Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Statistical Planning and Inference 125 (2004) 85–100

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Choice of a null distribution in resampling-based multiple testing

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

Division of Biostatistics, University of California, School of Public Health, Earl Warren Hall #7360, Berkeley, CA 94720-7360, USA

Received 27 November 2002; accepted 24 July 2003

Abstract

This paper investigates different choices of test statistic null distribution for resampling-based multiple testing in the context of single parameter hypotheses. We show that the test statistic null distribution for strongly controlling type I error may be obtained by projecting the true test statistic distribution onto the space of mean zero distributions. For common choices of test statistics, this distribution is asymptotically multivariate normal with the covariance of the vector influence curve for the parameter estimator. Applying the *ordinary* non-parametric or model-based bootstrap to mean zero centered test statistics produces an estimated test statistic null distribution which provides asymptotic strong control. In contrast, the usual practice of obtaining an estimated test statistic null distribution via an estimated data null distribution (e.g. null restricted bootstrap) only provides an asymptotically correct test statistic null distribution if the covariance of the vector influence curve is the same under the chosen data null distribution as under the true data distribution. This condition is the formal analogue of the subset pivotality condition (Westfall and Young, Resampling-based Multiple Testing: Examples and Methods for *p*-value adjustment, Wiley, New York, 1993). We demonstrate the use of our proposed ordinary bootstrap null distribution with a single-step multiple testing method which 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 method produces an asymptotically correct null distribution if (i) the sample sizes are equal or (ii) the populations have the same covariance structure.

MSC: primary 62H15; secondary 62H10

Keywords: Multiple testing; Strong control; Type I error; Permutation; Bootstrap

* Corresponding author.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2003.07.019

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 less conservative error rate control, such as control of the false discovery rate (Benjamini and Hochberg, 1995). In recent 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 the multivariate setting, though the results in this paper apply to multiple testing in general.

In Section 2, we formally define a statistical framework for multiple testing and provide definitions of important concepts in terms of the distributions of the data and test statistics. We reach the practical conclusion that "ordinary" bootstrap methods provide the asymptotically correct null distribution for multiple testing. Unlike the "null restricted" bootstrap, this approach does not require the subset pivotality condition given in Westfall and Young (1993, pp. 42–43), 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 demonstrate the practical use of the ordinary bootstrap estimated null distribution with a class of single-step multiple testing procedures MT(c) that provide conservative asymptotic control of any type I error rate which is a function of the distribution of the number of false positives of MT(c). We also illustrate that this method is equivalent to constructing an error-specific confidence region under the sampling distribution corresponding with the model-based bootstrap and checking if the hypothesized value is contained in it. If the data model is non-parametric, for example, then the sampling distribution is the ordinary non-parametric bootstrap distribution.

In Section 3, we consider the two sample problem. We compare different choices of test statistics and estimated null distributions algebraically and in simulations. When testing for a difference in means, the subset pivotality condition holds so that the ordinary and null restricted bootstrap methods are equivalent. We compare the bootstrap null distribution to the permutation null distribution and 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). We conclude with a discussion in Section 4.

2. Multiple testing procedures

Let $X_1, \ldots, X_n$ be i.i.d. $X \sim P \in \mathscr{P}$, where $\mathscr{P}$ 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

$$\mathrm{H}_{0,j}\colon\ \mu_j(P) = \mu_j^0, \quad j = 1,\ldots,p, \tag{1}$$

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

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 $\mathrm{IC}_j(X)$; that is,

$$\sqrt{n}(\mu_{j,n} - \mu_j) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_j(X_i \mid P) + \mathrm{o}_p(1), \tag{2}$$

then by the central limit theorem,

$$\sqrt{n}(\mu_n - \mu(P)) \underset{n\to\infty}{\overset{\mathrm{D}}{\to}} \mathrm{N}(0, \Sigma(P)), \tag{3}$$

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

$$Q_0(P) = \mathrm{N}(0, \Sigma(P)) \tag{4}$$

denote this limit distribution.

