Miscellanea

Statistical properties of an early stopping rule for resampling-based multiple testing

BY HUI JIANG

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.
jianghui@umich.edu

AND JULIA SALZMAN

Department of Statistics, 390 Serra Mall, Stanford University, Stanford, California 94305, U.S.A.
julia.salzman@stanford.edu

SUMMARY

Resampling-based methods for multiple hypothesis testing often lead to long run times when the number of tests is large. This paper presents a simple rule that substantially reduces computation by allowing resampling to terminate early on a subset of tests. We prove that the method has a low probability of obtaining a set of rejected hypotheses different from those rejected without early stopping, and obtain error bounds for multiple hypothesis testing. Simulation shows that our approach saves more computation than other available procedures.

Some key words: Bootstrap; Early stopping; False discovery rate control; Multiple hypothesis testing; Resampling.

1. INTRODUCTION

When testing hypotheses on complex datasets, applied statisticians typically use resampling schemes to estimate significance levels. Tests applied to massive data often require either a nontrivial time for each resampling step, computation of hundreds or thousands of individual significance tests, or both. When the time required to estimate resampled p -values is on the order of days, this computation can prevent efficient data analysis. For examples of prohibitive computation time in genetics, see [Siegmund et al. \(2011\)](#), [McLean et al. \(2010\)](#) and [Lin & Tang \(2011\)](#). This raises the question: can the computation time for testing hypotheses be reduced without compromising the power and level of a test of significance? Under natural assumptions, we answer this question affirmatively by developing an adaptive approach for early termination of p -value estimation simulations. Application of the rule to a large dataset substantially reduced the time required to compute p -values.

There is a rich literature on methods that improve the precision of bootstrap quantile estimators or their bias, such as balance or importance resampling ([Efron & Tibshirani, 1993](#); [Davison & Hinkley, 1997](#), Ch. 9), and work that addresses replication size or the power of such estimators ([Bickel & Freedman, 1981](#); [Hall, 1986](#); [Efron, 1987](#); [Davidson & MacKinnon, 2000](#); [Davison & Hinkley, 1997](#), Ch. 4). [Lin & Tang \(2011\)](#) use a stagewise procedure in which all hypotheses are first tested with small numbers of replications, after which small p -values are refined by further replications. [Guo & Peddada \(2008\)](#) use prespecified replication sizes to define candidate early stopping times. Then confidence intervals for bootstrap quantiles to determine stopping times are used to compute the false discovery rate using the procedure in [Benjamini & Hochberg \(1995\)](#).

In this paper, we introduce an adaptive stopping rule to reduce the computation time in resampling-based approaches, such as the bootstrap, when either false discovery rate control or precise estimates for small p -values are sought. Central to our procedure is a mapping of the p -value estimator to a stochastic process, which we use to provide theoretically tractable bounds on testing errors and computational savings. Compared to previous methods, our procedure does not require prespecified stagewise replication sizes and has greater flexibility and theoretical development. We also demonstrate greater computational savings of our approach in examples.

2. EARLY STOPPING PROCEDURE

For clarity, we will explain our methodology for the case of simple bootstrap p -value computations, although the results apply to other resampling methods; see §5. Following the notation of Davidson & MacKinnon (2000), consider the case of a parameter of interest τ for which we have a point estimate $\hat{\tau}$. If the distribution of $\hat{\tau}$ under the null hypothesis H_0 is unknown, the significance level for a test based on $\hat{\tau}$ can be estimated using simulation, for example with the bootstrap. For $j = 1, \dots, n$, τ_j^* is the j th bootstrapped value of $\hat{\tau}$ under H_0 . The estimated bootstrap p -value is

$$\hat{p}^*(\hat{\tau}) = n^{-1} \sum_{j=1}^n I(\tau_j^* > \hat{\tau}),$$

where $I(\cdot)$ is the indicator function. As $n \rightarrow \infty$, $\hat{p}^*(\hat{\tau})$, under regularity conditions, will tend to the ideal bootstrap p -value $p^*(\hat{\tau})$, which is unknown and conditioned on the data (Davidson & MacKinnon, 2000). For simplicity, we will use $\hat{p}^*$ and p^* to denote the estimated and ideal bootstrap p -values throughout this paper. The use of a point estimate $\hat{\tau}$ here is for ease of exposition, but use of an adjusted or pivotal estimator would generally be preferable in practice.

