University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 121

Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data

Katherine S. Pollard* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, kpollard@gladstone.ucsf.edu

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

http://biostats.bepress.com/ucbbiostat/paper121

Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data

Katherine S. Pollard and Mark J. van der Laan

Abstract

We define a general statistical framework for multiple hypothesis testing and show that the correct null distribution for the test statistics is obtained by projecting the true distribution of the test statistics onto the space of mean zero distributions. For common choices of test statistics (based on an asymptotically linear parameter estimator), this distribution is asymptotically multivariate normal with mean zero and the covariance of the vector influence curve for the parameter estimator. This test statistic null distribution can be estimated by applying the non-parametric or parametric bootstrap to correctly centered test statistics. We prove that this bootstrap estimated null distribution provides asymptotic control of most type I error rates. We show that obtaining a test statistic null distribution from a data null distribution, e.g. projecting the data generating distribution onto the space of all distributions satisfying the complete null), only provides the correct test statistic null distribution if the covariance of the vector influence curve is the same under the data null distribution as under the true data distribution. This condition is a weak version of the subset pivotality condition. We show that our multiple testing methodology controlling type I error is equivalent to constructing an error-specific confidence region for the true parameter and checking if it contains the hypothesized value. We also study the two sample problem and show that the permutation distribution produces an asymptotically correct null distribution if (i) the sample sizes are equal or (ii) the populations have the same covariance structure. We include a discussion of the application of multiple testing to gene expression data, where the dimension typically far exceeds the sample size. An analysis of a cancer gene expression data set illustrates the methodology.

1 Introduction

Multiple testing methods are hypothesis testing procedures designed to simultaneously test $p > 1$ hypotheses while controlling an error rate. Traditional approaches to multiple testing are reviewed by Hochberg and Tamhane [1987]. More recent developments in the field include resampling methods (Westfall and Young [1993]), step-wise procedures, and the false discovery rate (Benjamini and Hochberg [1995]). In the past few years, there has been increased interest in the field of multiple testing due to new technologies, such as gene expression arrays, that produce data for which (i) the dimension is much larger than the sample size, (ii) the variables (*e.g.*: genes) are often correlated, and (iii) some proportion of the null hypotheses is expected to be true. Gene expression studies have motivated us to better understand error control in multiple hypothesis testing, though the results in this paper apply to multiple testing in general. We discuss some implications specific to gene expression studies (where the dimension far exceeds the sample size) in Section 5.

Current multiple hypothesis testing methods aim to control a type I error rate under a data null distribution, defined by either (i) all null hypotheses being true (weak control) or (ii) any configuration of the null hypotheses being true (strong control). We propose a class of multiple testing procedures which are intermediate in strength and provide control of the chosen error rate under the *true* data generating distribution. We provide a multivariate normal null distribution for test statistics based on asymptotically linear estimators and show that control of the error rate under this null distribution guarantees asymptotic control.

We begin by formally defining the statistical framework for multiple testing in Section 2. We discuss specific choices of null distribution and methods of estimation. We reach the important practical conclusion that the standard bootstrap method provides the asymptotically correct null distribution for multiple testing. This approach does not require the subset pivotality condition given in Westfall and Young [1993], which is a condition needed to ensure that control under a data generating distribution satisfying the complete null gives the desired control under the true data generating distribution. We also generalize the equivalence of hypothesis testing and confidence regions to the multiple testing framework, illustrating that bootstrap-based estimated error rate specific confidence regions can be used for multiple testing without requiring the analyst to explicitly identify the null distribution of the test statistics. Specifically, our multiple testing method is equivalent with constructing a $1 - \alpha$ error specific confidence region and checking if

the hypothesized value is contained in it. In Section 4, we consider the two sample problem and compare different choices of test statistics and estimated null distributions algebraically and in simulations. We observe that the permutation distribution has the incorrect covariance *unless* (i) the two populations have the same covariance structure or (ii) the sample sizes are equal (*i.e.*: a balanced design). Section 5 discusses the case where the dimension far exceeds the sample size ($p \gg n$), including applications to gene expression studies. We then demonstrate the proposed methodology on a publicly available gene expression data set in Section 5.1. In Section 6, we offer some conclusions and topics for future research.

2 Multiple Testing Procedures

2.1 Data and Null Hypotheses

Let $X_1, \ldots, X_n$ be i.i.d. $X \sim P \in \mathcal{P}$, where $\mathcal{P}$ is a model, X is a p-dimensional vector, possibly including covariates and outcomes. Consider real valued parameters $\mu_j(P) \in \Re$, $j = 1, \ldots, p$. These parameters could be, for example, location parameters (*e.g.*: means/medians or differences between two population means/medians) or regression parameters (*e.g.*: association between expression and outcome in a linear/logistic model). Suppose we are interested in simultaneously testing the null hypotheses:

$$H_{0,j} : \mu_j(P) = \mu_j^0, j = 1, \ldots, p, \tag{1}$$

where the μ_j^0 are hypothesized null values, frequently zero.

We can then define a multiple testing procedure $MT(c)$ in terms of:

1. a vector T_n of test statistics T_{jn}, $j = 1, \ldots, p$,
2. a procedure $MT(c)$ given a vector $c \in \mathbb{R}^p$ defined by:

$$\text{Reject } H_{0j}, \text{ if } \mid T_{jn} \mid > c_j, \; j = 1, \ldots, p. \tag{2}$$

3. an error rate of $MT(c)$ that we wish to control at level α,
4. a vector function cut-off rule $c(Q, \alpha) \in \mathbb{R}^p$ such that if $T_n \sim Q$ then $MT(c(Q, \alpha))$ has an error rate exactly equal to α,
5. a null distribution Q_0 for the vector of test statistics such that $MT(c(Q_0, \alpha))$ has asymptotic control, and

6. an estimator Q_{0n} of Q_0 and corresponding estimated cut-offs $c_n = c(Q_{0n}, \alpha)$.

We discuss each of these components in more detail in the following Sections.

2.2 Test Statistics

Let μ_{jn} be an estimator of $\mu_j(P)$ based on $X_1, \ldots, X_n$, $j = 1, \ldots, p$. If μ_{jn} is asymptotically linear with influence curve $IC_j(X)$; that is,

$$\sqrt{n}(\mu_{j,n} - \mu_j) = \frac{1}{n}\sum_{i=1}^{n} IC_j(X_i|P) + op(1), \tag{3}$$

then by the central limit theorem,

$$\sqrt{n}(\mu_n - \mu(P)) \xrightarrow[n\to\infty]{D} N(0, \Sigma(P)), \tag{4}$$

where $\Sigma = \Sigma(P) = E(IC(X)IC(X)^\top)$ is the covariance of the vector influence curve $IC(X) = \{IC_j(X) : j = 1, \ldots, p\}$. Let

$$Q_0(P) = N(0, \Sigma(P)) \tag{5}$$

denote this limit distribution.

It follows that sensible choices of test statistics include:

$$T_{jn} \equiv \mu_{jn} - \mu_j^0, \tag{6}$$

$$T_{jn} \equiv \sqrt{n}(\mu_{jn} - \mu_j^0), \tag{7}$$

$$T_{jn} \equiv (\mu_{jn} - \mu_j^0)/sd(\mu_{jn}). \tag{8}$$

where $sd(\mu_{jn})$ is an estimate of $\sigma_j = \sqrt{VAR(IC_j(X))/n}$. Let $Q_n = Q_n(P)$ denote the true distribution of the vector of test statistics T_n under $X \sim P$. Let $\mathcal{M}_n = \{Q_n(P) : P \in \mathcal{P}\}$ denote the model for T_n implied by the data generating model $\mathcal{P}$.

In Section 2.8, we show that (5) is the asymptotically correct null distribution for the vector of test statistics (7) whenever μ_n is asymptotically linear. There is only one such distribution Q_0. We note that most choices of μ_n used in practice (*e.g.*: sample means, regression parameters) are in fact asymptotically linear. If one were to use the standardized test statistics (8), then the asymptotically correct null distribution would be $N(0, \rho(P))$, where $\rho(P)$ is the correlation (rather than covariance) matrix of $IC(X)$.

Standardizing test statistics so that the asymptotic marginal distributions of all T_{jn} are $N(0,1)$ (*e.g.*: dividing by $sd(\mu_{jn})$) is a useful tool when one wishes to use tabled null distributions. Figure 1 shows that in the gene expression context, however, finite sample estimates of marginal null distributions can be far from $N(0,1)$, even for standardized test statistics and reasonably large sample sizes. In particular, estimation of $sd(\mu_{jn})$ is known to be difficult in the gene expression context (Tusher et al. [2001], Rocke and Durbin [2001]). Furthermore, for most error rates multiple testing procedures with asymptotic control require estimating a multivariate distribution which is not identified by the p marginal distributions. Using a resampling-based multivariate distribution also eliminates the need to use standardized test statistics, except that standardized test statistics might approach their limit distribution faster (Hall [1992]). We revisit this issue in the simulations of Section 4.4, where we compare choices of test statistics.

2.3 Error Control

Multiple testing procedures can be assessed based on estimates of how many erroneous testing decisions they make.

2.3.1 Type I Error Rates

We assume the reader is familiar with the distinction between type I (false positive) and type II (false negative) errors in the standard univariate setting, where the typical approach is to control the type I error rate at a pre-specified level α and compare different procedures with type I error rate α based on their type II error rates (or power). Dudoit et al. [2002] compare different generalizations of type I error control to the multiple testing setting.