It follows that sensible choices of test statistics include

$$T_{jn} \equiv \mu_{jn} - \mu_j^0, \tag{5}$$

$$T_{jn} \equiv \sqrt{n}(\mu_{jn} - \mu_j^0), \tag{6}$$

$$T_{jn} \equiv (\mu_{jn} - \mu_j^0)/\mathrm{sd}(\mu_{jn}), \tag{7}$$

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

In Section 2.4, we show that (4) is the asymptotically correct null distribution for the vector of test statistics (6) 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 (7), then the asymptotically correct null distribution would be $\mathrm{N}(0, \rho(P))$, where $\rho(P)$ is the correlation (rather than covariance) matrix of $\mathrm{IC}(X)$.

2.1. Control of type I error rates

Define a two-sided multiple testing procedure $\mathrm{MT}(c)$ by

$$\text{Reject } \mathrm{H}_{0j} \quad \text{if } |T_{jn}| > c_j,\ j = 1,\ldots,p, \tag{8}$$

for a vector of cut-off values $c \in \mathbb{R}^p$. 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. Suppose c is given. Let $S_0 = \{j\colon \mu_j(P) = \mu_j^0\}$ be the unknown set of true negatives, and define the following random variables:

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

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

This notation is intended to emphasize that these quantities clearly depend on both c and a test statistic distribution Q. The quantity $V(c \mid Q)$ also depends on the set S_0, which is determined by the true distribution P. Notice that even when c and Q vary, the sum in Eq. (9) is always over the fixed set S_0. We will also sometimes use the notation $R(c \mid Z)$, where Z is the random variable of interest. If $Z \sim Q$, then $R(c \mid Z) = R(c \mid Q)$. Now, let $V_n(c) = V(c \mid Q_n(P))$ be the (unobserved) number of false positives of the testing procedure MT(c), and let $R_n(c) = R(c \mid Q_n(P))$ be the total number of rejected hypotheses. Both quantities are defined relative to the true test statistic distribution $Q_n(P)$.

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 $V_n(c)$. We will use the notation F_X to denote the cumulative distribution of a random variable X. Let α be a target error rate. We wish to arrange that $\theta(F_{V_n(c)}) \leqslant \alpha$, at least asymptotically. This is the *error rate* for MT(c). Given the distance measure $d(F_1, F_2) = \max_{j \in \{0,\ldots,p\}} |F_1(\{j\}) - F_2(\{j\})|$ 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:

$$\text{Monotonicity}: \ \text{if } F_1 \geqslant F_2 \quad \text{then } \theta(F_1) \leqslant \theta(F_2), \tag{11}$$

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

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

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

When $k = 1$, the gFWER is the usual family-wise error rate (FWER), as defined by Dudoit et al. (2003).

A common goal of multiple hypothesis testing is to control a chosen type I error rate θ at a target level α under the true data-generating distribution P, while maximizing power. In this paper, we define *strong control* to mean that a testing procedure provides control under P, that is $\theta(F_{V(c|Q_n(P))}) \leqslant \alpha$. This definition is an explicit statement about

the joint distribution $Q_n(P)$. Typically, strong control is defined in terms of subsets of null hypotheses (Hochberg and Tamhane, 1987; Westfall and Young, 1993). Strong control of family-wise error as defined in Westfall and Young (1993, p. 10) implies our definition $\theta(F_{V(c|Q_n(P))}) \leqslant \alpha$. We would like a testing procedure to have strong control, at least asymptotically. *Asymptotic strong control* means that the error rate $\alpha_n = \theta(F_{V(c|Q_n(P))})$ for a sample of size n has the property $\limsup_{n\to\infty} \alpha_n \leqslant \alpha$ under P.

2.2. Choice of null distribution

In this paper, we define a general framework for multiple testing based on a rule MT(c) (Equation (8)) and a vector $c = c(Q, \alpha, P_n)$ of cut-offs (Section 2.3). Multiple testing procedures can also be defined in terms of adjusted p-values (e.g. Westfall and Young, 1993; Yekutieli and Benjamini, 1999; Dudoit et al., 2003). Regardless of the particular testing method, the choice of test statistic null distribution is critical. We define the *null distribution* for a vector of test statistics T_n based on a asymptotically linear estimator μ_n as any zero centered distribution Q_{0n} such that $\lim_{n\to\infty} Q_{0n} = \mathrm{N}(0, \Sigma)$, where Σ is a covariance matrix and we call $\mathrm{N}(0, \Sigma)$ an *asymptotic null distribution*. We prove in Section 2.4 that for $T_n = \sqrt{n}(\mu_n - \mu^0)$, a single-step multiple testing procedure MT(c_0) with $c_0 \equiv c(Q_0, \alpha)$ provides asymptotic strong control of type I error. This shows that $Q_0 = \mathrm{N}(0, \Sigma(P))$ is an asymptotically correct null distribution for the vector of test statistics. 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.