Modern datasets often require m hypotheses to be tested, so m bootstrap p -values are to be estimated, one for each null hypothesis $H_{0,i}$ ($i = 1, \dots, m$). In many genomic applications, m is in the thousands or tens of thousands, as each gene requires the testing of an individual hypothesis. When m is large, even moderate values of n can make computation of bootstrap p -values prohibitive. In the context of multiple testing, the early stopping procedure makes the following assumptions.

Assumption 1. For each null hypothesis $H_{0,i}$, $\hat{\tau}_i$ is a point estimate of τ_i , the parameter of interest. To estimate the p -value based on $\hat{\tau}_i$, the distribution for $\hat{\tau}_i$ under $H_{0,i}$ is generated by a bootstrap simulation. The simulation produces n independent bootstrapped values $\tau_{i,1}^*, \dots, \tau_{i,n}^*$.

Assumption 2. For each $H_{0,i}$ and for each $k = 1, \dots, n$, define $\hat{p}_{i,k}^* = k^{-1} \sum_{j=1}^k I(\tau_{i,j}^* > \hat{\tau}_i)$. The standard estimated bootstrap p -value of the i th hypothesis is hence $\hat{p}_{i,n}^*$.

An example where Assumption 1 holds is when $\tau_{i,j}^*$ ($j = 1, \dots, n$) are simulated $\hat{\tau}_i$ under $H_{0,i}$ obtained by random sampling with replacement from the original dataset. Another example is parametric bootstrapping. The assumption of independence can be relaxed, but doing so is tangential to our main results.

Under the assumptions above, is it possible to bound the deviation of $\hat{p}_{i,k}^*$ from its expectation as a function of k ? Since the answer is yes, stopping the simulation before n iterations can save computation time, and the error made in estimating p -values can be quantified. We will show that since $\hat{p}_{i,k}^*$ can be mapped to a stochastic process, the early stopping procedure below can be used to accurately estimate bootstrap p -values when they are smaller than a prespecified threshold p_0 , an upper bound for the prespecified significance level of a statistical test performed, and stop if there is evidence from early simulations that $\hat{p}_{i,n}^*$ will exceed p_0 . The procedure for early stopping is outlined below:

Early stopping procedure. For a fixed hypothesis i , let p_0 be a p -value threshold of interest. Let a and c be constants independent of i satisfying $a > 0$ and $c > p_0/(1 - p_0)$. For $k = 1, \dots, n$, stop the simulation at the k th iteration if $\hat{p}_{i,k}^* > (a/k + c)/(1 + c)$, in which case denote $\hat{p}_{i,k}^*$ by $\hat{p}_i^{*,s}$. Otherwise, continue until $k = n$ and let $\hat{p}_i^{*,s} = \hat{p}_{i,n}^*$.

Remark 1. If the procedure stops before n iterations, $\hat{p}_i^{*,s} > p_0$.

Remark 2. The constants a , c , p_0 and n are assumed to be the same for all i .

PROPOSITION 1. *The procedure can be mapped to a stochastic process. For $k = 1, \dots, n$, let $\xi_k = I(\tau_{i,k}^* > \hat{\tau}_i) - cI(\tau_{i,k}^* \leq \hat{\tau}_i)$ and $X_k = \sum_{j=1}^k \xi_j$. As ξ_k and X_k will depend on i , there are m such independent random processes. The stopping criterion for the procedure $\hat{p}_{i,k}^* > (a/k + c)/(1 + c)$ reduces to $X_k > a$. For the rest of the paper, we let $T_a \wedge n$ denote the time at which the simulation stops. That is, $T_a \wedge n = \min\{\arg\min_k (X_k > a), n\}$.*

The early stopping procedure operates independently on each individual hypothesis. We state the procedure in the multiple hypothesis framework because it is in this scenario that the procedure's statistical properties and computational savings are of greatest interest. Future work will focus on extensions of this procedure to settings of dependence.

