

1 Motivation

Microarray and gene chip technologies allow researchers to monitor the expression of thousands of genes simultaneously. These gene expression studies are swiftly becoming very popular in biomedical research. Each experiment results in an observed data matrix X whose columns are n copies of a p -dimensional vector of gene expression measurements, where n is the number of observations and p is the number of variables (e.g: genes). Many interesting research questions have arisen from the fact that the dimension of the data far exceeds the sample size (typically $p > 5000$ and $n < 50$). For example, researchers frequently want to perform a statistical test for every gene in order to answer questions such as “Which genes are significantly differently expressed between two (or more) conditions?” or “Which genes have a significant association with an outcome or covariate?”. Ordering the genes based on statistical significance helps one to formulate biological hypotheses and allows one to prioritize the genes for follow-up experiments, such as drug target validation or *in situ* studies. In addition, identifying such statistically significant subsets of genes reduces the dimension of the data before further analysis, such as clustering or classification.

In order to make statements about the statistical significance of thousands of genes simultaneously, however, it is essential to appropriately account for multiple tests. Multiple hypothesis testing is a well studied problem, with applications in many fields. A number of the existing multiple testing procedures have been applied to gene expression data, but often the procedures are used incorrectly or the assumptions under which they work are violated. In this paper, we discuss the problem of multiple testing for gene expression data in a statistical framework, thereby illustrating which methods are suitable for this context.

2 Background

Given a family of hypotheses to test, a multiple testing procedure typically consists of (i) selecting the test statistics, (ii) choosing the desired type of error control, (iii) identifying and perhaps computing the appropriate null distribution of the test statistics, and (iv) making testing decisions given the observed test statistics and their null distribution. Traditional approaches to the multiplicity problem are reviewed by Hochberg and Tamhane (1987).

More recent developments in the field include resampling methods (Westfall and Young (1993)), step-wise procedures, and less conservative error rate control, such as control of the false discovery rate (Benjamini and Hochberg (2002)). Dudoit et al. (2002) discuss multiple testing in gene expression experiments, conduct a comparison study, and review some recent applications of multiple testing to expression data, noting where the methods have been used without proper consideration for the statistical framework of multiple testing. Multiple testing procedures in this context must account for

1. the dimension being much larger than the sample size,
2. the correlation structure between the genes,
3. the fact that genes do not share a common marginal null distribution,
4. the knowledge that some proportion of the genes should have significant affects (e.g.: different expression between groups or associations between expression and outcomes).

We wish to emphasize and elaborate upon these points.

The sheer number of genes in an expression study creates a multiplicity problem of a much larger scale than in more classical multiple testing situations (with typically only tens of tests). Because the number of tests is in the thousands for expression studies, computational issues are extremely important, particularly when resampling methods are used. This is an area that requires further research. Another implication of the large number of genes is that the researcher must carefully choose a type of error control which accurately accounts for the multiplicity. A precise definition of the type of error being controlled is an essential (and all too often ignored) component of any testing procedure.

In addition to the number of genes, several other characteristics of the gene expression distribution have implications for multiple testing. First, we generally expect some proportion of the genes (frequently 10-30%) to have significant effects (eg: differential expression or association with another variable). Hence, it is important to have *strong* control of the error rate (ie: control under any combination of true and false null hypotheses). This issue will be discussed further below. Second, we also expect that many of the genes' expression patterns will be correlated, because sets of genes involved in the same biological process are co-regulated. Therefore, it is essential to

consider the *joint* distribution of the gene’s test statistics when performing multiple testing. Third, we have observed that the marginal null distribution of a test statistic for each gene is not the same for every gene, even when standardized test statistics are used. In other words, for the sample sizes that we generally have in gene expression studies, we can not expect that the central limit theorem is applicable, so that the use of asymptotic (tabled) distributions is not appropriate. Consequently, it also no longer makes sense to assume a common threshold for testing all gene’s null hypotheses.

The goal of this paper is to investigate these issues further and to suggest some multiple testing procedures that are appropriate for use with gene expression data. We begin by describing the data and defining the multiple testing problem (Section 3) in a concise, but general, way. In Section 4, we outline the different components of multiple testing procedures within a clear, statistical framework and discuss the considerations that must be made when performing multiple testing with gene expression data. Section 5 discusses the specific example of the two sample multiple testing problem and compares our proposed null distributions with the permutation distribution algebraically (Sections 5.3 and 5.4) and in simulations (Section 5.5). Section 6 illustrates how multiple testing can be applied to real data. We conclude with a brief discussion (Section 7).

3 Data and Null Hypotheses

Let X_1, \dots, X_n be i.i.d. $X \sim P \in \mathcal{M}$, where \mathcal{M} is a model and X is a p -dimensional vector that may include gene expression, and possibly covariates and outcomes. One outcome of interest is survival, which may be subject to right censoring. Consider real valued parameters μ_1, \dots, μ_p , that is $\mu_j(P) \in \mathfrak{R}$. The gene-specific parameters μ_1, \dots, μ_p could be, for example, location parameters (e.g.: means/medians or differences between two population means/medians) or regression parameters (e.g.: association between \mathbf{x}_j and Y in a linear/logistic model). Typically, we are interested in simultaneously testing the null hypotheses:

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

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

4 Multiple Testing Procedures

Given i.i.d. X_1, \dots, X_n and a family of null hypotheses $\{H_{0,j}, j = 1, \dots, p\}$, we define a large class of multiple testing procedures (excluding FDR controlling procedures) by specific choices of test statistics, type of error control, null distribution, and testing method. We discuss each of these components and make some remarks about ensuring that an multiple testing procedure is indeed testing precisely the null hypotheses of interest.

4.1 Test Statistics

Let $\hat{\mu}_j$ be an efficient estimate or locally efficient estimate of μ_j . We refer to van der Laan and Robins (2002) for locally efficient estimators dealing with the curse of dimensionality. Then, an obvious choice of test statistics is:

$$T_j = \hat{\mu}_j - \mu_j^0, j = 1, \dots, p. \quad (2)$$

A statistic T_j that is sufficiently large in absolute value (or in a certain direction for a one-sided test) represents significant evidence against the null hypothesis $H_{0,j}$. Note that it is important to efficiently estimate the parameters of interest μ_j , since doing so provides test statistics T_j with maximal power.

Other choices of test statistics include $T_j = \sqrt{n}(\hat{\mu}_j - \mu_j^0)$ and $T_j = (\hat{\mu}_j - \mu_j^0)/sd(\hat{\mu}_j)$. If $\hat{\mu}_j$ is asymptotically linear with influence curve $IC_j(X)$, that is, $\hat{\mu}_j - \mu_j = \frac{1}{n} \sum_{i=1}^n IC_j(X_i) + op(\frac{1}{\sqrt{n}})$, and $sd(\hat{\mu}_j)$ is an estimate of $\sigma_j = \sqrt{VAR(IC_j(X))/n}$, then

$$T_j = \frac{\hat{\mu}_j - \mu_j^0}{sd(\hat{\mu}_j)} \xrightarrow[n \rightarrow \infty]{D} N(0, 1).$$

Standardizing test statistics so that the asymptotic marginal distributions of all T_j are $N(0, 1)$ is a useful tool when one wishes to use tabled null distributions. As we discuss in Section 4.4.1, marginal gene expression null distributions are far from being equal to $N(0, 1)$, even for standardized test statistics and reasonably large sample sizes. In particular, estimation of $sd(\hat{\mu}_j)$ is known to be difficult in the gene expression context (Tusher et al. (2001), Rocke and Durbin (2001)). Thus, we do not recommend using tabled distributions. This also eliminates the need to use standardized test statistics. We revisit this issue in the simulations of Section 5.5, where we compare the

statistics in Equation 2 to their standardized counterparts and show that it is easier to estimate the null distribution of non-standardized statistics (See Table 3).