2.2.1. Proposed estimated null distributions

In practice, we do not know the true distribution P, so Q_0 is unknown. Therefore, we use estimated cut-offs c_{0n}, which depend on an estimated null distribution $\hat{Q}_{0n}$. If $\hat{Q}_{0n}$ is a consistent estimator of Q_0, we can asymptotically control the error rate at level α up to the discreteness of $\hat{Q}_{0n}$. Resampling methods are a useful tool for estimation, because they produce estimated null distributions that (i) account for the correlation structure of the data, (ii) provide a *joint* null distribution, and (iii) can be used with both parametric and non-parametric models. Next, we present two resampling-based estimators for which asymptotic strong control is achieved under weak regularity conditions (see Section 2.4). We first give the specific estimators for the case that $T_n = \sqrt{n}(\mu_n - \mu^0)$ and then discuss adaptations for standardized test statistics.

Estimating $\Sigma(P)$: The first proposed estimator is $\tilde{Q}_{0n} = \mathrm{N}(0, \Sigma_n)$, 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. With standardized test statistics $T_n = (\mu_n - \mu^0)/\mathrm{sd}(\mu_n)$, the asymptotically correct null distribution is $\mathrm{N}(0, \rho(P))$, so that one can use the estimated null distribution $\mathrm{N}(0, \rho_n)$, where ρ_n is a consistent estimator of the correlation $\rho(P)$ of IC(X).

Ordinary bootstrap method: The second proposed estimator involves a simple bootstrap method. The "ordinary" bootstrap is applied to properly centered test statistics as

follows. Let $\tilde{P}_n$ be an estimator of the true data-generating distribution P according to the model $\mathscr{P}$ or the empirical distribution (i.e. model-based bootstrap or non-parametric 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 $Z_n^\#$. In practice, this is approximated by the empirical distribution of a large number B realizations from $Q_{0n}^\#$. Under regularity conditions, it is known that the bootstrap is consistent in the sense that $Z_n^\# \overset{\mathrm{D}}{\Rightarrow} Z \sim Q_0$ 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). If $T_n = (\mu_n - \mu^0)/\mathrm{sd}(\mu_n)$, then the bootstrap test statistics should also be standardized, for example $Z_{jn}^\# = (\mu_{jn}^\# - \tilde{\mu}_{jn})/\mathrm{sd}(\mu_{jn}^\#)$, where $\mathrm{sd}(\mu_{jn}^\#)$ is an estimate of $\sigma_j^\#/\sqrt{n} = \sqrt{\mathrm{VAR}(\mathrm{IC}_j(X^\#))}/\sqrt{n}$. Similarly, if $T_n = (\mu_n - \mu^0)$, then the bootstrap test statistics are not multiplied by $\sqrt{n}$: $Z_n^\# = (\mu_n^\# - \tilde{\mu}_n)$.

2.2.2. Comparison with current practice

Currently employed resampling-based multiple testing methodologies identify a test statistic null distribution $Q_n(P_0)$ implied by a choice P_0 of *data* null distribution satisfying $H_0^C = \bigcap_{j=1}^p \mathrm{H}_{0,j}$ and control 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, pp. 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, pp. 42–43) or specifically $\Sigma(P_0) = \Sigma(P)$ (Eq. (14)), which is violated in many applications. On the contrary, our proposed ordinary bootstrap method samples from an estimate of the true distribution, but centers the test statistics at zero. Therefore, we *always* consistently estimate the covariance matrix of the test statistics (even when Eq. (14) 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 the true distribution P onto the space $\mathscr{P}_0 = \{P \in \mathscr{P}\colon \mu = \mu^0\}$. 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:

$$P_0(P) = \arg \max_{P_0' \in \mathscr{M}_0, P_0' \ll \mu} \int \log\left(\frac{\partial P_0'(x)}{\partial \mu(x)}\right) \mathrm{d}P(x), \tag{13}$$