3. STATISTICAL PROPERTIES OF EARLY STOPPING

3.1. Statistical properties for hypothesis testing

PROPOSITION 2. *For constants p_0 and c defined in the early stopping procedure, choose $\theta > 0$ such that $p_0 \exp(\theta) + (1 - p_0) \exp(-c\theta) = 1$. For a fixed hypothesis i , suppose the ideal bootstrap p -value $p_i^* \leq p_0$. Then the probability that the early stopping procedure will terminate before n iterations is at most $\exp(-a\theta)$.*

Remark 3. In this paper, we specify c implicitly by specifying $\delta > 0$ satisfying $c = (1 + \delta)p_0/(1 - p_0)$. Guidelines for choosing δ so that a unique θ in Proposition 2 exists are given in §3.3.

The connection between a bootstrap p -value and the stochastic process shown in Proposition 1 is especially useful for investigating the computational savings in the context of multiple hypothesis testing, where bootstrapping is applied to jointly test a family of multiple hypotheses of similar purpose, and some control over false rejections is desired. In this paper, we consider two widely used multiple testing procedures for the control of false rejections: the control of the false discovery rate $\text{FDR} = E\{I(S + V > 0)V/(S + V)\}$ and the control of the familywise error rate $\text{FWER} = \text{pr}(V \geq 1)$, where V and S are the number of true null and true alternative hypotheses rejected. The false discovery rate controls the expected proportion of false rejections (Benjamini & Hochberg, 1995), and the familywise error rate controls the probability of making one or more false rejections; see Simes (1986), Rom (1990), Hommel (1988) and Hochberg (1988).

PROPOSITION 3. *For $i = 1, \dots, m$, let $\hat{p}_i^* = \hat{p}_{i,n}^*$ denote the p -value that would be computed if the simulation were run to completion. Without loss of generality, assume that $\{\hat{p}_i^*\}_{i=1}^m$ are in increasing order. For hypothesis j , if $\hat{p}_j^* \leq p_0$, then $\hat{p}_{(j)}^{*,s} \geq \hat{p}_j^*$, where $\hat{p}_{(j)}^{*,s}$ denotes the j th smallest $\hat{p}_j^{*,s}$.*

Remark 4. Proposition 3 implies that when rejected hypotheses have estimated p -values less than p_0 , the familywise error rate is no larger when applied to estimated p -values from early stopping compared to those from running the simulation to completion.

PROPOSITION 4. *Suppose that a multiple hypothesis testing procedure applied to $\{H_i\}_{i=1}^m$ rejects the i th hypothesis if $\hat{p}_i^* < \alpha$ and that this procedure controls the false discovery rate at level q . If the early stopping rule is employed with $p_0 \geq \alpha$ and the i th hypothesis is rejected if $\hat{p}_i^{*,s} < \alpha$, then the false discovery rate is controlled at level $q + m \exp(-a\theta)$.*

Straightforward adaptations of the argument used in Proposition 4 can be made to prove an analogous proposition for a large class of step-up procedures in Benjamini & Hochberg (1995) and Benjamini & Yekutieli (2001), or adaptations of the Benjamini–Hochberg procedures such as two-stage step-up procedures (Storey, 2002).

PROPOSITION 5. Suppose that m hypotheses are tested and the significance levels $\{\alpha_i\}_{i=1}^m$ corresponding to rejected hypotheses are such that $\alpha_i \leq p_0$ and result in control of the false discovery rate at level q . With the early stopping rule, the false discovery rate is controlled at level $q + m \exp(-a\theta)$.