4.2 Error Control

A multiple testing procedure will make testing decisions about all p null hypotheses $H_{0,j}$. In truth, some of the null hypotheses may be true and some may be false. We wish to assess the procedure based on an estimate of how many erroneous rejections it makes.

4.2.1 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) discuss and compare different generalizations of type I error control to the multiple testing setting. Let R be the total number of rejected hypotheses, let V be the (unobservable) number of false rejections, and let k be a user supplied constant. Some error rates include:

- PCER = $E(V)/p$: per-comparison error rate,
- PFER = $E(V)$: per-family error rate,
- FWER = $Pr(V \geq k)$: family-wise error rate,
- FDR = $\begin{cases} E(V/R) & R \geq 0 \\ 0 & R = 0 \end{cases}$: false discovery rate.

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 (Benjamini and Hochberg (2002)) is becoming a popular choice of error rate.

4.2.2 Strong 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, so we have to choose a way to compute the expectations and/or probabilities in the error rate. Weak control means that the error rate is controlled under a null distribution Q_0 satisfying the complete null hypothesis $H_0^C = \bigcap_{j=1}^p H_{0,j}$, i.e.: only if all hypotheses are true. In gene expression studies, however, it is usually expected that some of the hypotheses are false, so that weak control is not sufficient. Strong control means that the error rate is controlled under any combination of true and false hypotheses, so that in particular it is controlled under the true distribution P . Since control of the error rate under P is our goal, we would like a testing procedure to have strong control (at least asymptotically). While none of the multiple testing procedures in the literature in fact guarantees strong control for finite samples (and large p) without extreme parametric assumptions, we recommend using procedures which at least aim to have strong control as $n \rightarrow \infty$. We discuss asymptotic strong control further in the following Section 4.2.3.

The FDR method of Benjamini and Hochberg (2002) takes a different approach. This method does not have level α when the complete null H_0^C is true, but it does control $E(V/R)$ under the truth given certain assumptions about the data generating distribution. The methods proposed here can not handle strong control of the FDR. We note, however, that weak control of the FDR under $P_0 \in \mathcal{M}_0$ (as defined below in Section 4.3.1) is equivalent to weak control of the FWER.

4.2.3 Asymptotic Strong Control

Let α_n denote the error rate for a sample of size n and consider a target error rate α . Asymptotic strong control can then be defined as $\alpha_n \rightarrow \alpha$ as $n \rightarrow \infty$. In order to guarantee asymptotic strong control, we need convergence in distribution of the test statistics. There are a few situations in which this is the case. First, if the dimension of the data p were finite, then the usual central limit theorem would apply and the asymptotic distribution of the standardized test statistics (as $n \rightarrow \infty$) would be multivariate normal. In gene expression studies, however, the number of genes grows with the number of samples. Consider now the situation where $p > n$ so that $p = p(n) \rightarrow \infty$

faster than n . Under a parametric model for the observed data, one might be able to prove that the distribution of the test statistics is arbitrarily well approximated by a multivariate normal distribution for $n \rightarrow \infty$. Note that when p is much larger than n there is no multivariate central limit theorem, so that proving an approximation by a multivariate normal will only be possible with restrictive parametric assumptions on the observed data. In the gene expression context, however, we rarely believe such a parametric model. In practice, for any n there will typically be some genes for which the marginal distribution is not yet normal. Thus, without the existence of a multivariate normal approximation or limit distribution it is very doubtful that multiple testing procedures will have asymptotic strong control, but this is an area of future research.

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 $p \rightarrow \infty$ relative to n . Consider the following example studied by van der Laan and Bryan (2001), in which $\frac{n}{\log p} \rightarrow \infty$. Suppose the parameters of interest are the sample means and the test statistics are $T_j = \sqrt{n}(\hat{\mu} - \mu^0)$, $j = 1, \dots, p$. Let Σ denote the covariance matrix of X and let Σ_n be the empirical covariance matrix. Then, if the minimum eigen value of Σ is bounded away from zero, we have shown (van der Laan and Bryan (2001))

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

This does not, however, guarantee that $T \xrightarrow[\frac{n}{\log p} \rightarrow \infty]{D} N(0, \Sigma)$. If, in addition, one assumes $X \sim N(\mu, \Sigma)$, then $T \sim N(0, \Sigma)$, which suggests that in this parametric family one should control the error rate under $N(0, \Sigma_n)$. Using the results of van der Laan and Bryan (2001), one can now establish asymptotic strong control of such a procedure. This argument can be generalized to asymptotically linear test statistics $T = \{T_j : j = 1, \dots, p\}$ with influence curves $IC(X) = \{IC_j(X) : j = 1, \dots, p\}$ with negligible second order terms (Pollard and van der Laan (2001)).

Given these remarks, it seems unlikely that any multiple testing procedure can achieve asymptotic strong control in the gene expression context. Assuming a parametric model gives asymptotic strong control if the model is correct. But, in practice the model may be far from the truth (at least for some genes), so that one would do better using another choice of null distribution, such as an empirical null distribution. In the following sections, we propose multiple testing procedures which *aim* to have asymptotic strong control (of PCER, PFER, or PWER), that is they would achieve asymptotic strong control for finite p or $p \rightarrow \infty$ slow enough, and we compare these methods based on their finite sample performance.

4.3 Null Distribution

In order to decide if any of the observed test statistics are sufficiently unusual to reject the corresponding null hypotheses, we need to compare them to their distribution under the null (Equation 1). Since we are in a multiple testing setting, we need the *joint* null distribution of the test statistics.

4.3.1 Null Distribution as a Projection Parameter of P

In order to control the type I error, we need the null distribution of the data to resemble the true data generating distribution P except with the additional property that Equation 1 holds for all j . To understand what this means, first suppose that the complete null H_0^C is true. Then the true distribution P lies in the space $\mathcal{M}_0 = \{P \in \mathcal{M} : \mu_j(P) = \mu_j^0, j = 1, \dots, p\} \subset \mathcal{M}$ of all distributions in the model \mathcal{M} for which Equation 1 holds. For example, let P be parametrized by μ and variation independent nuisance parameters η . Since the nuisance parameters η are not constrained by the hypotheses $H_{0,j}$, then $\mathcal{M}_0 = \{P_{\mu,\eta} : \mu = \mu^0, \eta\}$. Now, in practice P probably does not lie in \mathcal{M}_0 , but we can not know if this is the case or not. Hence, we propose to use as null distribution a projection $P_0 = \Pi(P|\mathcal{M}_0)$ of P onto \mathcal{M}_0 using a particular distance metric $d(\cdot, \cdot)$.

This data null distribution P_0 will be as close to P as possible, and if H_0^C holds then $d(P, \mathcal{M}_0) \equiv \inf\{d(P, P_0) : P_0 \in \mathcal{M}_0\} = 0$. Suppose we use the Kullback-Leibler (K-L) distance. Then we have:

$$P_0 = P_0(P) = \arg \max_{P'_0 \in \mathcal{M}_0, P'_0 \ll \mu} \int \log \left(\frac{\partial P'_0(x)}{\partial \mu(x)} \right) dP(x). \quad (3)$$

Note this defines the optimal null distribution P_0 as a function of the true distribution P .