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 $Q_n(P_{0n}) \Rightarrow \mathrm{N}(0, \Sigma(P_0))$, this method provides asymptotic strong control if and only if

$$\Sigma(P_0) = \Sigma(P). \tag{14}$$

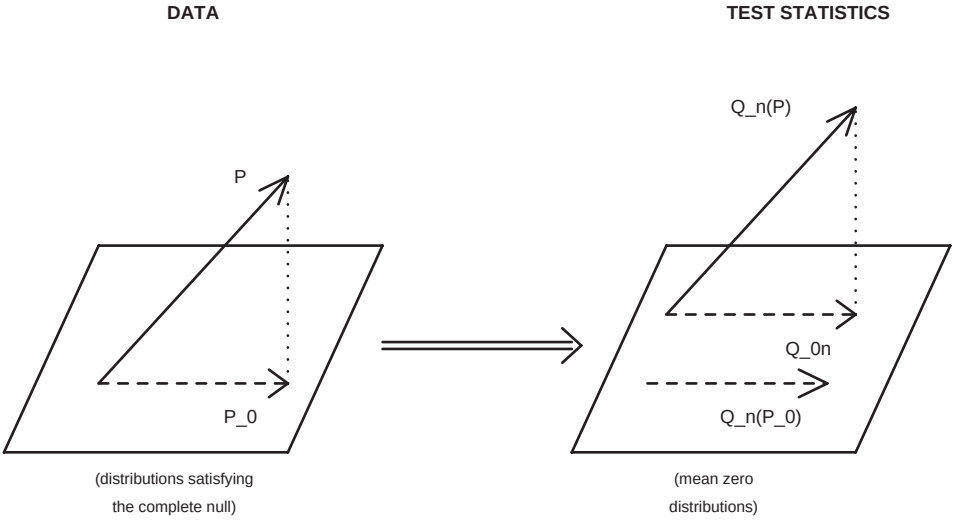

Fig. 1. A commonly used data null distribution is the projection $P_0 = \Pi(P \mid \mathscr{P}_0)$. While there are many data generating distributions satisfying H_0^C, most of these do not provide strong control. For example, with the single-step method of Section 2.3.1, $\theta(F_{R(c|Q_n(P^*))}) \leqslant \alpha$ for some $P^* \in \mathscr{P}_0$ does not imply $\theta(F_{V_n(c)}) \leqslant \alpha$. The true test statistic distribution implied by P and sample size n is $Q_n(P)$. A test statistic null distribution that provides asymptotic strong control for any P is the projection Q_{0n} of $Q_n(P)$ onto the space of mean zero distributions. Notice that $Q_n(P_0)$ (the distribution of the test statistics under P_0) may not be equivalent to Q_{0n}. Eq. (14) gives the condition under which these two distributions converge to the same asymptotic distribution. In practice, these null distributions are estimated.

This condition is the formal analog of the subset pivotality condition (Westfall and Young, 1993). Whenever Eq. (14) holds, it is correct to use the null restricted bootstrap distribution $Q_n(P_{0n})$ as well as our proposed ordinary bootstrap distribution $Q_{0n}^{\#}$, which is always correct. Fig. 1 is a symbolic comparison of the methods of using the projection of the true test statistic distribution and using the distribution of the test statistics under a projection of the data distribution.

Multiple testing for correlations (see Westfall and Young, 1993, pp. 43, 194–199) is a particular example of a case where $\Sigma(P_0) \neq \Sigma(P)$ so that $Q_n(P_{0n})$ and $Q_{0n}^{\#}$ are not asymptotically equivalent, and only $Q_{0n}^{\#}$ converges to the correct asymptotic distribution Q_0. Although the independence null distribution $Q_n(P_{0n}) \xrightarrow[n\to\infty]{\mathrm{D}} \mathrm{N}(0, I)$ estimates the covariance of the test statistics incorrectly, this distribution does yield conservative asymptotic control of the FWER ($k = 1$), even under H_0^C (Westfall and Young, 1993, pp. 194–199). It is not clear, however, that this result will hold for other error rates, such as the gFWER ($k > 1$). The results of Section 2.4 show that methods based on the ordinary bootstrap null distribution $Q_{0n}^{\#}$ provide asymptotic strong control of any type I error that is a function of the distribution of the number of false positives under any true data generating distribution P, allowing us to solve a wider range of testing problems than we could previously.