3.2. Computational savings

PROPOSITION 6. For a fixed hypothesis i , suppose that the ideal bootstrap p -value $p = p_i^*$ is a random variable uniformly distributed on the interval $[0, 1]$. Then

$$E_p\{E(T_a \wedge n)\} \leq np' + \frac{(a+1)(1-p_0)}{1+p_0\delta} \log\left(\frac{n}{a}\right), \quad (1)$$

where $T_a \wedge n$ denotes the stopping time defined in Proposition 1, $p' = \{a(1-p_0)/n + p_0(1+\delta)\}/(1+p_0\delta)$, δ is defined in Remark 3, and n , the maximum number of simulations, is fixed. The outer expectation is taken with respect to p .

Taking $n = 1000$, $\delta = 1$, $a = 10$, the right-hand side of (1) is roughly 150 when $p_0 = 0.05$ and roughly 79 when $p_0 = 0.01$, reducing the computation by a factor of 7 or 13, respectively. Further, (1) shows that the asymptotic factor for computation reduction using the stopping rule is $(1+p_0\delta)/\{p_0(1+\delta)\}$.

3.3. Choosing parameters

This section addresses how to choose parameters that govern decision rules for the early stopping p -values. There is a trade-off between computational savings and false rejections. For given p_0 and δ , to compute θ , we need to solve the equation $p_0 \exp(\theta) + (1-p_0) \exp\{-p_0(1+\delta)\theta/(1-p_0)\} = 1$. While a closed form solution to this equation involves solving a high degree polynomial equation, convenient numerical solutions are straightforward, and the existence, uniqueness and bounds for θ can be easily established.

PROPOSITION 7. Fix $0 < p_0 < 1/3$ and $0 < \delta < 1$. If

$$\delta < \frac{1-p_0}{p_0} \log_2 \frac{1-p_0}{1-2p_0} - 1,$$

then there exists a unique solution $\theta > 0$ satisfying the equation $p_0 \exp(\theta) + (1-p_0) \exp\{-p_0(1+\delta)\theta/(1-p_0)\} = 1$ with $\delta/2 \leq \theta \leq 2\delta$.

For $p_0 = 0.05$, the above condition requires $\delta < 0.48$, or $\delta < 0.45$ for $p_0 = 0.01$. The precise value of θ can be computed numerically. Figure 1 in the Supplementary Material shows the values and lower/upper bounds for θ under different combinations of p_0 and δ provided by Proposition 7.

4. SIMULATION AND APPLICATION

4.1. Simulation

The early stopping procedure in this paper is, to the best of our knowledge, the first early stopping procedure with theoretical finite sample guarantees in errors of single-hypothesis testing and on false discovery rate control for multiple testing. Thus, there is no way to compare its theoretical error rates to those of other methods in the literature. Here, we use simulations to compare the early stopping procedure with that of Guo & Peddada (2008), in which a number of prespecified replication sizes $B_0 \leq \dots \leq B_K$ for a bootstrap are used for candidate early stopping times. For each $k = 0, \dots, K$, B_k bootstrap replications are simulated, Clopper–Pearson confidence intervals for estimated bootstrap p -values are computed, and the procedure in Benjamini & Hochberg (1995) is applied to the upper and lower bounds of the confidence intervals, respectively. Hypotheses that are accepted or rejected for both applications of the Benjamini–Hochberg procedure are excluded from further bootstrapping and their current bootstrap p -value estimates

Table 1. Comparison between the Guo–Peddada procedure and the early stopping procedure over 100 replications

		Guo–Peddada			Early stopping		
		$B_0 = 125$	$B_0 = 250$	$B_0 = 500$	$\delta = 0.4$ $a = 5$	$\delta = 0.4$ $a = 10$	$\delta = 0.7$ $a = 5$
$\alpha = 0.01$	AveB	330	442	661	159	188	166
	Fold red.	6.1	4.5	3.0	12.6	10.6	12.0
	Degree con.	98%	98%	98%	97%	97%	98%
$\alpha = 0.05$	AveB	355	461	679	250	282	277
	Fold red.	5.6	4.3	2.9	8.0	7.1	7.2
	Degree con.	98%	98%	98%	98%	98%	98%

AveB is the average number of bootstrap replications for each test using the Guo–Peddada procedure or the early stopping procedure. Fold red., the ratio between the replication size 2000 of a full bootstrap procedure and AveB; Degree con., $1 - |R_1 - R_2|/R_2$, where R_1 and R_2 are respectively the numbers of rejected hypotheses using the Guo–Peddada procedure or the early stopping procedure and the Benjamini–Hochberg procedure on the ideal bootstrap p -values (Guo & Peddada, 2008).

are reported. All other hypotheses are included in the next round of B_{k+1} bootstrap replications. No theoretical properties of the computational savings due to the Guo–Peddada procedure are available.

