; Shafer et al., 2001). This new method illustrates that 11 of these positions are significant at the $\alpha = 0.05$ level, whereas the augmentation method was only able to identify 5 codons at that significance level. A further discussion of all of these biological findings are outlined in Birkner et al. (2005).

5 Summary

This paper has introduced a new multiple testing for controlling TPPFP(q), as well as a simulation study investigating its performance relative to previous proposals, and we used it to detect codons in the HIV-virus significantly associated with replication capacity of the virus. Our technique still fully uses the generally valid null-value shifted resampling based null distribution for the test-statistics, as generally proposed in our previous work (Pollard and van der Laan (2003) and Dudoit et al. (2004b)), and thereby avoids the need for the so called subset pivotality condition needed in the resampling based multiple testing literature presented in Westfall and Young (1993). Our method also uses the mixture model previously used to obtain FDR-procedures (Efron et al. (2001a)) to generate random guesses of the set of true null hypotheses, which are asymptotically degenerate at the set of true null hypotheses. We have provided a finite sample rational, and formal asymptotic results.

Our simulations show that the new method is significantly more powerful and controls the type-I error at a level much closer to the nominal level α than the competing methods in the important settings for which the number of tests is very large. The practical utility of our method was evidence in our data analysis which showed that our new procedure identified several codons with significant associations, which were not identified by the augmentation procedure or marginal p-value methods proposed in the literature.

References

Jason D. Barbour, Terri Wrin, Robert M. Grant, Jeffrey N. Martin, Mark R. Segal, Christos J. Petropoulos, and Steven G. Deeks. Evolution of Phenotypic Drug Susceptibility and Viral Replication Capacity during Long-Term Virologic Failure of Protease Inhibitor Therapy in Human Immunodeficiency Virus-Infected Adults. *Journal of Virology*, 76(21):11104–11112, 2002.

Merrill D. Birkner, Katherine S. Pollard, Mark J. van der Laan, and Sandrine Dudoit. Multiple Testing Procedures and Applications to Genomics. Technical Report 168, Division of Biostatistics, University of California, Berkeley, January 2005. URL `http://www.bepress.com/ucbbiostat/paper168`.

Merrill D. Birkner, Sandra E. Sinisi, and Mark J. van der Laan. Multiple Testing and Data Adaptive Regression: An Application to Hiv-1 Sequence Data. (161), October 2004. URL `http://www.bepress.com/ucbbiostat/paper161`.

Sandrine Dudoit, Mark J. van der Laan, and Merrill D. Birkner. Multiple Testing Procedures for Controlling Tail Probability Error Rates. Technical Report 166, Division of Biostatistics, University of California, Berkeley, December 2004a. URL `http://www.bepress.com/ucbbiostat/paper166`.

Sandrine Dudoit, Mark J. van der Laan, and Katherine S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004b. URL `http://www.bepress.com/sagmb/vol3/iss1/art13`. Article 13.

B. Efron, J.D. Storey, and R. Tibshirani. Microarrays, Empirical Bayes Methods, and False Discovery Rates. (218), 2001a.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 2001b.

C.R. Genovese and L. Wasserman. A Stochastic Process Approach to False Discovery Rates. Technical Report 762, Department of Statistics, Carnegie Mellon University, January 2003a. URL `http://www.stat.cmu.edu/cmu-stats`.

C.R. Genovese and L. Wasserman. Exceedance Control of the False Discovery Proportion. Technical Report 762, Department of Statistics, Carnegie Mellon University, July 2003b. URL `http://www.stat.cmu.edu/cmu-stats`.

Matthew J. Gonzales, Ilana Belitskaya, Kathryn M. Dupnik, Soo-Yon Rhee, and Robert W. Shafer. Protease and Reverse Transcriptase Mutation Patterns in HIV Type 1 Isolates from Heavily Treated Persons: Comparison of Isolates from Northern California with Isolates from Other Regions. *AIDS Reseach and Human Retroviruses*, 19(10):909–915, 2003.

E.L. Lehmann and J.P Romano. Generalizations of the Family-wise Error Rate. Technical report, Department of Statistics, Stanford University, 2003.

Katherine S. Pollard and Mark J. van der Laan. Resampling-based Multiple Testing: Asymptotic Control of Type I error and Applications to Gene Expression Data. Technical Report 121, Division of Biostatistics, University of California, Berkeley, June 2003. URL `http://www.bepress.com/ucbbiostat/paper121`.

Mark R. Segal, Jason D. Barbour, and Robert M. Grant. Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art2`. Article 2.

Robert W. Shafer, Kathryn M. Dupnik, Mark A. Winters, and Susan H. Eshleman. A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. In *HIV Sequencing Compendium*, pages 83–133. Theoretical Biology and Biophysics Group at Los Alamos National Laboratory, 2001.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. 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*, 3(1), 2004a. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. Technical Report 1, 2004b. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

P. H. Westfall and S. S. Young. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley and Sons, 1993.

Thomas D. Wu, Celia A. Schiffer, Matthew J. Gonzales, Jonathan Tyalor, Rami Kantor, Sunwen Chou, Dennis Israelski, Andrew R. Zolopa, W. Jeffrey Fessel, and Robert W. Shafer. Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different Protease Inhibitor Treatment. *Journal of Virology*, 77(8):4836–4847, 2003.