2.3. Choosing a cut-off vector

Given any test statistic distribution Q (e.g. a particular estimated null distribution), let $c = c(Q, \alpha, P_n) \in \mathbb{R}^p$ be a vector function cut-off rule such that if $T_n \sim Q$ then MT(c) has the property that $\theta(F_{R(c|Q)}) = \alpha$. For a one-sided test, only one tail of Q is used. Notice that MT(c) depends critically on the distribution Q through c. In the case of the single-step method of Section 2.3.1, c depends only on Q and α. The hope is that $\theta(F_{R(c|Q)}) = \alpha$ implies that $\theta(F_{V_n(c)}) \leqslant \alpha$. Step-wise methods typically also use the observed data P_n (e.g. the ordering of the observed test statistics T_n).

2.3.1. Example: a single-step method

One particular method for computing c is to select a common quantile of each marginal distribution of a test statistic null distribution Q_{0n}. Consider, for instance, a vector of cut-offs c satisfying

$$\Pr\left(\sum_{j=1}^{p} I\{|T_{jn}| > c_j\} \geqslant k\right) \leqslant \alpha, \quad T_n \sim Q_{0n}, \tag{15}$$

where k is a pre-specified number of false positives. For one-sided tests, the absolute value in Eq. (15) is omitted and the inequality depends on the direction of the test. In practice, we need to take B resamples from Q_{0n} and compute the cut-offs under the corresponding empirical distribution $\hat{Q}_{0n}$. With a sufficiently smooth resampled null distribution in hand (B large enough), these common quantiles can be fine-tuned to control the chosen error rate exactly under $\hat{Q}_{0n}$. The multiple testing procedure is now completely defined by a choice of estimated null distribution for the test statistics, since $c(\hat{Q}_{0n}, \alpha)$ does not depend on the observed data in this case. In Section 2.5, we observe that this single-step method is equivalent to constructing an error rate specific bootstrap confidence set.

Consider the following specific example of this method. Recall the ordinary bootstrap estimated null distribution $Q^{\#}_{0n}$ (Section 2.2.1). Define

$$R^{\#}_{0n}(c) \equiv R(c \mid Q^{\#}_{0n}) = \sum_{j=1}^{p} I(|Z^{\#}_{jn}| > c_j). \tag{16}$$

Now, let $c_{0n} = c(Q^{\#}_{0n}, \alpha)$ be the common quantiles of $Q^{\#}_{0n}$ such that $\theta(F_{R^{\#}_{0n}(c)}) = \alpha$. Then, the bootstrap-based multiple testing procedure MT(c_{0n}) asymptotically controls the error rate θ strongly at level α (see Section 2.4).

2.4. Asymptotic strong control theorem

We prove that the proposed class of single-step multiple testing procedures have conservative asymptotic strong control of the desired multiple testing type I error rate.

Theorem 1. *Consider the multiple testing framework defined above, with* $T_{jn} \equiv \sqrt{n}(\mu_{jn} - \mu^0_j)$, $j = 1, \ldots, p$ *and* $T_n \sim Q_n(P)$. *Suppose* $\theta(F) \in (0, 1)$ *is a type I error rate satisfying Assumptions* (11) *and* (12). *Define* $V_0(c) = V(c \mid Q_0)$, $R_0(c) = R(c \mid Q_0)$, *and a single-step*

cut-off rule $c_0 = c(Q_0, \alpha)$ *that does not depend on the observed data. Recall that* $V_n(c_0) = V(c_0 \mid Q_n(P))$. *Then the multiple testing procedure* $\mathrm{MT}(c_0)$ *has asymptotic strong control*:

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

In practice, the distribution $F_{R_0(c)}$ *of* $R_0(c)$ *is unknown, because* Q_0 *depends on the unknown data generating distribution* P. *Let* Q_{0n} *be an estimate of* Q_0. *Define* $c_{0n} \equiv c(Q_{0n}, \alpha)$ *and consider* $V_n(c_{0n}) = V(c_{0n} \mid Q_n(P))$. *Suppose that* $c_{0n} \to c_0$ *in probability for* $n \to \infty$. *Then*

$$\limsup_{n\to\infty} \theta(F_{V_n(c_{0n})}) \leqslant \alpha. \tag{18}$$

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 $c(Q_{0n}, \alpha) \xrightarrow{\mathrm{P}} c(Q_0, \alpha)$ 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 $c(Q_{0n}, \alpha) \xrightarrow{\mathrm{P}} c(Q_0, \alpha)$, and thereby the wished asymptotic strong control (18).