To compare the two procedures, we simulate the testing of $m = 1000$ hypotheses. Among all the hypotheses $i = 1, \dots, m$, 95% are assumed to be true null with ideal bootstrap p -value p_i^* simulated from $\text{Un}(0, 1)$. The remaining 5% of the tests are assumed to be true alternative with ideal bootstrap p -value p_i^* simulated from $\text{Un}(0, 0.0001)$. We do not simulate any specific bootstrap procedure in which data are resampled and test statistics are calculated because both the Guo–Peddada and the early stopping procedures are independent of the specific implementation. For the i th hypothesis and for the k th simulated bootstrap iteration, we draw a Bernoulli random variable with success probability p_i^* , resembling the event $\tau_{i,k}^* > \hat{\tau}_i$. To evaluate the errors made by the early stopping procedure in the same manner as those made by the Guo–Peddada procedure, the Benjamini–Hochberg procedure is run on the p -values estimated by the early stopping procedure, testing at false discovery rate levels $\alpha = 0.01$ and 0.05 . These and all other parameters are set as suggested in the simulations in Guo & Peddada (2008). The confidence level for the Clopper–Pearson exact confidence intervals used in the Guo–Peddada procedure is set to be 0.99 . The prespecified replication sizes used in the Guo–Peddada procedure are set to be $B_0 \leq \dots \leq B_K$ with $B_0 = 125, 250$ or 500 , $B_K = 2000$ and $B_{k+1} = 2B_k$ ($k = 0, \dots, K - 1$). In the early stopping procedure, we set $n = 2000$, $p_0 = \alpha$, $a = 5$ or 10 and $\delta = 0.04$ or 0.07 .

The simulation is repeated 100 times, the average performance of the two approaches is compared using the three criteria defined in Guo & Peddada (2008) and the simulation results are reported in Table 1. The Guo–Peddada and the early stopping procedures agree closely. For the conditions in the simulations, the Guo–Peddada procedure has a 3–6-fold reduction in computational requirements and the early stopping procedure has 7–12-fold reduction.

4.2. Genomic application

The following application demonstrates the effect of the proposed early stopping rule in reducing computation on a genomic dataset generated from a high-throughput sequencing experiment. The method developed in Salzman et al. (2011) was used to estimate isoform-specific gene expression in two conditions. To detect differential isoform usage of a gene, let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the vectors of point estimates of isoform-specific gene expression in two conditions, respectively. The test statistic for differential isoform usage is

$$T = \left\| \frac{\hat{\beta}_1}{\|\hat{\beta}_1\|_1} - \frac{\hat{\beta}_2}{\|\hat{\beta}_2\|_1} \right\|_1,$$

where $\|\cdot\|_1$ denotes the vector L_1 -norm. The estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ are computed by maximizing a Poisson likelihood, and the Poisson parameters are a function of a matrix of constants, called the sampling rate matrix. Details can be found in [Salzman et al. \(2011\)](#).

Proper detection of differential isoform usage depends on accurate estimation of the null distribution of T on a per gene basis, as this distribution varies as a function of the gene being estimated. To accurately estimate the null distribution of T , resampling of read counts needs to be performed on a per gene basis and for each bootstrap iteration a nontrivial computation must be made for the estimation of β_1^* and β_2^* by maximizing the likelihood function using convex optimization. The computation for the null distribution of T is intensive, since a small p -value of 10^{-3} necessarily requires thousands of bootstrap simulations. The scheme for resampling is given in the Supplementary Material.