The data null distribution P_0 directly implies a joint null distribution Q_0 for the test statistics. For example, in a shift experiment where the parameter of interest is a location parameter, such as the mean or median, then for a non-parametric model and test statistics given by Equation 2, $Q_0 = Q(\cdot - \mu^0)$, where Q is the underlying distribution of the test statistics. When one uses a resampling method to estimate P_0 , then Q_0 is estimated by the distribution of the test statistics computed on each of the resampled data sets.

4.3.2 Estimation of Null Distribution

In practice, we do not know the true distribution P so that we need to estimate P_0 in order to estimate Q_0 . We can use the substitution estimator

$$P_0(P_n) = \arg \max_{P'_0 \in \mathcal{M}_0, P'_0 \ll \mu_n} \int \log \left(\frac{\partial P'_0(x)}{\partial \mu_n(x)} \right) dP_n(x), \quad (4)$$

which can be defined as a maximum likelihood estimate $\hat{P}_{0,MLE}$ of P_0 under the model \mathcal{M}_0 , where μ_n is a user supplied dominant measure. Then, different choices of the model \mathcal{M} for P define different estimators. For example, if \mathcal{M} is non-parametric and μ_j is a location parameter of \mathbf{x}_j , then $\hat{P}_{0,MLE} = (P_n - \mu^0)$. A disadvantage of this estimator is that for small n the marginal distributions are very discrete with many ties. If $\mathcal{M} = N(\mu, \Sigma)$, then $\hat{P}_{0,MLE} = N(\mu^0, \Sigma_n)$. This estimator requires the parametric assumption that the data generating distribution is multivariate normal. For both approaches, the null distribution of the test statistics Q_0 is estimated by generating a large number B of bootstrap data sets from the chosen distribution $\hat{P}_{0,MLE}$ and computing the test statistics T_j^b for $b = 1, \dots, B$. Then, the distribution of the test statistics T_j^b over the B bootstrap data sets estimates Q_0 .

One advantage of using an estimate $\hat{Q}_{0,MLE}$ of Q_0 (derived from the estimated projection $\hat{P}_{0,MLE}$) as the null distribution is that the true Q_0 would give strong control of the error rate. Hence, if p were finite then the proposed $\hat{Q}_{0,MLE}$ would give asymptotic strong control as $n \rightarrow \infty$. And in general (i.e.: $p \gg n$), $\hat{Q}_{0,MLE}$ at least aims to have strong control in the limit, and we can compare different estimates of Q_0 based on their finite sample properties. A directly related benefit of this approach is that any nuisance parameters η

are not allowed to asymptotically vary from their true values so that under asymptotics $\hat{\eta} \rightarrow \eta$ any rejection can be directly attributed to the parameter of interest (e.g.: $\mu_j \neq \mu_j^0$ for some j) and not to η . As a concrete example, suppose that the parameters of interest are the p population means for each gene. Then, we want to use a null distribution with the same covariance structure as the true distribution but with all means set equal to a null value. If we ignore (or incorrectly specify) the covariance structure, then we might declare that some gene's mean is significantly different from its hypothesized null value when in fact it is not.

4.4 Testing Method

Given observed test statistics T_j and an appropriate estimate of their joint null distribution Q_0 , a multiple testing procedure is a rule for making a rejection decision for each $H_{0,j}$ while controlling the chosen type I error rate.

4.4.1 Common Threshold vs. Common Quantiles

In traditional testing settings, it is often assumed that n is large enough that the central limit theorem applies and therefore standardized test statistics have a common marginal distribution. In this case, a common threshold c is used to make the testing decision for every variable, such as rejecting $H_{0,j}$ whenever $|T_j| > c$. For a specified level α , c could be the $1 - \alpha$ quantile (unadjusted for multiple tests) or the $1 - \alpha/p$ quantile (Bonferroni adjustment) of the shared marginal distribution. A generalization of the common threshold approach is to use a common quantile of the marginal distribution. If the marginal distributions are identical, then this corresponds with a common threshold.

In the gene expression context, however, n is generally small and the marginal distributions for each gene are not the same. Figure 1 shows the empirical bootstrap null distributions for the t-statistics of four genes from a real gene expression data set. Clearly, a single distribution should not be used for all genes and in particular, many of the gene-specific test statistics have a distribution far from $N(0, 1)$. Hence, we strongly recommend not using tabled distributions for testing with gene expression data. The resampling-based joint null distributions discussed in Section 4.3 are more appropriate. Furthermore, multiple testing procedures that assume a common threshold c for all genes may be problematic. Single-step $\max T$ and $\min p$ procedures

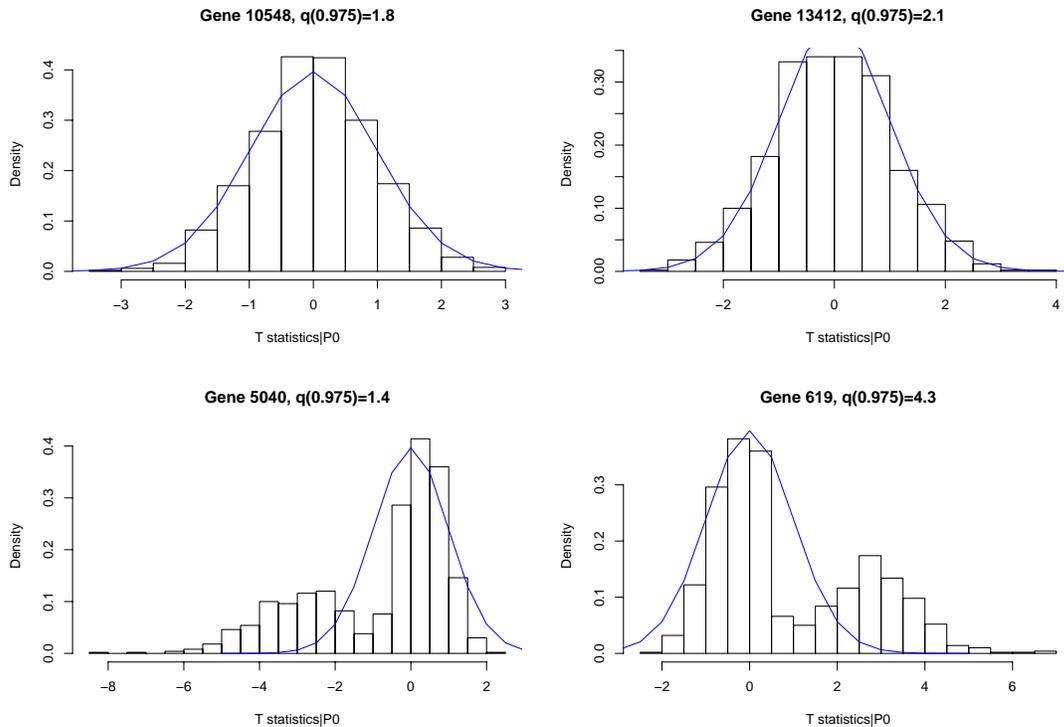


Figure 1: Histograms of null distributions of standardized t-statistics for four genes from the DLBCL data set of Alizadeh et al. (2000) computed by bootstrapping the centered empirical null distribution. The value of the 0.975 quantile of each distribution is given in the title. The 0.975 quantile of the T distribution is 2.0. 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.

(Westfall and Young (1993)), for example, do use the joint null distribution, but they correspond with employing a common threshold for all genes. We propose instead to employ testing procedures that use a common *quantile* for each gene, producing gene-specific thresholds c_j . The common quantile method can be further refined by using a step-down (or step-up) method to adjust the quantiles (See Table 1).