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

$$\theta(F_{V_n(c_0)}) \leqslant \theta(F_{R(c_0|Z_n)}),$$

where $R(c \mid Z_n) = \sum_{j=1}^{p} I(|Z_{jn}| > c_j)$. 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 $\Pr(V_n(c_{0n}) \neq V_n(c_0)) = \mathrm{O}(\delta_n)$, where $\delta_n = \max_{j=1,\ldots,p} |c_{0n,j} - c_{0,j}|$. 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

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

Thus,

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

2.5. Equivalence of multiple testing and confidence regions

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. Consider an error rate $\theta(\cdot)$ and the random variable $Z_n = \sqrt{n}(\mu_n - \mu(P))$. We define a θ-specific $100(1-\alpha)\%$ confidence region for $\mu(P)$ as the random region $\{\mu\colon \sqrt{n}|\mu_n - \mu| < c_n\}$, where c_n is chosen such that $\theta(F_{R(c_n|Z_n)}) = \alpha$. 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 this region is the usual $100(1-\alpha)\%$ simultaneous confidence region for $\mu(P)$.

Construction of an asymptotic confidence region requires estimating Q_0. Given an estimate Q_{0n} (e.g. bootstrap or multivariate normal), we let $\tilde{c}_n = c(Q_{0n}, \alpha)$ and obtain an asymptotically correct θ-specific $100(1-\alpha)\%$ confidence region for $\mu(P)$:

$$\left\{\mu\colon \mu_{jn} - \frac{\tilde{c}_{jn}}{\sqrt{n}} < \mu_j < \mu_{jn} + \frac{\tilde{c}_{jn}}{\sqrt{n}},\ j = 1,\ldots,p\right\}. \tag{19}$$

The corresponding multiple testing procedure consists of constructing such an error-specific confidence region for μ and checking if it contains the hypothesized value μ^0. This method is precisely equivalent to our single-step common quantile multiple testing procedure defined in Section 2.3.1. Therefore, one can perform multiple testing asymptotically controlling an error rate $\theta(\cdot)$ conservatively under the true distribution P by using the bootstrap distribution $Q_{0n}^{\#}$ to obtain $\tilde{c}_n = c(Q_{0n}^{\#}, \alpha)$ and following the rule:

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

Remark. Westfall and Young (1993, pp. 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).

3. 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

$$\mathrm{H}_{0,j}\colon \mu_j \equiv \mu_{1,j} - \mu_{2,j} = 0, \quad j = 1,\ldots,p. \tag{20}$$

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:

$$T_{jn} = (\mu_{jn} - 0)/\mathrm{sd}(\mu_{jn}),$$

$$D_{jn} = \mu_{jn} - 0.$$

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.

3.1. Models

Consider the following data models for this two sample problem:

(1) $\mathscr{P}_1$: $X \mid L=1 \sim P_1$ and $X \mid L=2 \sim P_2$, where P_1, P_2 can be arbitrary distributions,
(2) $\mathscr{P}_2$: $X \mid L=1 \sim P_0(\cdot - \mu_1)$ and $X \mid L=2 \sim P_0(\cdot - \mu_2)$, for a common unspecified distribution P_0 with mean zero.

Model $\mathscr{P}_2$ 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 Eq. (20). Other choices of models, which might be more parametric, could also be considered.

3.2. 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 2.2.1. 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 $\mathscr{P}_1$, then $\tilde{P}_n$ 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 $\mathscr{P}_2$, 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 $\mathscr{P}_2$ 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 $p \gg n$) under the model $\mathscr{P}_2$ when H_0^C is true (Lehmann, 1986; Puri and Sen, 1971). As noted above, $\mathscr{P}_2$ implies a stronger null model restriction, which is needed for an exact test.