In our experiments, using the conventional approach without early stopping, it takes five hours to estimate the p -values for 258 genes, each with 5000 simulations. The set of 258 genes are preselected because the computation would be prohibitive if it were run on the full set of over 15 000 expressed genes. Using our early stopping approach with parameters $p_0 = 0.05$, $a = 10$ and $\delta = 1$, the computing time is reduced to only 25 minutes, a 12-fold reduction. The computation is performed on a computer with a 2.66 GHz processor and 32 GB memory. Figure 2 in the Supplementary Material compares the estimated p -values with and without early stopping. Using a significance level of $p_0 = 0.05$, the same 19 genes are called significant in both cases.

5. DISCUSSION

The bounds in this paper are not tight because our purpose is to show how a simple rule and basic statistical theory can be applied to reduce computation time substantially with provable statistical properties. Other applications of our methodology, for example reducing computation times for small p -values, are also possible.

Our results extend to other resampling schemes where the sequentially generated p -value estimates from a simulation can be formulated as a random process, provided the early stopping theory we use here can be applied. This includes p -values from permutation tests, assuming that appropriate regularity conditions are met. Further efficiencies may be achieved by modifying parameter choices in our procedure, for example, by letting n , c , a , p_0 depend on i . Such modifications could have applied utility, and studying their applied and theoretical properties will be of future interest.

ACKNOWLEDGEMENT

We thank Jamie Bates and Patrick O. Brown for providing the sequencing data, and Wing Hung Wong, Xiaquan Wen and Raymond Chen for critical reading of the manuscript. This research was partially supported by grants from the National Institutes of Health and National Science Foundation, U.S.A.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes figures and the resampling scheme used in the genomic example in § 4.2.

APPENDIX

Proofs

LEMMA A1. Under the conditions in Proposition 2, $M_k(\theta) = \exp(\theta X_k)$ is a supermartingale.

Proof. Since $X_k = \sum_{j=1}^k \xi_j$, it suffices to show that for any j , $E\{\exp(\theta \xi_j)\} \leq 1$. By definition of θ , $1 = p_0 \exp(\theta) + (1 - p_0) \exp(-c\theta) \geq p_i^* \exp(\theta) + (1 - p_i^*) \exp(-c\theta) = E\{\exp(\theta \xi_j)\}$. $\square$

Proof of Proposition 2. Let $T_a \wedge n$ be the stopping time defined in Proposition 1. It suffices to show $\text{pr}(X_{T_a \wedge n} > a) \leq \exp(-a\theta)$. Since $\theta > 0$ and $T_a \wedge n$ is bounded, Markov's inequality yields $\text{pr}(X_{T_a \wedge n} > a) = \text{pr}(\theta X_{T_a \wedge n} > a\theta) \leq E\{\exp(\theta X_{T_a \wedge n})\} \exp(-a\theta) \leq \exp(-a\theta)$, where the second inequality is due to Lemma A1 and the optional stopping theorem. $\square$

Proof of Proposition 3. Recall that $\{\tau_{i,k}^*\}_{k=1}^n$ are independently and identically distributed bootstrapped values of $\hat{\tau}_i$ under null hypothesis $H_{0,i}$ in the i th hypothesis test. The simulations producing $\hat{p}_i^*$ and $\hat{p}_i^{*,S}$ can be coupled so that they are defined on the same probability space: for each hypothesis test i where the sampling stops after n_i iterations,

$$\hat{p}_i^{*,S} = \frac{1}{n_i} \sum_{k=1}^{n_i} I(\tau_{i,k}^* \geq \hat{\tau}_i), \quad \hat{p}_i^* = \frac{1}{n} \sum_{k=1}^n I(\tau_{i,k}^* \geq \hat{\tau}_i) \quad (i = 1, \dots, m).$$