The general approach is to reject $H_{0,j}$ if T_j lies in the tails of its marginal null distribution $Q_{0,j}$, where these tails are defined by constants c_j (or pairs of constants $c_{l,j}, c_{u,j}$ for asymmetric tests) chosen to account for multiple tests.

Frequently, we are interested in performing symmetric, two-sided tests, so that we reject $H_{0,j}$ whenever $|T_j| > c_j$. In the case of a one-sided test, we only use a single tail of the null distribution.

4.4.2 Computing Thresholds

For any choice of error rate and level α , we set c_j equal to an appropriate quantile of an estimate of the marginal null distribution $Q_{0,j}$. Suppose that we use a resampling-based null distribution (as defined in Section 4.3.2). Then, we select the vector of thresholds so that the genes share a common tail probability a under this null distribution. For a two-sided test, c_j will be the $1 - a$ quantile of the absolute value of gene j 's marginal null distribution. The way that we choose a depends on the type I error rate we wish to control. For example, consider the vector of thresholds $\{c_j : j\}$ defined by

$$\frac{1}{B} \sum_{b=1}^B I \left\{ \sum_{j=1}^p I\{|T_j^b| > c_j\} > k \right\} < \alpha, \quad (5)$$

where B is the number of independent bootstrap data sets resampled and k is a pre-specified number of false positives. When $k = 1$ this is the usual family-wise error rate, and when $k > 1$ this controls the expected number of false positives $E(V) \leq k$ with probability $1 - \alpha$.

4.4.3 Comparison with P-value Adjusting Methods

An alternative approach to multiple testing is to compute a p-value (i.e.: the probability of observing a statistic as or more extreme than T_j), adjust it for multiple tests, and simply report this adjusted p-value. Adjusted p-values are defined as the level of the entire testing procedure at which $H_{0,j}$ would just be rejected, given all of the statistics T_j . Rejection decisions can later be made by comparing the adjusted p-values to the level α of the test for different choices of α . Westfall and Young (1993), Yekutieli and Benjamini (1999), and Dudoit et al. (2002) discuss methods for computing adjusted p-values.

P-values have the advantage that a testing decision can be made later for any choice of α without keeping track of the entire null distribution. Another nice application is that p-values can be used to order the genes, even when they do not have the same marginal distributions so that one can

not assume that an observed statistic T_j has the same probability for two different values of j . The trade off is that computing unadjusted p-values and then performing a multiple testing adjustment on them usually requires much more computation than simply choosing thresholds c_j to control a multiple testing type I error rate directly. By choosing the thresholds directly from a sufficiently smooth null distribution (i.e.: enough resampling has been done), a sharp bound can be achieved for any choice of error rate. In the gene expression setting, the goal of testing is usually to select a subset of interesting genes for further analysis (e.g.: clustering or classification). In this case, it makes sense to examine a few different subsets (i.e.: choices of α) up front, but to then make a testing decision and stick to it for the remainder of the analysis. For this purpose, we recommend using thresholds instead of p-values.

Any multiple testing procedure for adjusting p-values can also be stated as a method for choosing thresholds. Table 1 contains formulas for a based on some popular multiple testing p-value adjustments. If the $1 - a$ quantile for each gene j is chosen from the joint null distribution $\{|T_j^b| : b = 1, \dots, B, j = 1, \dots, p\}$, then these methods are equivalent to computing unadjusted p-values from the 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 its 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.

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

Table 1: Formulas for computing thresholds based on several methods for p-value adjustment. In each case, the threshold c_j is the $1 - a$ quantile of the null distribution of $|T_j^b|$, where a is determined by the formula. For step-down methods, the $\{r_j\}$ are the order statistics of $\{|T_j|\}$ and $(p - r_j + 1) = \text{rank}(|T_j|)$.

5 Example: Two Sample Problem

As a specific example, consider the two sample multiple testing problem, where we observe p variables (e.g.: gene expression measurements) on each subject. The conclusions drawn here can be extended to other testing problems.

5.1 Data and Null Hypotheses

Suppose 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 expression vector $X_{ij}, j = 1, \dots, p$ for subject i and $L_i \in \{1, 2\}$ is a label indicating subject i 's group membership. The two populations can have different gene expression distributions, so that $P(X_i|L_i)$ is one of two different distributions P_1, P_2 depending on the value of L_i . Let $\mu_{1,j}$ and $\mu_{2,j}$ denote the means of gene j in populations 1 and 2, respectively. Suppose we are interested in testing

$$H_{0,j} : \mu_j \equiv \mu_{2,j} - \mu_{1,j} = 0, j = 1, \dots, p. \quad (6)$$

We can select test statistics, an error control rate, and a resampling-based null distribution as described in the preceding Section 4.

5.2 Permutation Null Distribution

Another approach to the two sample problem, which has been applied frequently in the gene expression literature, is to use permutations of the group labels to compute the null distribution. The assumption behind this approach is that the data are identically distributed in both populations, which is often a stronger assumption than we wish to make. In order to understand the permutation distribution within the framework presented in this paper, we first consider the two true null distributions corresponding with permutations and bootstrap resampling and then discuss implications for the corresponding estimated distributions.

Consider a model \mathcal{M} for this two sample problem (e.g.: non-parametric P_1 for population 1 and P_2 for population 2). First, suppose we are testing the null hypothesis $H_0 : P_1 = P_2$. In this case, we have $\mathcal{M}_0^1 = \{P_1 = P_2\}$ and $P_0^1(P) = \Pi(P|\mathcal{M}_0^1)$ (e.g.: as defined in Equation 3). So, $P_0^1(P)$ can be estimated with permutation resampling. Specifically, the distribution of

the test statistics over many (possibly all, depending on the sample sizes n_1, n_2 and computing power) permutations of the group labels L_i provides a null distribution estimate $P_0^1(P_n)$. Second, suppose we are testing the null hypotheses $H_{0,j}$ given by Equation 6. Then, we have $\mathcal{M}_0^2 = \{\mu_1 = \mu_2\}$ and $P_0^2(P) = \Pi(P|\mathcal{M}_0^2)$. It is clear that permutation resampling is not appropriate for the estimation of P_0^2 , since $\mu_{1,j} = \mu_{2,j}, j = 1, \dots, p$ does not imply $P_1 = P_2$. Instead, the bootstrap estimates defined in Section 4.3.2 are more appropriate choices for $P_0^2(P_n)$.

Thus, the permutation null distribution can be quite problematic for testing the hypotheses $H_{0,j}$ given by Equation 6. Another way to view this problem is to notice that the permutation null distribution lies in \mathcal{M}_0^2 but may be quite far from the projection of P onto \mathcal{M}_0^2 . Consequently, a null hypothesis $H_{0,j}$ may be rejected because the distribution of T_j under permutations depends on some nuisance parameter which is quite different from its permutation null value (an ‘‘average’’ over the two groups), rather than because $\mu_{1,j} \neq \mu_{2,j}$. The following algebraic comparison of null distributions makes this distinction clear. We compare these estimation approaches further in Section 5.5.

5.3 Algebraic Comparison of Permutation and Bootstrap Null Distributions

For simplicity, we suppose that $p = 2$, but note that conclusions about the covariance of two genes can be applied to any pairwise covariance when p is much larger. For gene j , denote the mean and variance of X_i by $(\mu_{1,j}, \sigma_{1,j}^2)$ in population 1 and by $(\mu_{2,j}, \sigma_{2,j}^2)$ in population 2. Let ϕ_1 be the covariance between the two genes in population 1 and ϕ_2 be the covariance between the two genes in population 2. We are interested in testing $H_{0,j} : \mu_j = \mu_{2,j} - \mu_{1,j} = 0, j = 1, 2$.