Let $S_0 = \{j : \mu_j(P) = \mu_j^0\}$ be the set of true negatives. Given a vector of cut-off values c, define the following random variables:

$$V(c|Q) = \sum_{j \in S_0} I(|\ T_{jn}\ | > c_j), \tag{9}$$

$$R(c|Q) = \sum_{j=1}^{p} I(|\ T_{jn}\ | > c_j), \text{ where } T_n \sim Q. \tag{10}$$

We use the absolute value of the test statistic $|\ T_{jn}\ |$ since we focus on two-sided tests here, but one-sided testing is also handled by our framework. The notation $V(c \mid Q)$ acknowledges that the distribution of $\sum_{j \in S_0} I(|\ T_{jn}\ | > c_j)$

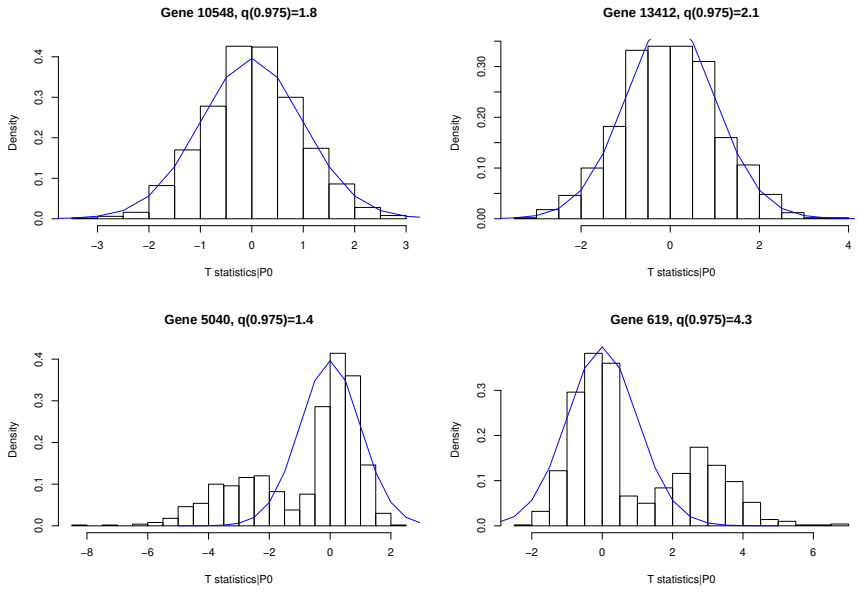

Figure 1: Histograms of null distributions of standardized t-statistics for four genes from the DLBCL data set of Alizadeh et al. [2000] computed by the non-parametric bootstrap. The value of the 0.975 quantile of each distribution is given in the title. The Student's T distribution with appropriate degrees of freedom ($df = 38$) is superimposed on each histogram, showing that the distributions can be heavy/light in the tails or quite skewed. The 0.975 quantile of the T distribution is 2.0.

is defined by the distribution of T_n. We will also some times use the notation $R(c \mid Z)$, where Z is the random variable of interest. If $Z \sim Q$, then $R(c \mid Z) = R(c \mid Q)$.

Let $V_n = V(c|Q_n(P))$ be the number of false positives of the testing procedure $MT(c)$, and let $R_n = R(c|Q_n(P))$ be the total number of rejected hypotheses. For a discrete distribution F on $\{0, \ldots, p\}$, define a real valued parameter $\theta(F) \in (0, 1)$ representing a particular type I error rate, where F represents a candidate for the distribution of V_n. We will use the notation F_X to denote the cumulative distribution of a random variable X. We wish to arrange that $\theta(F_{V_n}) \leq \alpha$, at least asymptotically. This is the error rate for $MT(c)$. Given the distance measure $d(F_1, F_2) = \max_{j \in \{0,\ldots,p\}} \mid F_1(\{j\}) - F_2(\{j\}) \mid$ for two such cumulative distribution functions F_1, F_2

on $\{0, \ldots, p\}$, we assume that this parameter $\theta(F)$ satisfies the following properties:

$$\text{Monotonicity: if } F_1 \geq F_2, \text{ then} \qquad \theta(F_1) \leq \theta(F_2) \tag{11}$$

$$\text{Uniform Continuity: if } d(F_n, G_n) \to 0, \text{ then} \qquad \theta(F_n) - \theta(G_n) \to 0 \tag{12}$$

Let $Z_n \equiv \sqrt{n}(\mu_n - \mu)$. We note that $V_n = \sum_{j=1}^{p} I(\mid Z_{jn} \mid > c_j, j \in S_0)$. Let k be a user supplied constant. Then, some error rates which are functions of the distribution F_{V_n} of V_n include:

- $\theta(F_{V_n}) = \int x dF_{V_n}(x)/p = E(V_n)/p$: per-comparison error rate (**PCER**),
- $\theta(F_{V_n}) = \int x dF_{V_n}(x) = E(V_n)$: per-family error rate (**PFER**),
- $\theta(F_{V_n}) = median(F_{V_n})$: median-based per-family error rate (**mPFER**),
- $\theta(F_{V_n}) = 1 - F_{V_n}(k-1) = Pr(V_n \geq k)$: generalized family-wise error rate (**gFWER**).

Note that when $k = 1$, the gFWER is the usual family-wise error rate (FWER).

In general, the per-family error rate is most conservative and the per-comparison error rate (ignoring the multiplicity problem) is the least conservative (Dudoit et al. [2002]). In the gene expression context, a less conservative error rate is often preferred since researchers view gene expression experiments as exploratory methods and are usually interested in obtaining a fairly large list of candidate genes, even if some proportion of these are likely to be false positives. For this reason, the false discovery rate (FDR) is becoming a popular choice of error rate (Benjamini and Hochberg [1995]). The FDR is a function of the distribution of V_n/R_n, and not simply F_{V_n}:

$$\theta = \begin{cases} E(V_n/R_n) & R_n \geq 0 \\ 0 & R_n = 0 \end{cases} \text{ : false discovery rate (\textbf{FDR})}$$

The FDR method of Benjamini and Hochberg [1995] only provides asymptotic control under independence or a particular type of dependence, and is therefore of a non-parametrically non-identifiable type-I error, and thus falls in a very different class of error rates than the ones we have studied. In particular, their FDR method does not have level α when the complete null hypothesis $H_0^C = \bigcap_{j=1}^{p} H_{0,j}$ is true, but it does control $E(V_n(c)/R_n(c))$ under the true data generating distribution *given* certain independent assumptions about this distribution. The methods we have proposed here

can not provide control of the FDR under the true distribution. We note, however, that weak control of the FDR under any test statistic distribution $Q_n(P_0)$ where P_0 satisfies the complete null is equivalent to weak control of the FWER.

2.3.2 Types of Error Rate Control

Error rates are defined under the true data generating distribution P, so that they depend on which hypotheses are in fact true. In practice, we do not know which hypotheses are true since we do not know either P or $Q_n(P)$, so we have to choose a way to compute the expectations and/or probabilities in the error rate. The goal of multiple hypothesis testing is to control the chosen error rate θ under the *true* data generating distribution P. We refer to this as "control under the true distribution" or simply "control". There are several approaches to this problem. Current methods control the error rate under a particular distribution for the test statistics $Q_n(P_0)$ implied by a choice P_0 of *data* null distribution. **Weak control** means that P_0 is a data generating distribution that satisfies the complete null hypothesis $H_0^C = \bigcap_{j=1}^{p} H_{0,j}$. One popular choice is $Q_n(P_0)$ (estimated by $Q_n(P_{0n})$) as defined in Section 2.7.3. There are many data generating distributions satisfying H_0^C, but most of these do not imply the correct null distribution for the test statistics. Equation (16) gives the condition under which $Q_n(P_0)$ is correct. **Strong Control** means that $\theta \leq \alpha$ under *any* choice of data generating distribution P_0 represented by one of the different configurations of the null hypotheses (referred to as control "under all configurations", Hochberg and Tamhane [1987]). **Asymptotic control** means that the error rate α_n for a sample of size n has the property $\limsup_{n\to\infty}\alpha_n \leq \alpha$ under P. Asymptotic strong and weak control are defined similarly. We discuss asymptotic control further in Section 2.8.

We have two critiques of current practice. First, in general (*i.e.* when some $H_{0,j}$ are true and some false), control under the true distribution is stronger than weak control but weaker than strong control, so that neither approach is ideal. Second, a test statistic null distribution derived via a data null distribution is only the correct distribution for multiple testing under certain conditions (*e.g.*: the subset pivotality condition of Westfall and Young [1993] or the weaker Equation (16) provided below). Hence, we propose the following multiple testing method, which is intermediate in strength and does not rely on a data null distribution.

2.4 Null Distribution

In order to decide if any of the observed test statistics are sufficiently unusual to reject the corresponding null hypotheses, we compare them to a *joint* null distribution for T_n. We prove in Section 2.8 that for $T_n = \sqrt{n}(\mu_n - \mu^0)$ the asymptotically correct null distribution is $Q_0 = N(0, \Sigma(P))$. It is interesting to note that Q_0 can be viewed as the Kullback-Leibler projection of the asymptotic distribution of T_n onto the space of multivariate distributions with mean zero (*i.e.*: the limit of the projection $T_n - E(T_n)$ of T_n). Hence, rejection decisions based on Q_0 can be directly attributed to the parameter of interest (*e.g.*: $\mu_j \neq \mu_j^0$ for some j) and not to other (nuisance) parameters. In practice, we do not know the true distribution P and hence must use an estimated test statistic null distribution. In Section 2.7, we present two resampling-based estimators for which asymptotic control is achieved under weak regularity conditions (see Section 2.8).

2.5 Cut-off Rule

Consider test statistics T_n, an error rate θ with target level α, and a two-sided multiple testing procedure $MT(c)$ defined by the decision rule:

$$\text{Reject } H_{0,j}, \text{ if } \mid T_{jn} \mid > c_j, j = 1, \ldots, p,$$

and the following method for choosing c. Given a null distribution Q, we let $c = c(Q, \alpha) \in \mathbb{R}^p$ be a vector function cut-off rule such that if $T_n \sim Q$ and F is the distribution of $R(c \mid Q)$, then $MT(c)$ has the property that $\theta(F) = \alpha$. For a one-sided test, only one tail of Q is used. Notice that $MT(c)$ depends critically on the choice of joint null distribution through c. One particular method for computing c is to select a common quantile of each marginal distribution from the null distribution Q. Consider, for instance, a vector of thresholds $\{c_j : j = 1, \ldots, p\}$ satisfying

$$Pr\left(\sum_{j=1}^{p} I\left\{|T_{jn}| > c_j\right\} > k\right) \leq \alpha, \; T_n \sim Q \tag{13}$$

where k is a pre-specified number of false positives. When $k = 1$ this is the usual FWER, and when $k > 1$ this controls $Pr(V_n > k) \leq \alpha$ under the distribution Q. In practice, we need to take B resamples from Q and compute the cut-offs under the corresponding empirical distribution. With a sufficiently smooth resampled null distribution Q in hand (B large enough),

these common quantiles can be fine-tuned to control the chosen error rate exactly under Q.

The multiple testing procedure is now completely defined by a choice of null distribution for the test statistics. We prove in Section 2.8 that for $T_n = \sqrt{n}(\mu_n - \mu^0)$ if we use $MT(c_0)$ with $c_0 \equiv c(Q_0, \alpha)$, then we have asymptotic control. This shows that $Q_0 = N(0, \Sigma(P))$ is the asymptotically correct null distribution. It is interesting to note that Q_0 can be viewed as the limit of the Kullback-Leibler projection of the distribution $Q_n(P)$ of T_n onto the space of mean zero distributions. In practice, we do not know the true distribution P, so Q_0 is unknown. Therefore, we use estimated cut-offs $c_{0n} = c(Q_{0n}, \alpha)$, which depend on an estimated null distribution Q_{0n}. If Q_{0n} is a consistent estimator of Q_0, we can asymptotically control the error rate at level α up to the discreteness of the resampled test statistic distribution.

In traditional testing settings, a common threshold is used to make the testing decision for every variable, *i.e.*: Reject $H_{0,j}$ if $|T_j| > c(\alpha)$ for a specified level α. The common quantile method is a generalization of this approach, which corresponds with a common threshold *only* if the marginal distributions have identical tail probabilities, which is not the case in many applications.