Without loss of generality, assume that $\{\hat{p}_i^*\}_{i=1}^m$ are in increasing order. Let $l = \arg\max_{1 \leq i \leq m} \hat{p}_i^* \leq p_0$. As stated in Remark 1, if the simulation stops early, then $\hat{p}_i^{*,S} > p_0$, which establishes that for all $i = 1, \dots, l$, $\hat{p}_i^{*,S} \geq \hat{p}_i^*$. Let $\hat{p}_{(j)}^{*,S}$ be the j th smallest p -value using the early stopping procedure. Since $\{\hat{p}_i^*\}_{i=1}^m$ are assumed to be in increasing order, it follows that $\hat{p}_j^{*,S} \geq \hat{p}_j^* \geq \hat{p}_t^*$ for all $j = 1, \dots, l$, $t = 1, \dots, j$. Hence, $\hat{p}_{(j)}^{*,S} \geq \hat{p}_j^*$. $\square$

Proof of Proposition 4. Recall that the false discovery rate is defined as $\text{FDR} = E\{I(S + V > 0) V / (S + V)\}$, where V and S are the number of true null and true alternative hypotheses rejected using bootstrap p -values estimated from all the n iterations. Let V' and S' be the number of true null and true alternative hypotheses rejected using bootstrap p -values estimated from the early stopping rule. Suppose l_1 of the true alternative and l_2 of the true null hypotheses, rejected using bootstrap p -values estimated from all the n iterations, are no longer rejected using bootstrap p -values estimated from the early stopping rule. By Proposition 3, no new hypotheses will be rejected using the early stopping rule. Then

$$\begin{aligned} \text{FDR}' &= E \left\{ I(S' + V' > 0) \frac{V'}{S' + V'} \right\} \\ &= E \left\{ I(S' + V' > 0) I(l_1 = 0) \frac{V'}{S' + V'} \right\} + E \left\{ I(S' + V' > 0) I(l_1 > 0) \frac{V'}{S' + V'} \right\} \\ &= E \left\{ I(S + V - l_1 - l_2 > 0) I(l_1 = 0) \frac{V - l_2}{S + V - l_1 - l_2} \right\} + E \{ I(l_1 > 0) \} \\ &\leq E \left\{ I(S + V > 0) \frac{V}{S + V} \right\} + E \{ I(l_1 > 0) \} \\ &\leq q + m \exp(-a\theta), \end{aligned}$$

where $E\{I(l_1 > 0)\} \leq m \exp(-a\theta)$ by Proposition 2 since the p -value of the rejected alternative hypothesis is assumed to be at most p_0 . $\square$

Proof of Proposition 5. The proof uses the same argument as in the proof of Proposition 4 to bound $E\{I(S' + V' > 0) V' / (S' + V')\}$. It is assumed that the step-up procedure with significance levels $\{\alpha_i\}_{i=1}^m$ has the property that if the i th hypothesis is rejected, $\hat{p}_i^* \leq \alpha_i \leq p_0$. On the event that no simulation for a p -value with the property that $\hat{p}_i^* \leq \alpha_i$ stops early, $\hat{p}_i^* \leq p_0$, $\hat{p}_i^{*,S} \leq p_0$ and $\hat{p}_i^{*,S} = \hat{p}_i^*$. On this event, $l_1 = 0$. By the same argument as in Proposition 4, $\text{pr}(l_1 > 0) = E\{I(l_1 > 0)\} \leq m \exp(-a\theta)$. $\square$

Proof of Proposition 6. Let p' satisfy the condition in the statement of the proposition. The expectation on the left-hand side of (1), as an integral with respect to p , can be broken into the integrals over $[0, p']$ and $(p', 1]$. The integral over $[0, p']$ is bounded by np' since $T_a \wedge n$ is bounded by n . To bound the integral over $(p', 1]$, consider the martingale $X_k - k(p + pc - c)$, where $p = p_i^* = \text{pr}(\tau_{i,k}^* > \hat{\tau}_i) = \text{pr}(\xi_k = 1)$. Since $p' > p_0(1 + \delta)/(1 + p_0\delta)$, for $p \in (p', 1]$, $E(\xi_k) = p + pc - c > 0$. Since $T_a \wedge n$ is a bounded

stopping time, the optional stopping theorem implies that $0 = E\{X_{T_a \wedge n} - (T_a \wedge n)(p + pc - c)\} \leq (a + 1) - E(T_a \wedge n)(p + pc - c)$, thus $E(T_a \wedge n) \leq (a + 1)/(p + pc - c)$. Integrating this expectation with respect to p over the interval $(p', 1]$ yields the result. $\square$