Bootstrap resampling is equivalent to generating n_1 new subjects from an estimate of P_1 and n_2 new subjects from an estimate of P_2 . Let (X_i^b, L_i^b) denote the bootstrap data. Note that $P(X_i^b|L_i^b = 1)$ is \hat{P}_1 and similarly $P(X_i^b|L_i^b = 2)$ is \hat{P}_2 . Permutation resampling corresponds with randomly re-assigning the labels $\{L_i : i = 1, \dots, n\}$ to the subjects. Let (X_i^*, L_i^*) denote the permuted data. Now $X_i^* \perp L_i^*$ so that the permutation gene expression distribution $g(X_i^*|L_i^*)$ does not depend on L_i^* . We can approximate the permutation distribution $g(X_i^*) = g(X_i^*|L_i^*)$ by a mixture of the distributions

\hat{P}_1, \hat{P}_2 with mixing proportions $p_1 = \frac{n_1}{n}$ and $p_2 = \frac{n_2}{n} = 1 - p_1$, where n_1, n_2 are fixed.

Suppose we use test statistics defined by Equation 2. Now, we can write T_j as $\sum_{i=1}^n \frac{I(L_i=2)X_i}{n_2} - \frac{I(L_i=1)X_i}{n_1}$, where $I(L_i = k)$ is the indicator function that equals one when $L_i = k$. The test statistics are $T_j^b = \sum_{i=1}^n \frac{I(L_i^b=2)X_i^b}{n_2} - \frac{I(L_i^b=1)X_i^b}{n_1}$ and $T_j^* = \sum_{i=1}^n \frac{I(L_i^*=2)X_i^*}{n_2} - \frac{I(L_i^*=1)X_i^*}{n_1}$, under the bootstrap and permutation null distributions, respectively. We know that the expected values of T_j^b and T_j^* are both zero for $j = 1, 2$. In the appendix, we derive expressions for the variances of T_j^b and T_j^* ($j = 1, 2$) and the covariances of (T_1^b, T_2^b) and (T_1^*, T_2^*) . Here we present the results:

$$\begin{aligned} \text{Var}(T_j^b) &= \frac{\sigma_{1,j}^2}{n_1} + \frac{\sigma_{2,j}^2}{n_2} \\ \text{Var}(T_j^*) &= \frac{\sigma_{1,j}^2}{n_2} + \frac{\sigma_{2,j}^2}{n_1} \\ \text{Cov}(T_1^b, T_2^b) &= \frac{\phi_1}{n_1} + \frac{\phi_2}{n_2} \\ \text{Cov}(T_1^*, T_2^*) &= \frac{\phi_1}{n_2} + \frac{\phi_2}{n_1} \end{aligned}$$

It is interesting to note that the roles of n_1 and n_2 are reversed under permutations. These expressions show us that under most values of the underlying parameters, the bootstrap and permutation distributions of T_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 $\phi_1 = \phi_2 \equiv \phi$, 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 $Z^* \perp L^*$ implies that 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.

Suppose we were to use the usual standardized t-statistics. By dividing the difference in means by an estimate of its variance, we expect that $\text{Var}(T_j^b) = \text{Var}(T_j^*) = 1$. So, we solve the problem of different variance terms under permutations. The covariances $\text{Cov}(T_1^b, T_2^b)$ and $\text{Cov}(T_1^*, T_2^*)$, however, are still not equivalent unless $n_1 = n_2$ or the correlation structures are the same in the two populations.

The results of this section can be immediately generalized to other two sample multiple testing problems.

5.4 Bias of Permutation Null Distribution

We have found that the permutation null distribution of standardized t-statistics does not have mean zero whenever $n_1 \neq n_2$. When $n_1 = n_2$ this bias disappears. The bias is largest when the sample sizes are very different or at least one of the sample sizes is very small. As an illustration, consider the following simple example. Let $n_1 = 2, n_2 = 50$ and suppose that the observations for gene j in population 1 are $(1, 3)$ while the observations in population 2 are a vector of zeros. It 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. The results for the difference in means and the t-statistic are (rounded to two decimal places for presentation):

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

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

We have confirmed these formulas by simulation.

The consequences of this finding are serious. Whenever there are genes whose means are not equal in a data set with unequal population sample sizes, then the permutation null distribution of the t-statistic will not have mean zero and should not be used to assess whether two population means are equal. Thus, whenever there are any differentially expressed genes, the permutation null distribution of the t-statistic is biased for unbalanced sample sizes. Furthermore, the genes need not be differentially expressed in truth, since the bias follows from the means being unequal in the observed data. In the simulations below, we see that the size of this bias depends on the magnitude of the difference in means and 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 when the two groups have unequal means. A simple improvement upon the permutation approach would be to use the mixture distribution described in the previous section (Section 5.3), which is nearly equivalent to the permutation distribution but will not have this bias.

5.5 Simulations

We have conducted simulations to understand the performance of different multiple testing procedures, including choice of test statistics, null distribution and error control. Results of these investigations have motivated some of the comments made in the preceding sections. Here we report results from a comparison of bootstrap and permutation null distributions for comparing means in the two sample problem defined in Section 5.1. 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 $Q_{0,MLE}$), 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 5.5.4, 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.

5.5.1 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 5.3 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. For each simulated data set, we compute two test statistics: the difference in means and the standardized t-statistic. In this section, we let D_j denote the difference in means and T_j the t-statistic for gene j so that we can distinguish the two test statistics. The null distributions of these statistics are estimated by (i) permutations, (ii) the non-parametric bootstrap (centered empirical distribution), and (iii) the parametric bootstrap (multivariate normal distribution with equal means). 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.

5.5.2 Comparison of Test Statistics

We have suggested that using the difference in means rather than standardized t-statistics makes sense in the gene expression context where the marginal distributions of the t-statistics are not identical (See Figure 1) and estimation of $sd(\hat{\mu}_j)$ is difficult. In this section we compare the two choices of test statistic based on the ease with which their null distributions can be estimated for reasonable sample sizes. Since all genes have the same marginal distribution in this simulation, we report the results for one gene and note that they are representative for all genes.

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap
	mean (sd) over $I = 200$ data sets		
$n_1 = 5, n_2 = 100$			
$MEAN(D_j)$	4.8e-5 (0.0086)	-1.2e-3 (0.029)	-2.2e-3 (0.029)
$MEAN(T_j)$	1.4e-2 (0.23)	3.4e-2 (0.66)	-3.0e-3 (0.040)
$n_1 = 10, n_2 = 200$			
$MEAN(D_j)$	5.2e-4 (0.0073)	2.3e-3 (0.024)	-2.1e-3 (0.026)
$MEAN(T_j)$	6.0e-3 (0.20)	-2.3e-2 (0.47)	-2.2e-3 (0.039)
$n_1 = 50, n_2 = 50$			
$MEAN(D_j)$	1.4e-4 (0.0087)	1.7e-3 (0.022)	6.3e-4 (0.023)
$MEAN(T_j)$	6.7e-4 (0.17)	-1.7e-2 (0.56)	1.3e-3 (0.037)

Table 2: Mean of the permutation, non-parametric bootstrap, and parametric bootstrap null distributions of the difference in means D_j and the t-statistic T_j for one gene. The true value is zero for all sample sizes and both statistics.