2.6 Comparison with P-value Adjusting Methods

An alternative approach to multiple testing is to compute marginal p-values (*i.e.*: the probability of observing a statistic as or more extreme than T_{jn}) and adjust these for multiple tests. Westfall and Young [1993],Yekutieli and Benjamini [1999] and Dudoit et al. [2002] review different methods for computing adjusted p-values. Some of the computational and practical advantages to using adjusted p-values (compared to thresholds) include:

1. no sorting is required for computation of adjusted p-values,
2. the target error rate α does not have to be chosen in advance,
3. p-values offer a measure of strength of evidence (versus an accept/reject decision),
4. p-values can be used to order the genes, even when they do not have the same marginal distributions.

On the other hand, by reducing the resampled null distribution Q_{0n} to marginal p-values, one looses the opportunity to control the error rate exactly at level α under Q_{0n}, contrary to the quantile-based method described above.

Stepwise p-value adjusting methods for controlling FWER allow one to achieve a level closer to α than single-step methods (*i.e.*: they are less conservative and more powerful) (Westfall and Young [1993],Dudoit et al. [2002]), but they are still not exact under Q_{0n}. In other words, step-down methods allow one to recover only some of the loss incurred by reducing Q_{0n} to marginal p-values. These procedures for adjusted p-values can also be stated as equivalent methods for choosing thresholds. Table 1 contains formulas for thresholds based on some popular multiple testing p-value adjustments. As with the corresponding p-value methods, these quantiles are relatively quick to compute, but do not allow one to control the error rate exactly under Q_{0n}. We also note that the step-down methods depend on the observed data, so they do not produce thresholds of the form $c = c(Q, \alpha)$, which only depend on the null distribution and level α. Hence, the theoretical results of Section 2.8 do not apply to such threshold rules.

	Bonferroni/Holm	Šidák	Westfall & Young
single-step	α/p	$1-(1-\alpha)^{1/p}$	$q(\alpha)$ of $\max_{l\leq p}\lvert T_l\rvert$
step-down	$\alpha/(p-r_j+1)$	$1-(1-\alpha)^{1/(p-r_j+1)}$	$q_j(\alpha)$ of $\max_{l\leq r_j}\lvert T_l\rvert$

Table 1: Formulas for computing thresholds based on several methods for p-value adjustment. In each case, the threshold c_j is the $1-\delta_j$ quantile of the null distribution of resampled test statistics $|T_{jn}^b|$, where is δ_j is determined by the given formula. For step-down methods, the $\{r_j\}$ are the order statistics of $\{|T_{jn}|\}$ and $(p - r_j + 1) = \text{rank}(|T_{jn}|)$. If the $1-\delta_j$ quantile for each gene $j = 1, \ldots, p$ is chosen from the estimated resampling-based joint null distribution $\{|T_{jn}^b| : b = 1, \ldots, B, j = 1, \ldots, p\}$, then these methods are equivalent to computing unadjusted marginal p-values from the estimated joint null distribution and then applying the corresponding procedure to obtain adjusted p-values. The single-step methods only use the marginal distribution of each gene to compute the threshold so that they are quick to compute, but do not give a very tight bound on the error whenever the genes are not independent. The formula for single-step $\max T$ shows that this p-value adjustment is equivalent to a common threshold (as is the single-step $\min P$ method).

In addition, it is the case that in many applications (*e.g.*: gene expression studies), the goal of testing is usually to select a subset of interesting variables (*e.g.*: genes) for further analysis, such as clustering or classification. Hence, it makes sense to examine a few different subsets (choices of α) up

front, but to then make a testing decision and stick to it for the remainder of the analysis. In this case, threshold-based methods make practical sense in addition to having the advantage of being able to control the error rate exactly under Q_{0n}.

2.7 Estimation of the Test Statistic Null Distribution

We present two resampling-based estimators of the asymptotically correct test statistic null distribution $Q_0 = N(0, \Sigma(P))$. For both estimators, asymptotic control is achieved under weak regularity conditions (see Section 2.8). We first give the specific estimators for $T_n = \sqrt{n}(\mu_n - \mu^0)$ and then discuss adaptations for standardized test statistics. We then compare these methods to the common approach based on first estimating a null distribution for the data.

2.7.1 Estimating $\Sigma(P)$

The first proposed estimator is $\tilde{Q}_{0n} = N(0, \Sigma_n)$, where Σ_n is an estimate of the covariance matrix $\Sigma(P)$ based on an estimate of the influence curve $IC(X)$. The null distribution of the test statistics is estimated by generating a large number B of resampled data sets from $\tilde{Q}_{0n}$. If Σ_n is an asymptotically consistent estimator of $\Sigma(P)$, then it follows that $\tilde{Q}_{0n}$ converges in distribution to Q_0, conditional on the data. If one were to use the standardized test statistics $T_n = (\mu_n - \mu^0)/sd(\mu_n)$, then the asymptotically correct null distribution is $N(0, \rho(P))$, and one can use $N(0, \rho_n)$ as an estimated null distribution, where ρ_n is a consistent estimator of the correlation $\rho(P)$ of $IC(X)$.

2.7.2 Bootstrap Method

The second proposed estimator involves a simple bootstrap method. Let $\tilde{P}_n$ be an estimator of the true data generating distribution P according to the model $\mathcal{P}$ or the empirical distribution (*i.e.*: model based bootstrap or nonparametric bootstrap). Let $\tilde{\mu}_n = \mu(\tilde{P}_n)$ and let $\mu_n^\#$ be the estimator μ_n but now applied to n i.i.d. copies $X_1^\#, \ldots, X_n^\#$ of $X^\# \sim \tilde{P}_n$. Let $Z_n^\# = \sqrt{n}(\mu_n^\# - \tilde{\mu}_n)$. We now estimate the distribution Q_0 with the distribution $Q_{0n}^\#$ of $Z_n^\#$. Under regularity conditions, it is known that the bootstrap is consistent in the sense that $Z_n^\# \stackrel{D}{\Rightarrow} Z \sim Q_0$ conditional on $\tilde{P}_n$, and hence $Q_{0n}^\#$ converges to Q_0 conditional on the data (*e.g.* van der Vaart and Wellner

[1996]). Define

$$R^{\#}_{0n}(c) \equiv R(c|Q^{\#}_{0n}) = \sum_{j=1}^{p} I(\mid Z^{\#}_{jn} \mid > c_j). \tag{14}$$

A bootstrap based multiple testing procedure controlling θ at level α is then defined by $MT(c_n)$, with $c(Q^{\#}_{0n}, \alpha)$ being a solution of $\theta\left(F_{R^{\#}_{0n}(c)}\right) = \alpha$. If $T_n = (\mu_n - \mu^0)/sd(\mu_n)$, then the bootstrap test statistics should also be standardized, for example $Z^{\#}_{jn} = (\mu^{\#}_{jn} - \tilde{\mu}_{jn})/sd(\mu^{\#}_{jn})$, where $sd(\mu^{\#}_{jn})$ is an estimate of $\sigma^{\#}_j = \sqrt{VAR(IC_j(X^{\#}))/n}$. Similarly, if $T_n = (\mu_n - \mu^0)$, then the bootstrap test statistics are not multiplied by $\sqrt{n}$: $Z^{\#}_n = (\mu^{\#}_n - \tilde{\mu}_n)$.

Note that this method can be easily generalized to two-sided tests. In this case, one uses the absolute value of the test statistic (*e.g.*: $T_n = \sqrt{n}|\mu_n - \mu^0|$) and computes an estimated null distribution $Q^{\#}_{0n}$ based on resampled test statistics $Z^{\#}_n = \sqrt{n}|\mu^{\#}_n - \tilde{\mu}_n|$.

2.7.3 Problems with Using a Data Null Distribution

Our method of resampling-based multiple testing is new. The current resampling-based multiple testing methodology identifies a null *data* distribution P_0 and controls the error rate under an estimator P_{0n} of P_0. For example, the prepivoting methods discussed in Beran [1988] utilize an estimated null hypothesis data model. Heteroscedastic bootstrapping (both parametric and non-parametric) is discussed in Westfall and Young [1993] (p.89-91, 123-125), where residuals are resampled (*e.g.*: the data is first centered around an estimate). This approach, often called "null restricted" bootstrap, requires the subset pivotality condition (Westfall and Young [1993] (p.42-43)) or specifically the weaker condition $\Sigma(P_0) = \Sigma(P)$ (Equation (16)), which is violated in many applications. On the contrary, our method samples from an estimate of the true distribution, but standardizes the test statistics correctly. Therefore, we *always* consistently estimate the covariance matrix of the test statistics (even when Equation (16) does not hold).

Formally, the method based on a data null distribution works as follows. An estimator of Q_0 is derived in two stages. First, one derives a data null distribution $P_0(P)$ by projecting P onto the space $\mathcal{P}_0 = \{P \in \mathcal{P} : \mu = \mu^0\}$. We illustrate below that a projection parameter $P_0(P)$ is necessary (but not sufficient) for this method to achieve control under P. A particular

candidate for such a $P_0(P)$ is the Kullback-Leibler projection:

$$P_0(P) = \arg\max_{P_0' \in \mathcal{M}_0, P_0' \ll \mu} \int \log\left(\frac{\partial P_0'(x)}{\partial \mu(x)}\right) dP(x), \tag{15}$$

where μ is a user supplied dominant measure. For example, in a shift experiment where the parameter of interest is a location parameter and the data model is non-parametric, one would use $P_0(P) = P(\cdot - \mu^0)$. The maximum likelihood estimator of $P_0(P)$ is $P_{0n} = P_0(P_n)$, where P_n denotes the empirical distribution of the data.

The second stage is to form an estimated test statistic null distribution $Q_n(P_{0n})$. Since $Q_n(P_{0n}) \Rightarrow N(0, \Sigma(P_0))$, this method provides asymptotic control if and only if

$$\Sigma(P_0) = \Sigma(P). \tag{16}$$

This condition is weaker than the subset pivotality condition (Westfall and Young [1993]), which requires that $\Sigma(P) = \Sigma(P^*)$ for any P^* corresponding with a configuration of the hypothesized parameters. In other words, Equation (16) requires that replacing μ by μ^0 does not affect the covariance matrix of the vector influence curve, while subset pivotality requires no change in the covariance matrix for *all* configurations of μ. In many examples, subset pivotality holds whenever Equation (16) is true, but in practice we do not need the stronger subset pivotality condition in order to have asymptotic control. Whenever Equation (16) holds, it is correct to use the null restricted bootstrap ($Q_n(P_{0n})$) as well as our proposed ordinary bootstrap ($Q_{0n}^{\#}$), which is always correct. The following example helps to illustrate when $Q_n(P_{0n})$ is not asymptotically equivalent to $Q_{0n}^{\#}$ so that using $Q_n(P_{0n})$ is not correct, but $Q_{0n}^{\#}$ still provides asymptotic control.

