Daniel B. Rubin*

Evaluations of the Optimal Discovery Procedure for Multiple Testing

DOI 10.1515/ijb-2015-0027

Abstract: The Optimal Discovery Procedure (ODP) is a method for simultaneous hypothesis testing that attempts to gain power relative to more standard techniques by exploiting multivariate structure [1]. Specializing to the example of testing whether components of a Gaussian mean vector are zero, we compare the power of the ODP to a Bonferroni-style method and to the Benjamini-Hochberg method when the testing procedures aim to respectively control certain Type I error rate measures, such as the expected number of false positives or the false discovery rate. We show through theoretical results, numerical comparisons, and two microarray examples that when the rejection regions for the ODP test statistics are chosen such that the procedure is guaranteed to uniformly control a Type I error rate measure, the technique is generally less powerful than competing methods. We contrast and explain these results in light of previously proven optimality theory for the ODP. We also compare the ordering given by the ODP test statistics to the standard rankings based on sorting univariate p-values from smallest to largest. In the cases we considered the standard ordering was superior, and ODP rankings were adversely impacted by correlation.

Keywords: optimal discovery procedure, multiple testing, ranking

1 Introduction

One of the most common problems in both applied and theoreticial statistics involves multiple comparisons of Gaussian means. Suppose we observe $Z_1 \sim N(\mu_1, 1), \ldots, Z_m \sim N(\mu_m, 1)$ and seek to test the multiple null hypotheses H_i: $\mu_i = 0$ against the two-sided alternative hypotheses that $\mu_i \neq 0$. Conceptually, the overwhelming majority of multiple testing procedures involve first ordering the hypotheses based on the degree of evidence for nonzero means, and secondly deciding upon a cutoff somewhere along this ordering to determine statistical significance or non-significance. The ranking in the first step is traditionally determined by ordering the univariate p-values, or equivalently by absolute z-scores $|Z_1|, \ldots, |Z_m|$. The majority of research surrounding this problem, which is too extenstive to summarize, has therefore pertained to the second step of determining where to split the list of p-values.

Storey [1] formulated a novel method for this problem termed the Optimal Discovery Procedure (ODP). Unlike most existing approaches, this technique forms a new set of test statistics $T_1, \ldots, T_m$ to determine a new ordering, such that the statistic for an individual hypothesis depends on all of the observed z-scores. The ODP was shown to maximize the number of expected true positives among a large class of procedures with equal or greater control of the expected number of false positives. While perhaps counterintuitive that an ordering contradicting the usual one could ever be acceptable (why should a z-score closer to zero provide greater evidence for a nonzero mean?), Storey notes an analogy between this hypothesis testing result and well-known results on borrowing strength in multivariate point estimation [2].

After the ODP is used to define test statistics $T_1, \ldots, T_m$, the method requires a user-specified cutoff λ such that the ith hypothesis is rejected when $T_i > \lambda$. This threshold is chosen to control a multivariate measure of the Type I error rate, with higher values of the threshold providing greater protection against false positives. In Section 2 we present definitions and notations to formally specify the method.

In Sections 3 and 4 we present several cautionary results related to the ODP. Specifically, if the cutoff λ must be chosen large enough to control common Type I error rate measures (such as the expected number of false positives or the false discovery rate) uniformly over all data generating distributions, the ODP will

***Corresponding author: Daniel B. Rubin,** U.S. Food and Drug Administration, Silver Spring, MD, USA, E-mail: daniel.rubin@fda.hhs.gov

generally be less powerful than well-known competing methods providing similar guarantees. This loss of power is shown through both theoretical results and numerical comparisons, and is illustrated on two well-known microarray datasets. We discuss why our findings do not contradict the theoretical and numerical optimality results in Storey [1], but why the interpretations differ.

In Section 5 we conclude our assessment of the ODP by recalling the aforementioned two steps in a multiple testing method: ranking the significance of hypotheses and then cutting this ranking. To decouple our evaluation of the ODP test statistics from the thornier issue of the significance threshold, we directly compare the rankings to the standard ordering of univariate p-values. We show through numerical examples that in some cases the ODP can lead to worse rankings than ordering p-values from smallest to largest, and that the ranking quality can degrade with correlation between z-scores.

2 Setup and optimal discovery procedure

Although the ODP can be formulated more generally, for conceptual exposition we follow Storey [1] in considering two-sided testing of Gaussian means with unit variances, and note that many testing problems can be reduced to this framework. Suppose we observe multivariate $Z \sim N(\mu, \Sigma)$, with $Z = (Z_1, ..., Z_m)$, $\mu = (\mu_1, ..., \mu_m)$, and $\Sigma_{i,i} = 1$, and perform m tests, where null hypotheses are H_i: $\mu_i = 0$, and the alternative hypotheses are that $\mu_i \neq 0$. A (non-randomized) multiple testing procedure is a function ϕ: $\mathbb{R}^m \to \{0,1\}^m$, where $\phi(Z) = [\phi_1(Z), ..., \phi_m(Z)]$, meaning hypothesis i is rejected when $\phi_i(Z) = 1$ and is not rejected if $\phi_i(Z) = 0$.