Table 2 shows the average value of the mean of the null distributions of D_j and T_j for one gene with several sample sizes. The true means are both zero for all sample sizes. Both test statistics are unbiased with observed means close to zero, though the means for D_j tend to be smaller in absolute value and slightly less variable than those for T_j . Table 3 shows results for the variance of the same null distributions. The true variances for each sample size and statistic are given in the last column. We see that it is very difficult to estimate the variance of T_j 's null distribution with the non-parametric bootstrap. This is another argument in favor of using D_j with bootstrap null distributions. Note that the permutation distribution of $VAR(D_j)$ is

far from the truth, as predicted by the formulas in Section 5.3. We have also looked at estimates of the pair-wise covariances between test statistics under each null distribution. Both bootstrap null distributions estimate $COV(D_j, D_{j'}) (j \neq j')$ accurately, whereas the permutation distribution does not. The parametric bootstrap is the only distribution that accurately estimates $COV(T_j, T_{j'}) (j \neq j')$; the nonparametric bootstrap estimate is larger than the true value, whereas the permutation estimate is much smaller than the true value.

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap	True Value
	mean (sd) over $I = 200$ data sets			
$n_1 = 5, n_2 = 100$				
$VAR(D_j)$	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
$n_1 = 10, n_2 = 200$				
$VAR(D_j)$	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
$n_1 = 50, n_2 = 50$				
$VAR(D_j)$	0.085 (0.030)	0.45 (0.51)	0.53 (0.64)	0.102
$VAR(T_j)$	1.19 (0.20)	8.56 (17.26)	1.51 (0.41)	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 for one gene. The true values are from formulas (approximate for the t-statistics, Moore and McCable (2002)) and have been confirmed by simulation.

5.5.3 Comparison of Null Distributions

The results in the previous section allow us to compare the three choices of null distribution in terms of how easy they are to estimate and how close they are to the true null distribution.

- PERMUTATIONS: As expected, the permutation distribution estimates of $VAR(D_j)$, $COV(D_j, D_{j'})$ and $COV(T_j, T_{j'})$ are far from the true values and close to those given by the formulas in Section 5.3. The

permutation distribution estimates for $VAR(T_j)$, however, are much closer to the true values as we predict. When $n_1 = n_2$, permutations perform well.

- **NON-PARAMETRIC:** The centered empirical distribution is fairly close to the true Q_0 for D_j , but its estimate of the variance of T_j is quite variable and occasionally very far from the truth. This problem is a consequence of there being many ties in the resampling when one or both of the populations has a small sample size. These ties lead to very small variance estimates in denominators of the t-statistics, which produce unrealistically large bootstrap T_j^b . Smoothing the empirical distribution reduces this problem.
- **PARAMETRIC:** Overall, the parameters of the multivariate normal bootstrap null distribution are the closest to the true values and are quite accurate even in very unbalanced designs. Thus, we see that you gain immensely if you guess the model correctly and use a parametric bootstrap null distribution. In Section 6.3, however, we see that the multivariate normal is not necessarily a good choice of model with real gene expression data.

5.5.4 Error 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 5 to control $P(V > 10) \leq \alpha = 0.05$, where V 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. 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 underestimation. 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, as expected, the permutation null distribution does the worst job of controlling both error rates. Both the non-parametric and the parametric bootstrap methods perform fairly well, though they tend to be conservative for T_j and anti-conservative for D_j . When the simulation

is repeated with more similar covariance structures in the two populations, both methods control the error rate perfectly.

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap
$n_1 = 5, n_2 = 100$			
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
$n_1 = 50, n_2 = 50$			
D_j	0.425	0.155	0.135
T_j	0.13	0.04	0.06

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

5.5.5 Differentially Expressed Genes

We repeated the simulation giving ten of the genes in population 2 non-zero means (0.5, 1.0, 1.5, ..., 5.0). Most of the results are similar to the first simulation, but we note a few exceptions. First, we confirm the result of Section 5.4 that genes with non-zero differences in means do not have mean zero in the permutation null distribution of T_j when $n_1 \neq n_2$. Figure 2 shows the expected values of T_j under permutations across $I = 200$ independent data sets with $n_1 = 5, n_2 = 100$. We see that this bias increases with the value of the mean in population 2. Second, we also see a similar bias in the estimated variance of both T_j and D_j for genes with unequal means under permutations. Third, estimated error rates tend to be slightly larger when there are some false null hypotheses. Finally, the methods with the largest error rates have the most power. Table 5 shows the estimated power of each null distribution. 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.

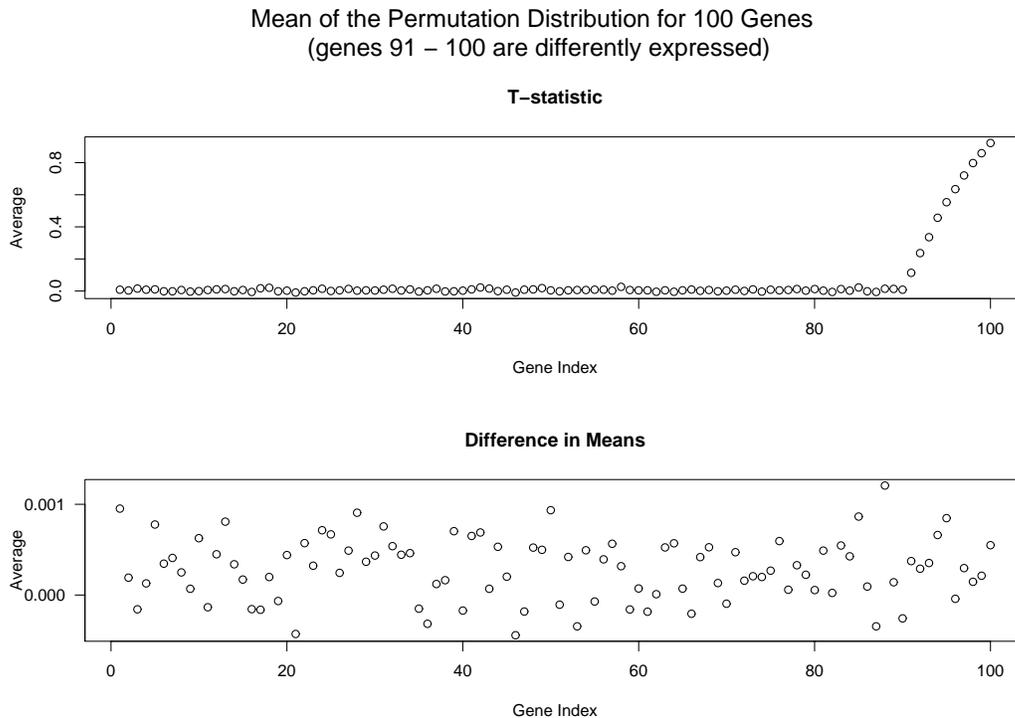


Figure 2: Mean of the permutation null distributions of the difference in means and the t-statistic for simulated data. Genes 91 to 100 have increasing non-zero means between 0.05 and 5 in population 2. The average value of the mean over $I = 200$ independent data sets is plotted for each gene. The mean of the null distribution should be zero for all genes.

6 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 $\log_2(20)$. Our goal is to identify and then cluster clones with significantly different mean expression levels between the Activated and

	Permutation	Non-parametric Bootstrap	Parametric Bootstrap
$n_1 = 5, n_2 = 100$			
D_j	0.87	0.69	0.64
T_j	0.60	0.18	0.39

Table 5: Estimates of the power when controlling the error rate $P(V > 10) < 0.05$ for the permutation, non-parametric bootstrap, and parametric bootstrap null distributions of D_j and T_j . $I = 200$ independent data sets are used, so that we can expect the error in the estimates to be on the order of 0.05.