Table 1
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

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

In contrast, the bootstrap method implied by model $\mathscr{P}_1$ provides only asymptotic control at level α when H_0^C is true.

3.3. Implications for the permutation test

3.3.1. 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. Denote the variance of X_j $(j=1,2)$ by $\sigma^2_{1,j}$ in population 1 and by $\sigma^2_{2,j}$ in population 2. Let ϕ_1 be the covariance between the two variables (X_1, X_2) 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 1, derivations in Pollard and van der Laan, 2003).

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) $\sigma^2_{1,j} = \sigma^2_{2,j} \equiv \sigma^2_j$ $(j=1,2)$ 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 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 $T_{jn} = (\mu_{jn} - \mu_j^0)/\mathrm{sd}(\mu_{jn})$, 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.

3.3.2. 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 (Fig. 2), while for the bootstrap method under model $\mathscr{P}_1$ the bias

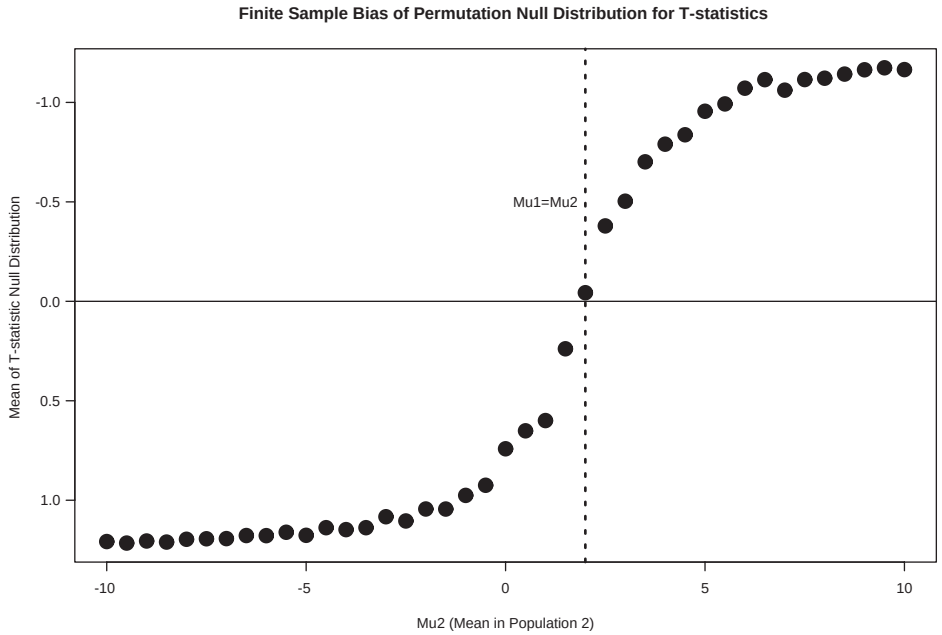

Fig. 2. Mean of the permutation null distribution of the standardized two sample t-statistic for simulated data. Population 1 consists of $n_1 = 2$ observed values, fixed to be 1 and 3. Population 2 consists of $n_2 = 50$ observations from a $N(m, 0.1)$ distribution. The mean of the null distribution is plotted versus different values of the mean m in Population 2 (i.e. as a function of the difference in means). The vertical line marks where the difference in means is truly zero. The mean of the null distribution is close to zero here (as it should be for all m), but increases in magnitude with the difference in means. All 1326 possible permutations were performed exactly.

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.

3.4. Simulations

We have conducted simulations to understand the performance of different multiple testing procedures for the two sample problem. The main findings are summarized here. We refer the reader to Pollard and van der Laan (2003) for a description of data generation and more extensive results.

3.4.1. 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 3.3, 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. 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 3.3. 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 $\mathrm{sd}(\mu_n)$) only helps when the estimate $\mathrm{sd}(\mu_n)$ 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 $\mathrm{sd}(\mu_n)$, which depends on (i) the data-generating model (i.e. model-based estimation versus non-parametric estimation) and (ii) the variance of the data.