Proof of Proposition 7. Let $f(\theta) = p_0 \exp(\theta) + (1 - p_0) \exp\{-p_0(1 + \delta)\theta/(1 - p_0)\} - 1$. We have $f(0) = 0$, $f'(\theta) = p_0 \exp(\theta) + (1 - p_0) \exp\{-p_0(1 + \delta)\theta/(1 - p_0)\} \left\{ -\frac{p_0(1 + \delta)}{1 - p_0} \right\} \Big|_{\theta=0} = -p_0\delta < 0$.

The conditions of the proposition imply that $f(\log 2) > 0$, establishing the existence of a solution $\theta \in (0, \log 2)$ by the mean value theorem. Further, the conditions $0 < p_0 < 1/3$ and $0 < \delta < 1$ imply that $\delta \leq (1 - 2p_0)/p_0$, that is, $p_0(1 + \delta)/(1 - p_0) \leq 1$. Therefore, both θ and $p_0(1 + \delta)\theta/(1 - p_0)$ are within $(0, \log 2)$. Taylor's theorem implies that there exists $\xi \in (0, x)$ so that $\exp(x) = 1 + x + (x^2/2) \exp(\xi)$. Thus, if $0 \leq x \leq \log 2$, then $1 + x + x^2/2 \leq \exp(x) \leq 1 + x + x^2$. Similarly, if $0 \leq x \leq \log 2$, then $1 - x + x^2/4 \leq \exp(-x) \leq 1 - x + x^2/2$. Using the above inequalities derived from Taylor's theorem to bound the terms $\exp(\theta)$ and $\exp\{-p_0(1 + \delta)\theta/(1 - p_0)\}$ in $f(\theta)$ yields

$$\delta \left\{ 1 + \frac{(1 + \delta)^2 p_0}{2(1 - p_0)} \right\}^{-1} \leq \theta \leq 2\delta \left\{ 1 + \frac{(1 + \delta)^2 p_0}{2(1 - p_0)} \right\}^{-1} \leq 2\delta. \quad (\text{A1})$$

Combining (A1) with the inequality

$$\left\{ 1 + \frac{(1 + \delta)^2 p_0}{2(1 - p_0)} \right\}^{-1} \geq \left\{ 1 + \frac{(1 + \delta)}{2} \right\}^{-1} \geq \frac{1}{2}$$

implies that $\delta/2 \leq \theta \leq 2\delta$. $\square$

REFERENCES

- BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300.
- BENJAMINI, Y. & YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.
- BICKEL, P. J. & FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–217.
- DAVIDSON, R. & MACKINNON, J. G. (2000). Bootstrap tests: How many bootstraps? *Economet. Rev.* **19**, 55–68.
- DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- EFRON, B. (1987). Better bootstrap confidence intervals. *J. Am. Statist. Assoc.* **82**, 171–200.
- EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton: Chapman and Hall/CRC.
- GUO, W. & PEDDADA, S. (2008). Adaptive choice of the number of bootstrap samples in large scale multiple testing. *Statist. Appl. Genet. Molec. Biol.* **7**, Art. 13, 21.
- HALL, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Ann. Statist.* **14**, 1453–62.
- HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–2.
- HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–6.
- LIN, D.-Y. & TANG, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–67.
- MCLEAN, C. Y., BRISTOR, D., HILLER, M., CLARKE, S. L., SCHAAR, B. T., LOWE, C. B., WENGER, A. M. & BEJERANO, G. (2010). GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501.
- ROM, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663–5.
- SALZMAN, J., JIANG, H. & WONG, W. H. (2011). Statistical modeling of RNA-Seq data. *Statist. Sci.* **26**, 62–83.
- SIEGMUND, D. O., YAKIR, B. & ZHANG, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Statist.* **5**, 645–68.
- SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–4.
- STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 479–98.

[Received September 2010. Revised July 2012]