Example: **Testing for zero correlation**

Let $X_1, \ldots, X_n$ be i.i.d. $X \sim P$, where P is a p-variate normal distribution. Suppose we are interested in testing whether the correlations between all variables are zero: $H_{jk} : \rho_{jk} = 0$, for $j = 1, \ldots, p$ and $k = j+1, \ldots, p$. Commonly used test statistics are $\sqrt{n}$ times the sample correlations. Westfall and Young [1993] study this problem (p.43), and note that the joint distribution of a pair of test statistics depends on the correlation between the corresponding variables, so that subset pivotality fails. Equivalently, changing the hypothesized parameters changes the asymptotic covariance of the vector influence curve for the sample correlations, which is not the same under P as under a multivariate normal distribution P_0 for which H_{jk} is true for all (j, k). We wish to asses the performance of the null restricted

bootstrap ($Q_n(P_{0n})$) and our proposed ordinary bootstrap ($Q_{0n}^{\#}$). Clearly, neither procedure provides asymptotic strong control. The ordinary bootstrap does provide asymptotic control, however, since $Q_{0n}^{\#}$ is the distribution of $\sqrt{n}(\rho_n^{\#} - \tilde{\rho}_n)$, where $\rho_n^{\#}$ is the vector of sample correlations in the bootstrap sample and $\tilde{\rho}_n$ is the sample correlation in the original sample, which converges to Q_0. The null restricted bootstrap, in contrast, does not provide asymptotic control. This example illustrates that requiring strong control is too much and not necessary, since our proposed method controls the error rate under the true data generating distribution, which is all that one cares about.

2.8 Asymptotic Control Theorem

We prove that the proposed class of multiple testing procedures have asymptotic control of the wished multiple testing type I error rate.

Theorem 1. *Given data and null hypotheses defined in Section 5.1, consider a parameter $\mu_j = \mu_j(P) \in \mathbb{R}$ with an asymptotically linear estimator μ_{jn}, $j = 1, \ldots, p$. Let $T_{jn} \equiv \sqrt{n}(\mu_{jn} - \mu_{j0})$, $j = 1, \ldots, p$ and $T_n \sim Q_n = Q_n(P)$. Suppose that we use a multiple testing procedure $MT(c)$ as defined in Section 2.5. Then, consider a type I error rate $\theta(F) \in (0, 1)$ satisfying Assumptions (11) and (12) of Section 2.3.1. Let $Z_n \equiv \sqrt{n}(\mu_n - \mu)$ and let $Z \sim Q_0 \equiv N(0, \Sigma(P))$ be the limit (in distribution) of Z_n. We define the following random variables in terms of the distribution of Z: $V_0(c) = V(c \mid Q_0)$ and $R_0(c) = R(c \mid Q_0)$. Let $c = c(Q, \alpha) \in \mathbb{R}^p$ be a vector function of a p-variate distribution Q and α satisfying $\theta(F_{R(c|Q)}) = \alpha$. Let $c_0 = c(Q_0, \alpha)$ and define $V_n(c_0) = V(c_0 \mid Q_n)$. Then the multiple testing procedure $MT(c_0)$ has asymptotic control:*

$$\limsup_{n \to \infty} \theta\left(F_{V_n(c_0)}\right) \leq \alpha. \tag{17}$$

Since the distribution Q_0 of Z is unknown, the distribution of $F_{R_0(c)}$ of $R_0(c)$ is unknown. Consequently, we will need to estimate Q_0. Let Q_{0n} be an estimate of the distribution Q_0 and define $c_{0n} \equiv c(Q_{0n}, \alpha)$. Let $V_{0n}(c) = V(c \mid Q_{0n})$. Suppose that $c_{0n} \to c_0$ in probability for $n \to \infty$. Then

$$\limsup_{n \to \infty} \theta\left(F_{V_{0n}(c_{0n})}\right) \leq \alpha. \tag{18}$$

Suppose that the mapping $Q \to c(Q, \alpha)$ is continuous in the sense that point-wise convergence of the multivariate cumulative distribution of Q_{0n} to the multivariate cumulative distribution of Q_0, at each point, implies $c(Q_{0n}, \alpha) \xrightarrow{P} c(Q_0, \alpha)$ as $n \to \infty$. Under this condition, we have that convergence in distribution of the estimator Q_{0n} to Q_0, conditional on the empirical distribution P_n, implies $c(Q_{0n}, \alpha) \xrightarrow{P} c(Q_0, \alpha)$, and thereby the wished asymptotic control (18).

Proof. We will first prove (17). Recall that $Z \sim Q_0 \equiv N(0, \Sigma(P))$ is the limit (in distribution) of $Z_n \equiv \sqrt{n}(\mu_n - \mu)$. By (11) we have:

$$\theta\left(F_{V_n(c_0)}\right) \leq \theta\left(F_{R(c_0|Z_n)}\right),$$

where $R(c \mid Z_n) = \sum_{j=1}^{p} I(\mid Z_{jn} \mid > c_j)$. By assumption, we have that for $n \to \infty$, the multivariate c.d.f. of Z_n converges to the multivariate c.d.f. $Z \sim Q_0$ at each point. This implies that $d(F_{R(c_0|Z_n)}, F_{R(c_0|Z)}) \to 0$. By the continuity assumption (12) this implies

$$\theta(F_{R(c_0|Z_n)}) \to \theta(F_{R(c_0|Z)}) = \alpha.$$

This proves (17).

It remains to prove (18). It is easy to show that $Pr(V_n(c_{0n}) \neq V_n(c_0)) = O(\delta_n)$, where $\delta_n = \max_{j=1,\ldots,p} \mid c_{0n,j} - c_{0,j} \mid$. Since by assumption $\delta_n \to 0$ in probability, this proves that $Pr(V_n(c_{0n}) = V_n(c_0)) \to 1$ for $n \to \infty$, and thus that $d(F_{V_n(c_{0n})}, F_{V_n(c_0)}) \to 0$. By the uniform continuity (12), this implies that

$$\theta(F_{V_n(c_{0n})}) - \theta(F_{V_n(c_0)}) \to 0 \text{ for } n \to \infty.$$

Thus,

$$\begin{aligned}
\limsup_{n\to\infty} \theta\left(F_{V_n(c_{0n})}\right) &= \limsup_{n\to\infty} \theta\left(F_{V_n(c_{0n})}\right) - \theta\left(F_{V_n(c_0)}\right) \\
&\quad + \limsup_{n\to\infty} \theta\left(F_{V_n(c_0)}\right) \\
&\leq 0 + \limsup_{n\to\infty} \theta\left(F_{V_n(c_0)}\right) \\
&\leq \alpha, \text{ by (17). } \square
\end{aligned}$$

3 Equivalence of Multiple Testing and Confidence Regions

We present a generalization of the equivalence of hypothesis testing and confidence regions, which is multivariate and allows for any choice of error

rate. Let $F_{R(c|Z_n)}$ denote the distribution of $R(c \mid Z_n) = \sum_{j=1}^{p} I(\mid Z_{jn} \mid > c_j)$, where $Z_n = \sqrt{n}(\mu_n - \mu(P))$. Let c_n be chosen such that the error rate $\theta(F_{R(c_n|Z_n)}) = \alpha$. Then, the random region $\{\mu : \sqrt{n}|\mu_n - \mu| < c_n\}$ or

$$\left\{\mu : \mu_{jn} - \frac{c_{jn}}{\sqrt{n}} < \mu_j < \mu_{jn} + \frac{c_{jn}}{\sqrt{n}}, j = 1, \ldots, p\right\} \tag{19}$$

is a θ-specific $(1-\alpha)\%$ confidence region for $\mu(P)$. This is a generalization of the definition of a simultaneous confidence region to any choice of error rate. If $\theta(\cdot)$ is the FWER, then the region defined by (19) is a $(1-\alpha)\%$ *simultaneous* confidence region for $\mu(P)$.

In practice, we do not know the distribution $F_{R(c|Z_n)}$. We can estimate it with the distribution $F_{R(c|Z_n^{\#})}$ of $R(c \mid Z_n^{\#})$, where $Z_n^{\#} \sim Q_{0n}^{\#}$ is the bootstrap random variable $\sqrt{n}(\mu_n^{\#} - \mu_n)$. Let $\tilde{c}_n = c(Q_{0n}^{\#}, \alpha)$. Then,

$$\left\{\mu : \mu_{jn} - \frac{\tilde{c}_{jn}}{\sqrt{n}} < \mu_j < \mu_{jn} + \frac{\tilde{c}_{jn}}{\sqrt{n}}, j = 1, \ldots, p\right\} \tag{20}$$

is an asymptotically correct θ-specific $(1-\alpha)\%$ confidence region for $\mu(P)$. Our multiple testing procedure $MT(\tilde{c}_n)$ defined in Section 2 is equivalent with:

$$\text{Reject } H_{0,j} \text{ if } \mid \sqrt{n}(\mu_{jn} - \mu_j^0) \mid > \tilde{c}_{jn}, \text{ for } j = 1, \ldots, p.$$

In other words, one can perform multiple testing controlling an error rate $\theta(\cdot)$ by using the bootstrap distribution $Q_{0n}^{\#}$ to define a θ-specific confidence region and then checking for every $j = 1, \ldots, p$ if $T_{jn} = \sqrt{n}|\mu_{jn} - \mu_j^0| > \tilde{c}_{jn}$. Equivalently, the multiple testing procedure $MT(\tilde{c}_n)$ equals:

$$\text{Reject } H_{0j} \text{ if } \mu_j^0 \text{ is outside the interval } \left[\mu_{jn} - \frac{\tilde{c}_{jn}}{\sqrt{n}}, \mu_{jn} + \frac{\tilde{c}_{jn}}{\sqrt{n}}\right], \text{ for } j = 1, \ldots, p.$$

REMARK: Westfall and Young [1993] (p.82-83) note the equivalence between multiple testing with the *null restricted* bootstrap controlling FWER and constructing a simultaneous confidence interval based on a *null restricted* bootstrap. This particular equivalence requires the subset pivotality condition (Westfall and Young [1993]).

4 Two Sample Problem

As a specific example, consider the two sample multiple testing problem. Suppose that we observe n_1 observations from population 1 and n_2 from

population 2. We can think of the data as (X_i, L_i), where X_i is the multivariate vector $X_{ij}, j = 1, \ldots, p$ for subject i and $L_i \in \{1, 2\}$ is a label indicating subject i's group membership. Let $\mu_{1,j}$ and $\mu_{2,j}$ denote the means of variable j in populations 1 and 2, respectively. Suppose we are interested in testing

$$H_{0,j} : \mu_j \equiv \mu_{1,j} - \mu_{2,j} = 0, j = 1, \ldots, p. \tag{21}$$

We can define a procedure $MT(c)$ as described in Section 2. We will use the notation D_n for the non-standardized test statistics so that we can compare them with the standardized t-statistics:

$$\begin{aligned} T_{jn} &= (\mu_{jn} - 0)/sd(\mu_{jn}), \\ D_{jn} &= \mu_{jn} - 0. \end{aligned}$$

First, we examine different choices of data models, and then we investigate the implications that each choice of model has in terms of the performance of the implied testing procedure.