3.4.2. Choice of estimated null distribution

For both D_n and T_n, we compare three choices of test statistic null distribution estimators: (i) non-parametric bootstrap $Q_{0n}^{\#}$, (ii) permutation $Q_n(P_{0n})$, and (iii) parametric bootstrap $Q_n(P_{0n})$ (i.e. the data null distribution P_{0n} is a normal distribution with zero difference in means and estimated covariance). Eq. (14) 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. 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 3.3. 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 $\mathrm{sd}(\mu_{jn})$ non-parametrically. Second, ties in the resampling can result in very small estimates of $\mathrm{sd}(\mu_{jn})$. 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 $\mathrm{sd}(\mu_{jn})$) is much less variable, so that the asymptotic results of Hall (1992) will apply. In terms of error rate control, 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).

4. Discussion

Our main contribution is to note that the "ordinary" bootstrap of mean zero centered test statistics can be used as a null distribution to asymptotically control type I error strongly without need for the subset pivotality condition. In particular, we claim 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. 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 $Q_0 = \mathrm{N}(0, \Sigma(P))$ (Eq. (4)). Our theorem shows that under weak regularity conditions, a class of estimators of Q_0 provides asymptotic strong control of most type I error rates with single-step multiple testing methods under *any* true data generating distribution P (regardless of subset pivotality). In future work we will show that this result can be extended to step-down methods based on the same bootstrap estimated null distributions. Using a data null distribution P_0 to obtain a test statistic null distribution, in contrast, only provides a consistent estimator of Q_0 under certain conditions, which we formalize in Eq. (4). A cut-off vector based on an estimate of P_0 will typically depend much less on the true data generating distribution (since it does not use an estimate of the covariance $\Sigma(P)$) than cut offs based on our proposed bootstrap null distribution $Q_{0n}^{\#}$, which will be very responsive to the true P and therefore better estimate the covariance of the test statistic null distribution. It is precisely because of this responsiveness of the cut-off vector as a function of the true distribution P that we obtain control under all distributions.

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} has the *incorrect* covariance unless $\Sigma(P) = \Sigma(P_0)$ or, interestingly, if the design is balanced (i.e. equal sample sizes in the two groups). It is a very powerful fact, however, 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). In our limited simulation study, the standardized t-statistic T_n performed poorly compared to D_n when $\mathrm{sd}(\mu_n)$ was variable (e.g. non-parametric bootstrap with a small sample size).

In this paper, we have focused on asymptotics for fixed dimension p and $n \to \infty$. In this case, the usual central limit theorem applies, and $\mathrm{N}(0, \Sigma(P))$ 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 ($p \gg n$). In Pollard and van der Laan (2003), we consider the example studied by van der Laan and Bryan (2001) and Pollard and van der Laan (2002), in which $n/(\log p) \to \infty$. While it is possible to show uniform consistency of the mean and covariance of the data in this setting, these results do not constitute a multivariate

central limit theorem. It is a topic of future research to investigate the precise conditions under which the multivariate normal approximation $N(0, \Sigma(P))$ is valid.

Acknowledgements

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. We also acknowledge the very helpful suggestions of the referees and guest editors.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. 57, 289–300.

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. J. Amer. Statist. Assoc. 83 (403), 687–697.

Dudoit, S., Shaffer, J., Boldrick, J., 2003. Multiple hypothesis testing in microarray experiment. Statist. Sci. 18 (1), 71–103.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer, Berlin.

Hochberg, Y., Tamhane, A., 1987. Multiple Comparison Procedures. Wiley, New York.

Lehmann, E., 1986. Testing Statistical Hypotheses. Springer, Berlin.

Pollard, K., van der Laan, M., 2002. Statistical inference for simultaneous clustering of gene expression data. In: Dension, D., Hansen, M., Holmes, C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification. Springer, Berlin, pp. 305–320.

Pollard, K., van der Laan, M., 2003. Resampling-based multiple testing: asymptotic strong control of type I error and applications to gene expression data. Working Paper 121, University of California, Berkeley, Division of Biostatistics, Working Paper Series. http://www.bepress.com/webbiostat/paper121.

Puri, M., Sen, P., 1971. Nonparametric Methods in Multivariate Analysis. Wiley, New York.

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

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

Westfall, P., Young, S., 1993. Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York.

Yekutieli, D., Benjamini, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Statist. Plann. Inference 82, 171–196.