GC groups.

6.1 Multiple Testing

We compute standardized t-statistics 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 family-wise error and compare the clones identified as having significantly different means between the two groups using: (i) Equation 5 common quantiles (for gene-specific thresholds) with the bootstrap distribution, (ii) single-step Bonferroni common quantiles with the bootstrap distribution, (iii) Equation 5 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.

Table 6 shows how many of the $p = 13,412$ null hypotheses are rejected using each method. Interestingly, Equation 5 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 5, we believe that the permutation subset is likely to be larger and the bootstrap subset to be slightly smaller than the true subset.

Method	Null Distribution	Rejections
Equation 5 common quantiles	bootstrap	186
Bonferroni common quantiles	bootstrap	186
Equation 5 common quantiles	permutations	287
Bonferroni common quantiles	permutations	287
Bonferroni common threshold	t-distribution	32

Table 6: 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).

We repeat this analysis using the (unstandardized) difference in means 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 the difference in means tends to be more anti-conservative than the t-statistic. 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.

6.2 Clustering

We choose to use the subset of 186 clones selected with the bootstrap null distribution for further analysis. Using the uncentered correlation (or cosine-angle) metric, we apply a hierarchical clustering algorithm called HOPACH (van der Laan and Pollard (2001)) 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.

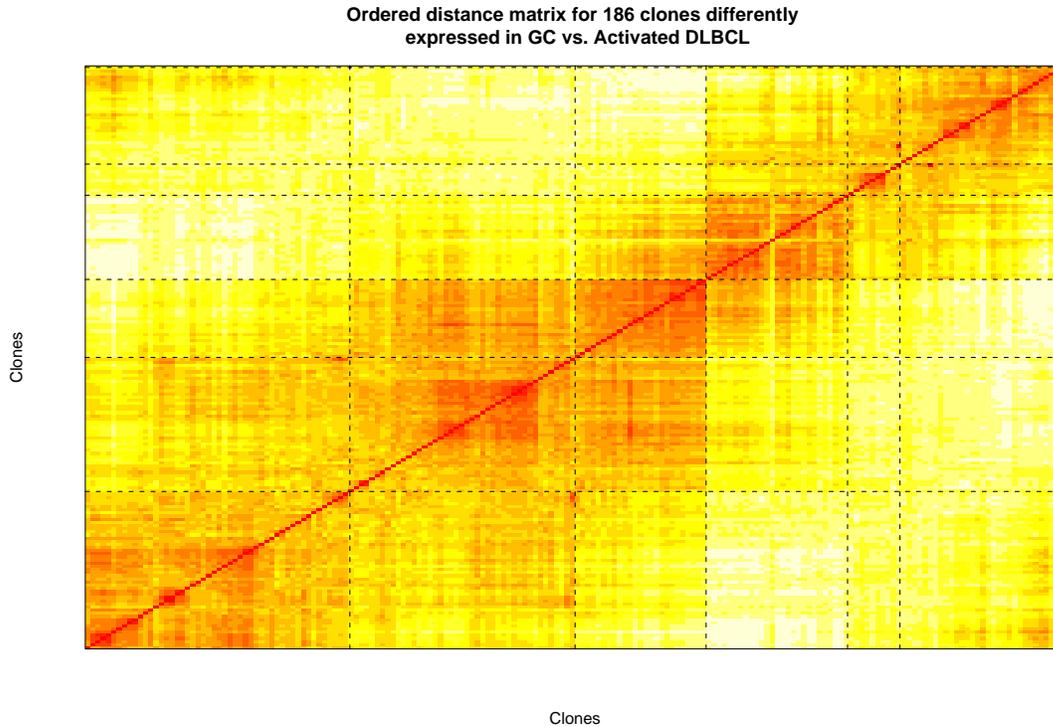


Figure 3: Uncentered correlation pairwise distance matrix of the 186 clones differently expressed in GC versus Activated DLBCL. The clones are ordered according to the final level of the HOPACH hierarchical tree. The dotted lines mark the boundaries between the six main clusters identified in the first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

6.3 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 Activated and n_2 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 5 to control the FWE $P(V > 10) \leq \alpha = 0.05$. 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 parametric bootstrap is no longer the best method, since the model is not normal.

	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 for permutation, non-parametric bootstrap and parametric bootstrap null distributions of D_j and T_j . In each case, Equation 5 was used to adjust for multiple tests. The target error rate is $\alpha = 0.05$

We also repeat the simulation with ten genes whose means are non-zero in population 2 (as in Section 5.5). Error control rates are similar to those in Table 7, and power is very high (at least 0.88 for all null distributions).

7 Discussion

In this paper, we describe multiple testing for high dimensional data sets, such as gene expression data, and propose an optimal data joint null distribution P_0 (defined as a projection parameter of the underlying distribution) for assessing the significance of observed statistics. We illustrate that resampling methods can be used to compute maximum likelihood estimates of P_0 . By choosing thresholds directly from a sufficiently smooth estimated

null distribution, a sharp bound can be achieved for any choice of error rate. Current computing power does, however, limit the number B of resampled data sets that one can practically use. Since we ran such a large number of simulations, it was necessary to use $B = 1000$, although we acknowledge that our results could have been more accurate with larger B . For example, all of the rejections in the data analysis had unadjusted p-values equal to zero. If the tails of the null distributions had more observations, some of these likely would have been non-zero and may even have been non-significant after multiple testing adjustment. In practice, one usually estimates the resampled distribution only once and B should be chosen as large as possible.

We examine the practice of using a common threshold for testing all null hypotheses. We note that a common threshold only makes sense if we believe that every gene has the same marginal distribution, so that the tail probability defined by the threshold is the same for all genes. The use of a resampling-based joint null distribution allows one to avoid assuming equality of marginal distributions (as is necessary with tabled distributions). We propose that one use a common quantile rather than a common threshold for each gene. In the case of step-down methods, a series of quantiles based on the ranks of the observed statistics is produced. Equivalently, one could compute p-values for all genes using the marginal distributions from the resampling-based joint null distribution and then adjust these p-values for multiple tests. We suggest that when one has the entire joint null distribution in hand, it makes more sense to simply control the desired error rate directly rather than compute p-values.

For the two sample multiple testing problem, we critique the popular approach of using a permutation null distribution to estimate P_0 and illustrate with algebraic examples, simulations, and a data analysis when this distribution is quite far from the desired P_0 . It is important to remark, however, that there are some situations, such as randomized clinical trials, where you believe in equality of the distributions of two or more groups. Also, when the groups have equal sample sizes in the two sample problem, the pooled variance and covariance estimates from the permutation approach are in fact better than the unpooled estimates from the bootstrap approach. In these cases, a permutation distribution may be appropriate. In general, however, one is interested in testing hypotheses about specific parameters (e.g.: equality of means) of two or more groups' distributions without believing that the entire distributions are identical. For these testing situations, we recommend using a bootstrap null distribution, and the non-parametric bootstrap is the

best choice when the parametric form of the underlying gene expression distribution is not known.

We also compare two different choices of test statistic, the difference in means and the standardized t-statistic. Since the permutation estimator is problematic with both statistics, we compare them based on the ease with which their null distributions can be approximated with bootstrap estimators. We observe (as have others in the literature) that t-statistics can have poor estimates of the standard deviations in the denominator so that their bootstrap estimated null distributions have inflated variances. This is particularly true of the non-parametric bootstrap, where ties can lead to very large t-statistics, resulting in conservative testing methods which might not make any rejections (but will control the error rate). We also note that there is no need to use standardized statistics when one is not assuming that all genes have a common null distribution. Hence, we suggest that the difference in means may be an appropriate choice of test statistic with bootstrap null distributions.