4.1 Models

Consider the following data models for this two sample problem:

1. $\mathcal{P}_1$: $X|L = 1 \sim P_1$ and $X|L = 2 \sim P_2$, where P_1, P_2 can be arbitrary distributions,

2. $\mathcal{P}_2$: $X|L = 1 \sim P_0(\cdot - \mu_1)$ and $X|L = 2 \sim P_0(\cdot - \mu_2)$, for a common non-parametric distribution P_0 with mean zero.

Model $\mathcal{P}_2$ makes a much stronger assumption, specifically that under the null hypotheses, the data are identically distributed in the two populations. If we were testing the hypothesis $H_0 : P_1 = P_2$, then this would clearly be a good choice of model, but it may be a poor choice for testing Equation (21). Other choices of models, which might be more parametric, could also be considered.

4.2 Bootstrap Null Distributions

Each of the models implies a different null distribution for the test statistics. Suppose we use the bootstrap estimator $Q_{0n}^{\#}$ as described in Section 2.7.2. For both of the models, we estimate μ_1, μ_2 with the sample means μ_{1n_1}, μ_{2n_2}. If we assume model $\mathcal{P}_1$, then $\hat{P}_n$ is the empirical distribution of (X_i, L_i), and we resample n_1 observations from population 1

and n_2 observations from population 2 *separately* to form the bootstrap samples $X_1^\#, L_1^\#, \ldots, X_n^\#, L_n^\#$. Then, $Q_{0n}^\#$ is the empirical distribution of $Z_n^\# = \sqrt{n}(\mu_{1n_1}^\# - \mu_{2n_2}^\# - (\mu_{1n_1} - \mu_{2n_2}))$. If we assume model $\mathcal{P}_2$, then we first estimate P_0 by making centered observations $X_i - \mu_{1n_1}$ if $L_i = 1$ and $X_i - \mu_{2n_2}$ if $L_i = 2$ and forming the empirical distribution P_{0n} of the *combined* sample of centered observations. Then, we resample n_1 observations from P_{0n} and add μ_{1n_1} and n_2 observations from P_{0n} and add μ_{2n_2} to form the bootstrap samples $X_1^\#, L_1^\#, \ldots, X_n^\#, L_n^\#$. Again, $Q_{0n}^\#$ is the empirical distribution of $Z_n^\#$.

We note that this procedure for $\mathcal{P}_2$ is equivalent to forming a combined empirical distribution of the X_i $(i = 1, \ldots, n)$ and using the distribution of $\sqrt{n}$ times the difference in the sample means when we draw n_1 samples and set $L_i = 1$ and n_2 samples and set $L_i = 2$. This is the resampling (with replacement) analogue of the commonly used permutation test. Remarkably, permutation tests are known to be exact (even for $p >> n$) under the model $\mathcal{P}_2$ (Lehmann [1986] and Puri and Sen [1971]). As noted above, $\mathcal{P}_2$ implies a stronger null model restriction, which is needed for an exact test. In contrast, the bootstrap method implied by model $\mathcal{P}_1$ is only approximate. Note also that the exactness of the permutation test is conditional on the observed data, so that the unconditional significance of an "exact" level α permutation test is less than or equal to α (*i.e.*: it is unconditionally conservative). In other words, even for a finite sample size an "exact" test controls the error rate conservatively (not exactly) in the sense that the error rate $\theta \leq \alpha$.

4.3 Implications for the Permutation Test

4.3.1 Covariance

For simplicity, we suppose that $p = 2$, but note that conclusions about the covariance of two variables can be applied to any pairwise covariance when p is much larger. For variable j, denote the variance of X_i by $\sigma_{1,j}^2$ in population 1 and by $\sigma_{2,j}^2$ in population 2. Let ϕ_1 be the covariance between the two variables in population 1 and ϕ_2 be the covariance between the two variables in population 2. We have derived formulas for the variance of D_j $(j = 1, 2)$ and the covariance of the two test statistics D_1, D_2 under both models (Table 2, derivations in Appendix).

These expressions show us that under most values of the underlying parameters, the bootstrap and permutation distributions of D_j are not equivalent. But, when (i) $n_1 = n_2$ or (ii) $\sigma_{1,j}^2 = \sigma_{2,j}^2 \equiv \sigma_j^2$ $(j = 1, 2)$ and

$\mathcal{P}_1$	$Var(D_j)$	$\frac{\sigma^2_{1,j}}{n_1} + \frac{\sigma^2_{2,j}}{n_2}$
$\mathcal{P}_2$		$\frac{\sigma^2_{1,j}}{n_2} + \frac{\sigma^2_{2,j}}{n_1}$
$\mathcal{P}_1$	$Cov(D_1, D_2)$	$\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}$
$\mathcal{P}_2$		$\frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}$

Table 2: Formulas for the variance and covariance of the difference in means statistic under two different models. It is interesting to note that the roles of n_1 and n_2 are reversed under permutations.

$\phi_1 = \phi_2 \equiv \phi$, then they are the same. Thus, unless one of these conditions holds we recommend using a bootstrap distribution since it preserves the correlation structure of the original data. When a study is "balanced" ($n_1 = n_2$), however, these results suggest that one should use the equivalent permutation distribution, because the variances and covariances are the same for both populations and estimates of these "pooled" values (which make use of all n subjects) are more efficient. Notice that if we were to use the usual standardized t-statistics $T_{jn} = (\mu_{jn} - \mu_j^0)/sd(\mu_{jn})$, despite the fact that the variances are equal under both models, the covariances are still not equivalent unless $n_1 = n_2$ or the correlation structures are the same in the two populations.

4.3.2 Bias

We have also found that resampling-based estimated null distributions of standardized t-statistics do not have mean zero whenever $n_1 \neq n_2$, unless the observed difference in means is zero. For the permutation method, this bias depends on the observed difference in means (Figure 2), while for the bootstrap methods the bias is independent of the observed difference. This finite sample bias arises from using a variance estimate in the denominator of the t-statistics, and disappears in simulations when the estimate is replaced by the true variance. In small, heavily unbalanced samples, one should be aware that this bias could be relatively quite large. We found that there is also a bias in the estimation of the variance of both the difference in means and the t-statistic in unbalanced designs whenever the two groups have unequal observed means.

As an illustration, consider the following very simple example. Let $n_1 = 2, n_2 = 50$ and suppose that the observations for variable j in population 1 are $(1, 3)$ while the observations in population 2 are a vector of zeros. It

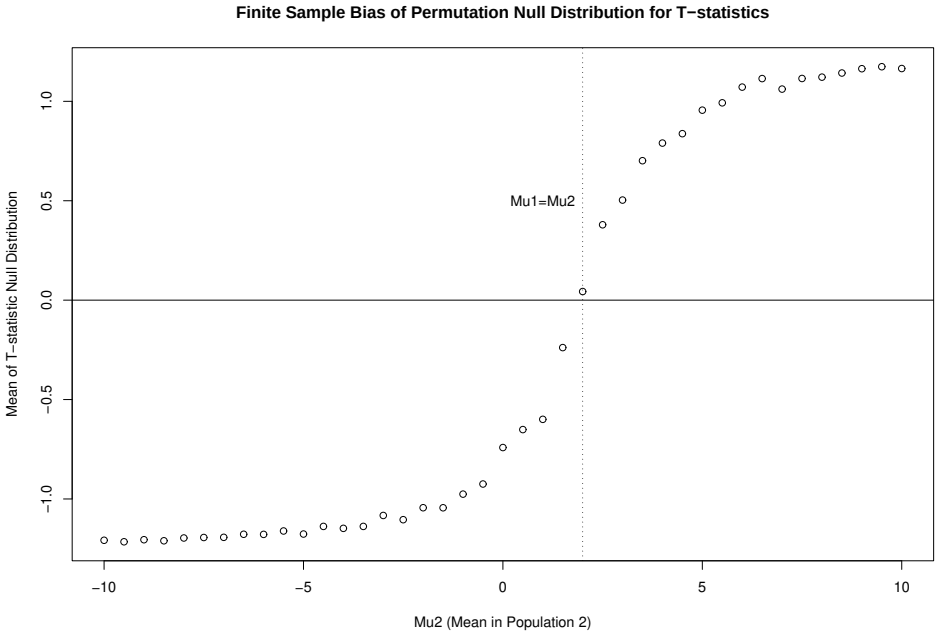

Figure 2: Mean of the permutation null distribution of the standardized two sample t-statistic for simulated data. Population 1 consists of $n_1 = 2$ subjects with observed values 1 and 3. Population 2 consists of $n_2 = 50$ subjects with observed values normally distributed with standard deviation 0.1 and different choices of mean. The mean of the null distribution is plotted versus the mean in Population 2 (*i.e.*: as a function of the difference in means since the mean in Population 1 is constant). The vertical line marks where the difference in means is truly zero. The mean of the null distribution is close to zero here, but increases in magnitude with the difference in means. The mean of the null distribution should be zero for all data sets. All 1326 possible permutations were performed exactly.

is easy to enumerate all of the possible permutations for this data set and compute the expected value of any test statistic under this null distribution

exactly. The results for the difference in means and the t-statistic are:

$$E(\mu_1 - \mu_2) = \frac{\binom{2}{2} * 2 + \binom{50}{1} * 0.44 + \binom{50}{1} * 1.48 - \binom{50}{2} * 0.08}{\binom{52}{2}} = 0$$

$$E(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}) = \frac{\binom{2}{2} * 2 + \binom{50}{1} * 0.87 + \binom{50}{1} * 0.99 - \binom{50}{2} * 1.27}{\binom{52}{2}} = -1.104$$

4.4 Simulations

We have conducted simulations to understand the performance of different multiple testing procedures for the two sample problem. In our evaluation of the different methods, we focus on estimation of the null distribution (e.g.: mean and variance of the test statistic under different choices of Q_{0n}), since accurately estimating Q_0 is essential if resulting inferences are to be correct. We also report estimates of the error control rates in Section 4.4.4, though we note that at most $I = 200$ data sets are used in each simulation so that the margin of error is almost as large as the level α that we are trying to estimate.

4.4.1 Data and Null Distributions

The following approach was used to generate simulated data sets. First, we simulate n_1 observations from a p-variate normal distribution with equal means $\mu_1 = 0$, equal variances $\sigma_1^2 = 0.1$, and all pairwise correlations $\rho_1 = 0$. Second, we simulate n_2 observations from a p-variate normal distribution with equal means $\mu_2 = 0$, equal variances $\sigma_2^2 = 5$ and all pairwise correlations $\rho_2 = 0.9$. The values of all parameters are chosen in light of the results from Section 4.3 as an extreme case of unbalanced groups in terms of sample size, variance, and correlation. We have examined different sample sizes and dimensions, but focus here on the results for $p = 100$ and several choices of n_1, n_2 representing unbalanced, nearly balanced and perfectly balanced designs. It would be an interesting area of future research to look at a wide range of covariance structures and sample sizes in order to try to understand the relative contributions of variance, correlation, and sample size to error control in finite samples. We know that the difference in covariance structures between the two populations will cause problems for the permutation method when $n_1 \neq n_2$, and our goal is to study the effect for several finite sample sizes (n_1, n_2).

For each simulated data set, we compute two test statistics: the difference in means D_n and the standardized t-statistic T_n. The null distributions

of these statistics are then estimated by (i) permutation-based $Q_n(P_{0n})$, (ii) the non-parametric bootstrap $Q_{0n}^{\#}$, and (iii) the parametric bootstrap-based $Q_n(P_{0n})$ (*i.e*: $P_{0n} = N(0, \Psi_n)$, where Ψ_n is the observed data covariance matrix). Notice that in (iii) we use the correct parametric distribution for the data. Equation (16) holds for the data generating distribution in the simulations, so we expect all three estimators to perform well asymptotically. The goal is to examine their finite sample performance. In each case, $B = 1000$ independent resampled data sets are used. Since we know the true distribution P in this simulation, we can compare parameters of the estimated null distributions to their true values.

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap	True Value
	mean (sd) over $I = 200$ data sets			
$n_1 = 5, n_2 = 6$				
$VAR(D_j)$	0.97 (0.40)	0.67 (0.28)	0.80 (0.48)	0.85
$VAR(T_j)$	1.21 (0.030)	3.26 (0.80)	1.56 (0.12)	1.62
$n_1 = 100, n_2 = 5$				
$VAR(D_j)$	0.071 (0.034)	0.84 (0.60)	1.038 (0.73)	1.001
$VAR(T_j)$	1.34 (0.18)	16.58 (21.08)	1.96 (0.21)	1.996
$n_1 = 200, n_2 = 10$				
$VAR(D_j)$	0.052 (0.030)	0.65 (0.50)	0.78 (0.64)	0.5005
$VAR(T_j)$	1.23 (0.18)	8.95 (13.69)	1.65 (0.49)	1.285
$n_1 = 19, n_2 = 20$				
$VAR(D_j)$	0.26 (0.075)	0.23 (0.070)	0.25 (0.074)	0.26
$VAR(T_j)$	1.05 (0.047)	1.14 (0.075)	1.11 (0.057)	1.11
$n_1 = 50, n_2 = 50$				
$VAR(D_j)$	0.101 (0.02)	0.100 (0.02)	0.102 (0.02)	0.102
$VAR(T_j)$	1.02 (0.04)	1.05 (0.05)	1.04 (0.05)	1.041

Table 3: Variance of the permutation, non-parametric bootstrap, and parametric bootstrap null distributions of the difference in means D_j and the t-statistic T_j. Since all variables have the same marginal distribution in this simulation, we report the results for one and note that they are representative for all variables. The true values are from formulas (approximate for the t-statistics, Moore and McCabe [2002]) and have been confirmed by simulation.

4.4.2 Choice of Test Statistic

We compare D_n and T_n based on the ease with which their null distributions can be estimated. For most models there are consistent finite sample estimators of the null distributions of both test statistics, although it is known that the null distribution of pivotal statistics (such as T_n) can be estimated with less asymptotic error than that of D_n in many cases (Hall [1992]). In our simulations, we observed the finite sample bias of the estimated null distributions of T_n noted in Section 4.3, while null distributions of both test statistics had observed means close to zero when the observed difference in means between the two samples was close to zero. The covariance structure of the test statistic null distributions was more difficult to estimate (See Table 3). In particular, the variance of T_n's null distribution is usually much too large with the non-parametric bootstrap estimator (resulting in conservative error rate control). In addition, whenever $n_1 \neq n_2$ the permutation estimates of the variance and correlation of the null distribution of D_n and the correlation (but not the variance) of the null distribution of T_n are far from the truth, as predicted by the formulas in Section 4.3. Thus, it is certainly interesting to do multiple testing with D_n in addition to T_n.

We suggest that D_n may be a better choice at small sample sizes and with non-parametric data generating models, whereas T_n is often preferable with larger sample sizes or more parametric models. In other words, pivoting (*i.e.*: dividing by $sd(\mu_n)$) only helps when the estimate $sd(\mu_n)$ is close to a constant (*e.g.*: asymptotically). How fast it becomes beneficial to pivot (as $n \to \infty$) is determined by the variance of $sd(\mu_n)$, which depends on (i) the data generating model (*i.e.*: model-based estimation versus non-parametric estimation) and (ii) the variance of the data.

4.4.3 Choice of Estimated Null Distribution

For both D_n and T_n, we compare the three choices of test statistic null distribution estimators. The comparison is based on the ability of each method to estimate the true null distribution and consequently to control error rates of interest. The most striking finding is that when $n_1 = n_2$, the permutation method performs very well even when the covariance structures are unbalanced, as predicted by the algebraic results in Section 4.3. Predictably, using a parametric bootstrap estimate of the data null distribution P_0 performs well when the model is correct, but quite poorly otherwise. The non-parametric bootstrap generally performs better for D_n than for T_n for two reasons. First, the bootstrap method estimates $sd(\mu_n)$

non-parametrically. Second, ties in the resampling can result in very small estimates $sd(\mu_n)$. Smoothing the empirical distribution does reduce this problem. Both of these factors contribute to the bootstrap method producing highly variable and unrealistically large resampled t-statistics. In contrast, the permutation-based test statistic (which uses a pooled estimate $sd(\mu_n)$) is much less variable, so that the asymptotic results of Hall [1992] will apply.

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap
$n_1 = 5, n_2 = 6$			
D_j	0.090	0.24	0.20
T_j	0.11	0.045	0.075
$n_1 = 5, n_2 = 100$			
D_j	0.67	0.15	0.12
T_j	0.095	0.0050	0.035
$n_1 = 10, n_2 = 200$			
D_j	0.77	0.12	0.10
T_j	0.085	0.015	0.025
$n_1 = 19, n_2 = 20$			
D_j	0.045	0.080	0.070
T_j	0.055	0.035	0.045
$n_1 = 50, n_2 = 50$			
D_j	0.080	0.085	0.090
T_j	0.080	0.065	0.065

Table 4: Estimates $\hat{\alpha}$ of the error rate $P(V_n > 10)$ over $I = 200$ independent data sets with $p = 100$ variables for the permutation, non-parametric bootstrap, and parametric bootstrap null distributions of D_n and T_n. We can expect the error in the estimates to be on the order of 0.05. The target error rate is $\alpha = 0.05$.

4.4.4 Error Rate Control

Since the two population mean vectors are equal, we know that any rejected null hypotheses are false positives, so we can estimate error rates. We report results from using Equation (13) with $k = 10$ to control $P(V_n > 10) \leq \alpha = 0.05$, where V_n is the number of false positives. Results for other error rates followed similar patterns. Table 4 shows the estimates of α over $I = 200$

independent data sets, where the thresholds are computed independently for each data set. A few interesting points emerge. First, conservative error control is associated with overestimating $VAR(T_j)$ (causing the upper quantiles c_j to be too large) and conversely, failure to control the error rate is due to under-estimation. Second, the direction of the bias in $\hat{VAR}(T_j)$ has consequences in terms of the size of the bias of $\hat{\alpha}$. In particular, the skewedness of type I error means that bias due to an underestimate of the variance is much larger in magnitude than the bias due to a similarly sized overestimate of the variance. Finally, the parametric and non-parametric bootstrap methods tend to be conservative for T_n and anti-conservative for D_n, whereas the permutation method tends to be anti-conservative for both statistics (but particularly for D_n).

We have also conducted simulations with some differences in means not truly zero. Estimated error rates tend to be slightly larger when there are some false null hypotheses. Also, the methods with the largest error rates have the most power. In practice, one might want to use a cost function that accounts for both type I and type II errors in order to optimize both the error rate and power.

5 Applications to Gene Expression Data Analysis

In this paper, we have focused on asymptotics for fixed dimension p and $n \to \infty$. Under these asymptotics, the usual central limit theorem applies, and $N(0, \Sigma(P))$ is the correct test statistic null distribution. In many applications, such as gene expression studies, however, the number of variables is typically always much larger than the number of samples. We present a few preliminary ideas on this topic. First, it is clear that some error rates should be harder to control than others because they depend on the most extreme gene(s) (*e.g.*: family-wise error). Second, parameters whose estimators have second order terms (*e.g.*: regression coefficients) will make error control harder than with sample means. Third, what we can say about the asymptotic distribution of the test statistics depends on the rate at which $p \to \infty$ relative to n.

When $p \gg n$, there is no multivariate central limit theorem. Hence, proving an approximation by a multivariate normal will only be possible with restrictive parametric assumptions on the observed data, though we rarely believe such a parametric model for the data in the gene expression context. We consider the example studied by van der Laan and Bryan [2001] and Pollard and van der Laan [2002], in which $\frac{n}{\log p} \to \infty$. Let (μ, Σ) denote

the mean and covariance of the data X. Then, if the minimum eigen value of Σ is bounded away from zero, van der Laan and Bryan [2001] have shown that when $\frac{n}{\log p} \to \infty$

1. $\max_{i,j} |\Sigma_{n,i,j} - \Sigma_{i,j}| \to 0$,
2. $\max_{i,j} |\Sigma^{-1}_{n,i,j} - \Sigma^{-1}_{i,j}| \to 0$.

This uniform consistency result is very different from a central limit theorem and does not guarantee that $\sqrt{n}(\mu_n - \mu) \xrightarrow[\frac{n}{\log p} \to \infty]{D} N(0, \Sigma)$. It does show us that when $X \sim N(\mu, \Sigma)$ one should control the error rate under the test statistic null distribution $N(0, \Sigma_n)$. Furthermore, in general, for $X \sim P$ one might reasonably choose to use one of the consistent estimators of $N(0, \Sigma)$ discussed in this paper as a null distribution for multiple testing. Note, however, that for any n there will typically be some genes whose marginal distribution is not yet normal (*i.e.*: the central limit theorem does not yet apply). It is a topic of future research to investigate the precise conditions under which the multivariate normal approximation $N(0, \Sigma(P))$ is valid.

5.1 Data Analysis

We apply resampling-based multiple testing methods to a publicly available data set (Alizadeh et al. [2000]). Expression levels of $13,412$ clones (relative to a pooled control) were measured in the blood samples of 40 diffuse large B-cell lymphoma (DLBCL) patients using cDNA arrays. According to Alizadeh et al. [2000], the patients belong to two molecularly distinct disease groups, 21 Activated and 19 Germinal Center (GC). We log the data (base 2), replace missing values with the mean for that gene, and truncate any expression ratio greater than 20-fold to $\log_2(20)$.