We conclude by commenting that the methods presented here have applications in other high dimensional contexts. In addition, these multiple testing methods can be directly extended to the equivalent problem of multi-dimensional confidence estimation.

References

- Alizadeh, A., M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, T. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenberger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, and L. Staudt (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.
- Benjamini, Y. and Y. Hochberg (2002). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stistical Society* 12, 7–29.
- Dudoit, S., J. Shaffer, and J. Boldrick (2002). Multiple hypothesis testing in microarray experiments. Technical Report 110, Group in Biostatistics, University of California. Submitted.

- Hochberg, Y. and A. Tamhane (1987). *Multiple Comparison Procedures*. John Wiley & Sons.
- Moore, D. and G. McCabe (2002). *Introduction to the Practice of Statistics*. W.H. Freeman & Company.
- Pollard, K. and M. van der Laan (2001, July). Statistical inference for two-way clustering of gene expression data. Technical Report 96, Group in Biostatistics, University of California. To appear in MSRI Nonlinear Estimation and Classification Workshop Proceedings.
- Rocke, D. and B. Durbin (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology* 8(6), 557–569.
- Tusher, V., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *PNAS* 98, 5116–5121.
- van der Laan, M. and J. Bryan (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* 2, 1–17.
- van der Laan, M. and K. Pollard (2001, May). Hybrid clustering of gene expression data with visualization and the bootstrap. Technical Report 93, Group in Biostatistics, University of California. To appear in JSPI.
- van der Laan, M. and J. Robins (2002). *Unified methods for censored longitudinal data and causality*. New York: Springer.
- Westfall, P. and S. Young (1993). *Resampling-based Multiple Testing: Examples and methods for p-value adjustment*. John Wiley & Sons.
- Yekutieli, D. and Y. Benjamini (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 82, 171–196.

APPENDIX: Details of proofs in Section 5.3

The derivations of expressions for (i) the variances of T_j^b and T_j^* ($j = 1, 2$) and (ii) the covariances of (T_1^b, T_2^b) and (T_1^*, T_2^*) are similar, and both make use of the double expectation theorem. For simplicity, assume that the null hypotheses hold for both genes, so that the means for the two populations are zero vectors $\mu_1 = \mu_2 = (0, 0)$. Consider gene j . The variance of the difference in means test statistic under bootstrap resampling is:

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

Similarly, the variance of the test statistic under permutations is:

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

Note that in the permutation derivation, the variance of Z^* is $1/n(\sigma_{1,j}^2 n_1 + \sigma_{2,j}^2 n_2)$ for both values of L^* , since $Z^* \perp L^*$. It is interesting to note that the final expression for the variance under permutations resembles that under bootstrap resampling, except with the roles of n_1 and n_2 reversed.

Now, consider the covariance between the test statistics for the two genes. Under bootstrap resampling we have:

$$\begin{aligned}
Cov(T_1^b, T_2^b) &= E(T_1^b * T_2^b) \\
&= E\left(\left(\sum_{i=1}^n \frac{I(L_i^b = 2)Z_{1,i}^b}{n_2^2} - \frac{I(L_i^b = 1)Z_{1,i}^b}{n_1^2}\right) * \left(\sum_{i=1}^n \frac{I(L_i^b = 2)Z_{2,i}^b}{n_2^2} - \frac{I(L_i^b = 1)Z_{2,i}^b}{n_1^2}\right)\right) \\
&= E\left(\sum_{i=1}^n \left(\frac{I(L_i^b = 2)Z_{1,i}^b}{n_2^2} - \frac{I(L_i^b = 1)Z_{1,i}^b}{n_1^2}\right) * \left(\frac{I(L_i^b = 2)Z_{2,i}^b}{n_2^2} - \frac{I(L_i^b = 1)Z_{2,i}^b}{n_1^2}\right)\right) \\
&= nE\left(\left(\frac{I(L^b = 2)Z_1^b}{n_2^2} - \frac{I(L^b = 1)Z_1^b}{n_1^2}\right) * \left(\frac{I(L^b = 2)Z_2^b}{n_2^2} - \frac{I(L^b = 1)Z_2^b}{n_1^2}\right)\right) \\
&= nE\left(\frac{I(L^b = 1)Z_1^b Z_2^b}{n_1^2} + \frac{I(L^b = 2)Z_1^b Z_2^b}{n_2^2}\right) \\
&= nE\left(E\left(\frac{Z_1^b Z_2^b}{n_1^2} | L^b = 1\right) * P(L^b = 1) + E\left(\frac{Z_1^b Z_2^b}{n_2^2} | L^b = 2\right) * P(L^b = 2)\right) \\
&= n\left(\frac{\phi_1}{n_1^2} \frac{n_1}{n} + \frac{\phi_2}{n_2^2} \frac{n_2}{n}\right) \\
&= \frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}.
\end{aligned}$$

Under permutations we have:

$$\begin{aligned}
Cov(T_1^*, T_2^*) &= E(T_1^* * T_2^*) \\
&= E\left(\left(\sum_{i=1}^n \frac{I(L_i^* = 2)Z_{1,i}^*}{n_2^2} - \frac{I(L_i^* = 1)Z_{1,i}^*}{n_1^2}\right) * \left(\sum_{i=1}^n \frac{I(L_i^* = 2)Z_{2,i}^*}{n_2^2} - \frac{I(L_i^* = 1)Z_{2,i}^*}{n_1^2}\right)\right) \\
&= E\left(\sum_{i=1}^n \left(\frac{I(L_i^* = 2)Z_{1,i}^*}{n_2^2} - \frac{I(L_i^* = 1)Z_{1,i}^*}{n_1^2}\right) * \left(\frac{I(L_i^* = 2)Z_{2,i}^*}{n_2^2} - \frac{I(L_i^* = 1)Z_{2,i}^*}{n_1^2}\right)\right) \\
&= nE\left(\left(\frac{I(L^* = 2)Z_1^*}{n_2^2} - \frac{I(L^* = 1)Z_1^*}{n_1^2}\right) * \left(\frac{I(L^* = 2)Z_2^*}{n_2^2} - \frac{I(L^* = 1)Z_2^*}{n_1^2}\right)\right) \\
&= nE\left(\frac{I(L^* = 1)Z_1^* Z_2^*}{n_1^2} + \frac{I(L^* = 2)Z_1^* Z_2^*}{n_2^2}\right) \\
&= nE\left(E\left(\frac{Z_1^* Z_2^*}{n_1^2} | L^* = 1\right) * P(L^* = 1) + E\left(\frac{Z_1^* Z_2^*}{n_2^2} | L^* = 2\right) * P(L^* = 2)\right) \\
&= n\left(\frac{1/n(\phi_1 n_1 + \phi_2 n_2)}{n_1^2} \frac{n_1}{n} + \frac{1/n(\phi_1 n_1 + \phi_2 n_2)}{n_2^2} \frac{n_2}{n}\right) \\
&= \frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}.
\end{aligned}$$

Note that in the permutation derivation, the covariance of Z_1^* and Z_2^* is $1/n(\phi_1 n_1 + \phi_2 n_2)$ for both values of L^* , since $Z^* \perp L^*$. Again, it is interesting to note that the final expression for the covariance under permutations resembles that under bootstrap resampling, except with the roles of n_1 and n_2 reversed.