5.1.1 Testing for a Difference in Means

Our goal is to identify and then cluster clones with significantly different mean expression levels between the Activated and GC groups. We compute standardized t-statistics T_{jn} for each gene. We use permutation and non-parametric bootstrap methods to compute joint null distributions of the t-statistics. We choose to control the usual FWER ($k = 1$) and compare the clones identified as having significantly different means between the two

groups using: (i) Equation (13) common quantiles (for gene-specific thresholds) with the non-parametric bootstrap distribution, (ii) single-step Bonferroni common quantiles with the non-parametric bootstrap distribution, (iii) Equation (13) common quantiles with the permutation distribution, (iv) single-step Bonferroni common quantiles with the permutation distribution, and (v) Bonferroni adjusted common threshold with the tabled t-distribution for each marginal distribution.

Method	Null Distribution	Rejections
Equation (13) common quantiles	bootstrap	186
Bonferroni common quantiles	bootstrap	186
Equation (13) common quantiles	permutations	287
Bonferroni common quantiles	permutations	287
Bonferroni common threshold	t-distribution	32

Table 5: Number of rejected null hypotheses (out of $p = 13,412$) for five different choices of thresholds and null distribution. All 32 of the genes in the t-distribution subset are in both the permutation and the bootstrap subset, and the bootstrap and permutation subsets have 156 genes in common. Data are from Alizadeh et al. [2000].

Table 5 shows how many of the $p = 13,412$ null hypotheses are rejected using each method. Interestingly, Equation (13) and single-step Bonferroni common quantiles produce the same subset of clones (for both the bootstrap and the permutation null distributions), though this need not be the case since the single-step Bonferroni quantiles are always smaller. We see that the variances of the t-statistics across the $B = 1000$ samples tend to be smaller in the permutation distribution compared to the bootstrap distribution, resulting in the larger number of rejected null hypotheses with permutations. Based on the results of Section 4.4, we believe that the permutation subset is likely to be larger and the bootstrap subset to be smaller than the true subset. We believe that the permutation subset is likely to be closer to the true subset, since it makes use of a pooled variance estimate in T_n and $n_1 \approx n_2$.

We repeat this analysis using the difference in means D_n as the test statistic. For all of the resampling approaches, more clones are selected than with the t-statistics. This result confirms our observation in the simulations that D_n tends to be more anti-conservative than T_n. We also repeat the analysis with two random Activated patients removed so that the design

is perfectly balanced. Slightly fewer genes are significantly different between the two groups, but setting $n_1 = n_2 = 19$ did not change the results significantly.

5.1.2 Testing for an Association with Disease Group Using Logistic Regression

One might also be interested in testing for an association between gene expression and an outcome Y of interest, such as survival or disease group. In this case, a regression model $E(Y \mid X_j) = m(X_j \mid \beta_j)$ (*e.g.*: linear or logistic regression) is fit for every gene $j = 1, \ldots, p$, producing a vector of observed regression coefficients β_n which measure the association between gene expression and the outcome. The usual test statistics can be used (with $\mu_j = \beta_j$ as the parameter) to test the hypotheses $H_{0,j} : \beta_j = 0$, $j = 1, \ldots, p$ (or more generally, $H_{0,j} : \beta_j = \beta_j^0$). The bootstrap method of Section 2.7.2 can then be used to estimate the test statistic null distribution, using appropriate resampled random variables (*e.g.*: $Z_n^{\#} = \sqrt{n}(\beta_n^{\#} - \beta_n)$ for test statistics $\sqrt{n}(\beta_n - 0)$).

We apply the non-parametric bootstrap method to the data set of Alizadeh et al. [2000], with disease group (Activated versus GC) as a binary outcome and a logistic regression model. This is an example of a case that illustrates the simplicity of the bootstrap method. Despite the fact that (i) the outcome is not a linear function of gene expression and (ii) the error may not be independent of gene expression, the bootstrap can be applied directly without concern about the form of the test statistic distribution. In contrast, the usual resampling-based multiple testing methods (*e.g.*: permutations or resampling residuals as proposed by Westfall and Young [1993]) do not work, because the assumptions under which they are appropriate do not hold. Table 6 contains the number of genes that are significantly associated with disease group. The finding that the number of rejected null hypotheses is the same for $k = 1, 10, 50$ is partially due to the discreteness of the resampled null distribution (with $B = 1000$ resamples). By resampling more times (*e.g.*: $B = 10000$), a sharper bound can be achieved.

5.1.3 Clustering

We choose to use the subset of 186 clones selected with the bootstrap null distribution as having a significant difference in means for further analysis. Using the uncentered correlation (or cosine-angle) metric, we apply a hierarchical clustering algorithm called HOPACH (van der Laan and Pol-

$k =$	1	10	50	100	200
Rejections	303	303	303	471	553

Table 6: Logistic Regression Parameters. Number of rejected null hypotheses (out of $p = 13,412$) using the non-parametric bootstrap estimated null distribution and controlling the gFWER $P(V_n > k)$ for different choices of k, where V_n is the number of false positives. The test statistics used are $\sqrt{n} * (\beta_n - 0)$. Fine-tuned common quantiles $\{c_j : j = 1, \ldots, p\}$ are computed using Equation (13) with the estimated null distribution in order to control the gFWER at level $\alpha = 0.05$. Data are from Alizadeh et al. [2000].

lard [2003]) to identify the main clusters of clones and order the clones in a sensible way. Figure 3 shows the clone-by-clone distance matrix ordered according to the final level of the HOPACH tree. The six main clusters identified in the first level of the tree are marked. One of these clusters has an expression profile that is significantly associated with survival time in a multiplicative intensity model and a cox proportional hazards model. Investigating the relationship between expression and survival in this data set is an area of future work.

5.1.4 Real Data Simulations

We conduct some additional simulations using 100 randomly selected genes from the data set of Alizadeh et al. [2000] centered to all have mean zero in the Activated and GC groups as the true data generating distribution. The idea is to make use of a real data set in order to (i) avoid assumptions about the parametric form of the underlying distribution and (ii) have a more realistic covariance structure between the genes. We treat the 21 Activated and 19 GC patients as the population and randomly sample $n_1 < 21$ Activated and $n_2 < 19$ GC patients from it to create an "observed" data set $I = 200$ times. We estimate the null distributions of the t-statistic and the difference in means, each resampling $B = 1000$ times. In each case, we use Equation (13) to control the gFWER $P(V > 10) \leq \alpha = 0.05$. We repeat the simulation for three choices of (n_1, n_2). Overall, the permutation distribution does the worst job and the non-parametric bootstrap the best job of controlling the error rate. Notice that the normal distribution parametric bootstrap is no longer the best method, since the data model is not normal.

We also repeat the simulation with ten genes whose means are non-zero in population 2 (as in Section 4.4). Error control rates are similar to those

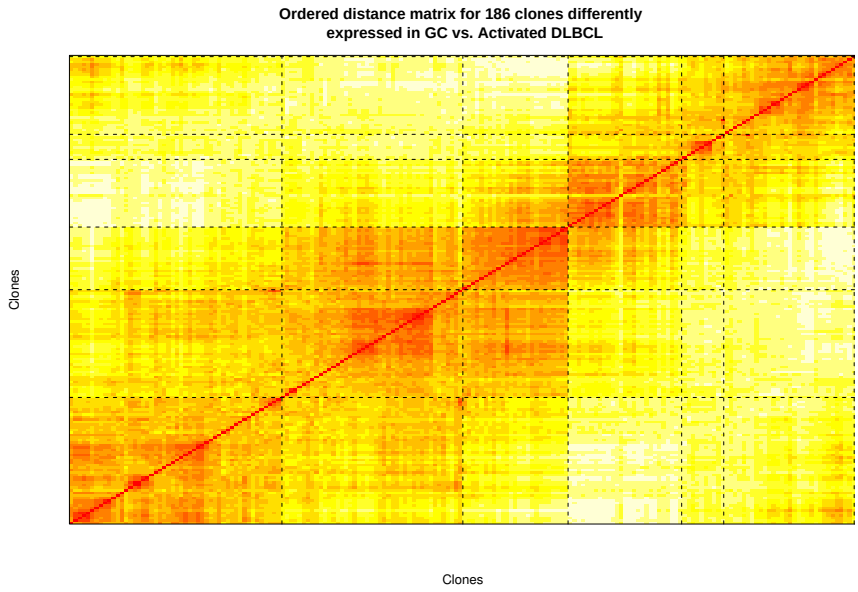

Figure 3: Uncentered correlation pairwise distance matrix of the 186 clones differently expressed in GC versus Activated DLBCL. The clones are ordered according to the final level of the HOPACH hierarchical tree. The dotted lines mark the boundaries between the six main clusters identified in the first level of the tree. Red corresponds with smallest and white with largest distance. Data are from Alizadeh et al. [2000].

in Table 7, and power is very high (at least 0.88 for all null distributions).

6 Discussion

Defining a formal statistical framework for hypothesis testing in multivariate settings has lead us to a better understanding of the correct null distribution for testing multiple hypotheses simultaneously. First, we have learned that for common choices of test statistics one should use a null distribution which is a projection of the true test statistic distribution on the space of mean zero distributions. Second, when the test statistics are based on asymptotically linear estimates μ_n of the parameter of interest $\mu(P)$, then the asymptotically correct test statistic null distribution is $N(0, \Sigma(P))$, where $\Sigma(P)$ is the covariance of the vector influence curve of μ_n. Third, our theorem shows

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap
$n_1 = 5, n_2 = 15$			
D_j	0.21	0.025	0.085
T_j	0.020	0.025	0.020
$n_1 = 9, n_2 = 11$			
D_j	0.13	0.050	0.065
T_j	0.015	0.065	0.015
$n_1 = 10, n_2 = 10$			
D_j	0.17	0.060	0.070
T_j	0.020	0.055	0.035

Table 7: Estimates $\hat{\alpha}$ of the error rate $Pr(V > 10)$ over $I = 200$ independent simulated data sets with $p = 100$ genes for permutation, non-parametric bootstrap and parametric bootstrap null distributions of D_j and T_j. In each case, Equation (13) was used to adjust for multiple tests. The target error rate is $\alpha = 0.05$

that under weak conditions, a class of estimators of the test statistic null distribution provides asymptotic control of most type I error rates for any data generating distribution P. A standard bootstrap method produces one such estimator. In particular, the bootstrap approach does not require the subset pivotality condition. Using a data null distribution P_0 to obtain a test statistic null distribution, in contrast, only provides asymptotic control when the subset pivotality condition of Westfall and Young [1993] holds, or according to our formal definition, when $\Sigma(P) = \Sigma(P_0)$.

In the context of testing for a difference in means in the two sample problem, we have illustrated that the commonly used method of estimating a test statistic null distribution $Q_n(P_{0n})$ via a permutation data null distribution P_{0n} indeed has the *correct* covariance if $\Sigma(P) = \Sigma(P_0)$ or, interestingly, if the design is balanced (*i.e.*: equal sample sizes in the two groups). It is a very powerful fact that whenever $n_1 = n_2$, the permutation method provides an estimated test statistic null distribution which is asymptotically correct and may in fact be more efficient for small sample sizes (by using pooled estimates of the covariance matrix). However, the permutation method suffers from a bias that depends on the observed difference in the means. In our limited simulation study, the standardized t-statistic T_n worked poorly compared to D_n when $sd(\mu_n)$ was variable (*e.g.*: non-parametric bootstrap

with a small sample size).

7 Acknowledgment

This research has been supported by a grant from the Life Sciences Informatics Program with industrial partner biotech company Chiron Corporation. We thank Sandrine Dudoit and Peter Westfall for the insightful discussions and helpful comments resulting in improvements of the manuscript.

References

A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, T.Jr. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenberger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stistical Society*, 57:289–300, 1995.

R. Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. Technical Report 110, Group in Biostatistics, University of California, Sept 2002. Submitted.

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, 1992.

Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, 1987.

E.L. Lehmann. *Testing Statistical Hypotheses*. Springer, 1986.

D.S. Moore and G.P. McCabe. *Introduction to the Practice of Statistics*. W.H. Freeman & Company, 2002.

K.S. Pollard and M.J. van der Laan. Statistical inference for simultaneous clustering of gene expression data. In D.D. Denison, M.H. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, pages 305–320. Springer-Verlag, 2002.

M.L. Puri and P.K. Sen. *Nonparametric Methods in Multivariate Analysis*. Wiley, 1971.

D.M. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8(6):557–569, 2001.

V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116–5121, 2001.

M.J. van der Laan and J.F. Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2:445–461, 2001.

M.J. van der Laan and K.S. Pollard. Hybrid clustering of gene expression data with visualization and the bootstrap. *Journal of Statistical Planning and Inference*, 117:275–303, 2003.

A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

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

D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196, 1999.

APPENDIX: Derivations of formulas in Section 4.3

The derivations of expressions for (i) the variances of D_j $(j = 1, 2)$ and (ii) the covariance of (D_1, D_2) are similar, and both make use of the double expectation theorem. For simplicity, assume that the null hypotheses hold for both variables, so that the means for the two populations are zero vectors $\mu_1 = \mu_2 = (0, 0)$. Consider variable j.

Recall the models $\mathcal{P}_1$ and $\mathcal{P}_2$ defined in Section 4.1. The distribution $P^b \in \mathcal{P}_1$ is defined by $X^b \mid L^b = 1 \sim P_1$ and $X^b \mid L^b = 2 \sim P_2$. The distribution $P^* \in \mathcal{P}_2$ is defined by $X^* \sim P_0$, $L^* \perp X^*$, and $P(L^* = 1) = 0.5$. Let D_j^b denote the test statistic based on n i.i.d. observations of $(X^b, L^b) \sim P^b \in \mathcal{P}_1$. Let D_j^* denote the test statistic based on n i.i.d. observations of $(X^*, L^*) \sim P^* \in \mathcal{P}_2$. Asymptotically, the distribution of D_j^* equals the distribution of the permutation test statistic. Our bootstrap estimate of the distribution of D_j (Section 2.7.2) converges to the distribution of D_j^b, while the permutation estimate of the distribution of D_j converges to the distribution of D_j^*.

The variance of the difference in means test statistic D_j under P^b is:

$$
\begin{aligned}
Var(D_j^b) &= E((D_j^b)^2) - E(D_j^b)^2 \\
&= E((D_j^b)^2) \\
&= E\left(\sum_{i=1}^{n} \frac{I(L_i^b = 2)(X_i^b)^2}{n_2^2} + \frac{I(L_i^b = 1)(X_i^b)^2}{n_1^2}\right) \\
&= nE\left(E\left(\frac{I(L^b = 2)(X^b)^2}{n_2^2} + \frac{I(L^b = 1)(X^b)^2}{n_1^2} | L^b\right)\right) \\
&= nE\left(E\left(\frac{I(L^b = 2)(X^b)^2}{n_2^2} + \frac{I(L^b = 1)(X^b)^2}{n_1^2} | L^b = 1\right) * P(L^b = 1)\right) \\
&\quad + nE\left(E\left(\frac{I(L^b = 2)(X^b)^2}{n_2^2} + \frac{I(L^b = 1)(X^b)^2}{n_1^2} | L^b = 2\right) * P(X^b = 2)\right) \\
&= n\left(\frac{\sigma_{1,j}^2}{n_1^2}\frac{n_1}{n} + \frac{\sigma_{2,j}^2}{n_2^2}\frac{n_2}{n}\right) \\
&= \frac{\sigma_{1,j}^2}{n_1} + \frac{\sigma_{2,j}^2}{n_2}.
\end{aligned}
$$

Similarly, the variance of the test statistic D_j under P^* is:

$$\begin{aligned}
Var(D_j^*) &= E((D_j^*)^2) - E(D_j^*)^2 \\
&= E((D_j^*)^2) \\
&= E\left(\sum_{i=1}^{n} \frac{I(L_i^*=2)(X_i^*)^2}{n_2^2} + \frac{I(L_i^*=1)(X_i^*)^2}{n_1^2}\right) \\
&= nE\left(E\left(\frac{I(L^*=2)(X^*)^2}{n_2^2} + \frac{I(L^*=1)(X^*)^2}{n_1^2}|L^*\right)\right) \\
&= nE\left(E\left(\frac{I(L^*=2)(X^*)^2}{n_2^2} + \frac{I(L^*=1)(X^*)^2}{n_1^2}|L^*=1\right) * P(L^*=1)\right) \\
&\quad +nE\left(E\left(\frac{I(L^*=2)(X^*)^2}{n_2^2} + \frac{I(L^*=1)(X^*)^2}{n_1^2}|L^*=2\right) * P(L^*=2)\right) \\
&= n\left(\frac{1/n(\sigma_{1,j}^2 n_1 + \sigma_{2,j}^2 n_2)}{n_1^2}\frac{n_1}{n} + \frac{1/n(\sigma_{1,j}^2 n_1 + \sigma_{2,j}^2 n_2)}{n_2^2}\frac{n_2}{n}\right) \\
&= \frac{\sigma_{1,j}^2}{n_2} + \frac{\sigma_{2,j}^2}{n_1}.
\end{aligned}$$

Note that in this derivation, the variance of X^* is $\frac{1}{n}(\sigma_{1,j}^2 n_1 + \sigma_{2,j}^2 n_2)$ for both values of L^*, since X^* is independent of L^*. It is interesting to note that the final expression for the variance of D_j^* resembles that of the variance of D_j, except with the roles of n_1 and n_2 reversed.

Now, consider the covariance between the test statistics for the two genes. Under P^b we have:

$$\begin{aligned}
Cov(D_1^b, D_2^b) &= E(D_1^b * D_2^b) \\
&= E\left(\left(\sum_{i=1}^{n} \frac{I(L_i^b = 2)X_{1,i}^b}{n_2^2} - \frac{I(L_i^b = 1)X_{1,i}^b}{n_1^2}\right)\right. \\
&\quad \left. * \left(\sum_{i=1}^{n} \frac{I(L_i^b = 2)X_{2,i}^b}{n_2^2} - \frac{I(L_i^b = 1)X_{2,i}^b}{n_1^2}\right)\right) \\
&= E\left(\sum_{i=1}^{n}\left(\frac{I(L_i^b = 2)X_{1,i}^b}{n_2^2} - \frac{I(L_i^b = 1)X_{1,i}^b}{n_1^2}\right)\right. \\
&\quad \left. * \left(\frac{I(L_i^b = 2)X_{2,i}^b}{n_2^2} - \frac{I(L_i^b = 1)X_{2,i}^b}{n_1^2}\right)\right) \\
&= nE\left(\left(\frac{I(L^b = 2)X_1^b}{n_2^2} - \frac{I(L^b = 1)X_1^b}{n_1^2}\right)\right. \\
&\quad \left. * \left(\frac{I(L^b = 2)X_2^b}{n_2^2} - \frac{I(L^b = 1)X_2^b}{n_1^2}\right)\right) \\
&= nE\left(\frac{I(L^b = 1)X_1^b X_2^b}{n_1^2} + \frac{I(L^b = 2)X_1^b X_2^b}{n_2^2}\right) \\
&= nE\left(E\left(\frac{X_1^b X_2^b}{n_1^2}|L^b = 1\right) * P(L^b = 1)\right. \\
&\quad \left. + E\left(\frac{X_1^b X_2^b}{n_2^2}|L^b = 2\right) * P(L^b = 2)\right) \\
&= n\left(\frac{\phi_1}{n_1^2}\frac{n_1}{n} + \frac{\phi_2}{n_2^2}\frac{n_2}{n}\right) \\
&= \frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}.
\end{aligned}$$

Under P^* we have:

$$\begin{aligned}
Cov(D_1^*, D_2^*) &= E(D_1^* * D_2^*) \\
&= E\left(\left(\sum_{i=1}^{n} \frac{I(L_i^*=2)X_{1,i}^*}{n_2^2} - \frac{I(L_i^*=1)X_{1,i}^*}{n_1^2}\right)\right. \\
&\quad \left. * \left(\sum_{i=1}^{n} \frac{I(L_i^*=2)X_{2,i}^*}{n_2^2} - \frac{I(L_i^*=1)X_{2,i}^*}{n_1^2}\right)\right) \\
&= E\left(\sum_{i=1}^{n}\left(\frac{I(L_i^*=2)X_{1,i}^*}{n_2^2} - \frac{I(L_i^*=1)X_{1,i}^*}{n_1^2}\right)\right. \\
&\quad \left. * \left(\frac{I(L_i^*=2)X_{2,i}^*}{n_2^2} - \frac{I(L_i^*=1)X_{2,i}^*}{n_1^2}\right)\right) \\
&= nE\left(\left(\frac{I(L^*=2)X_1^*}{n_2^2} - \frac{I(L^*=1)X_1^*}{n_1^2}\right)\right. \\
&\quad \left. * \left(\frac{I(L^*=2)X_2^*}{n_2^2} - \frac{I(L^*=1)X_2^*}{n_1^2}\right)\right) \\
&= nE\left(\frac{I(L^*=1)X_1^*X_2^*}{n_1^2} + \frac{I(L^*=2)X_1^*X_2^*}{n_2^2}\right) \\
&= nE\left(E\left(\frac{X_1^*X_2^*}{n_1^2}|L^*=1\right) * P(L^*=1)\right. \\
&\quad \left. + E\left(\frac{X_1^*X_2^*}{n_2^2}|L^*=2\right) * P(L^*=2)\right) \\
&= n\left(\frac{1/n(\phi_1 n_1 + \phi_2 n_2)}{n_1^2}\frac{n_1}{n} + \frac{1/n(\phi_1 n_1 + \phi_2 n_2)}{n_2^2}\frac{n_2}{n}\right) \\
&= \frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}.
\end{aligned}$$

Note that in the permutation derivation, the covariance of X_1^* and X_2^* is $\frac{1}{n}(\phi_1 n_1 + \phi_2 n_2)$ for both values of L^*, since Z^* is independent of L^*. Again, it is interesting to note that the final expression for the covariance under P^* resembles that under P^b, except with the roles of n_1 and n_2 reversed.