There are many multivariate generalizations of the Type I and II error rates. In line with Storey's optimality results and power comparisons we initially consider controlling Type I errors through expected false positives (EFP) and assessing power through expected true positives (ETP). For the data generating distribution defined by (μ, Σ) these quantities are

$$\text{EFP}(\mu, \Sigma) = E_{(\mu,\Sigma)}\left[\sum_{i=1}^{m} 1(\mu_i = 0)\phi_i(Z)\right],$$
$$\text{ETP}(\mu, \Sigma) = E_{(\mu,\Sigma)}\left[\sum_{i=1}^{m} 1(\mu_i \neq 0)\phi_i(Z)\right].$$

Storey's theoretical results for the ODP allow arbitrary dependence between test statistics, in that "no restrictions are placed on the probabilistic dependence between the tests." Hence, we do not restrict covariance matrix Σ when considering procedures controlling this multivariate Type I error rate measure.

In what follows we also consider procedures that control the false discovery rate (FDR) of Benjamini and Hochberg [3] due to its popularity for large datasets, its interpretability across experiments with different numbers of measurements, and the discussion in Storey [1] of using the ODP in combination with FDR control. The FDR is defined as the expected ratio of false rejections to total rejections:

$$\text{FDR}(\mu) = E_{\mu}\left[\frac{\sum_{i=1}^{m} 1(\mu_i = 0)\phi_i(Z)}{\max\left(1, \sum_{i=1}^{m} \phi_i(Z)\right)}\right]. \tag{1}$$

Because most literature on the FDR is based on independent test statistics, we assume the Gaussian z-scores $Z_1, ..., Z_m$ are uncorrelated when comparing FDR-controlling multiple testing procedures, and hence have suppressed the covariance matrix in the notation of eq. (1).

Based on applying a multivariate generalization of the Neyman-Pearson likelihood ratio to this case of testing multiple Gaussian means, the ODP is defined Storey [1, eq. (5)] in terms of the following test statistic for the ith null hypothesis:

$$T_i^* = \sum_{j=1}^{m} \exp(-0.5|Z_i - \mu_j|^2 + 0.5Z_i^2). \tag{2}$$

The procedure rejects the ith null hypothesis when $T_i^* > \lambda$, where λ is a common user-specified threshold for all test statistics that defines a tradeoff between Type I and II errors. The technique is an example of a single thresholding procedure, meaning a method that defines a test statistic function for each hypothesis and applies a common rejection cutoff. Storey proves that the test statistics defined by eq. (2) maximize the expected number of true positives $\mathrm{ETP}(\mu, \Sigma)$ among all single thresholding procedures with an equal or lesser number of expected false positives $\mathrm{EFP}(\mu, \Sigma)$.

Unfortunately, this method cannot be used in practice because the test statistics depend on the unknown means that are to be tested. Consequently, Storey recommends estimating each unknown mean μ_j by the unbiased Z_j and plugging these values into the above formula. The resulting procedure Storey [1, eq. (6); Storey et al. [4], eq. (3.2)] rejects the ith null hypothesis $H_i : \mu_i = 0$ in favor of the alternative hypothesis $\mu_i \neq 0$ when the test statistic

$$T_i = \sum_{j=1}^{m} \exp(-0.5Z_j^2 + Z_j Z_i)$$

exceeds the common threshold λ. In the remainder of the paper, we thus refer to this implementable method as the ODP.

A nonstandard feature of the method is that the test result for hypothesis i depends not only on the z-score Z_i, but also on the z-scores for all other hypotheses. Hence, one test may reject while another with a larger univariate z-score fails to reject. Storey draws parallels between this borrowing of information and the James-Stein paradox in point estimation [2], in which the naive estimator of a multivariate Gaussian mean is outperformed by a shrinkage estimator that uses all data points to estimate each component of the mean vector.

In this work we only consider the simplest implementation of the ODP in which the rejection threshold λ is fixed. This restriction is due to analytical tractability, and the previous development of optimality theory for this case. However, Storey [1] and Storey et al. [4] also discuss several modifications that can involve estimating test statistic densities under null and alternative hypotheses, estimating the number of true null hypotheses, and using bootstrap resampling to more adaptively estimate significance thresholds. Additional research would be required to characterize how our results translate to these extensions.

3 Evaluating power under EFP control

As a benchmark technique to compare against the ODP, we consider the simple Bonferroni-style method that rejects the ith null hypothesis when

$$|Z_i| \geq c,$$

for some cutoff c that is common across all tests and is chosen to control a Type I error rate. In what follows we refer to this method as the uniformly most powerful (UMP) unbiased procedure, due to the fact that each of the m individual tests are UMP unbiased [5]. Unlike the ODP this benchmark method does not attempt to borrow information across the z-scores when testing each individual null hypothesis.

Storey suggests the ODP is more powerful than the UMP unbiased procedure, in part through numerical comparisons that can be implemented as follows (e.g., Figure 2 in Storey [1]):

1. Select a data generating distribution, determined by some (μ, Σ).
2. Select a level q at which to control the expected number of false positives $\mathrm{EFP}(\mu, \Sigma)$.
3. Choose cutoffs $c = c(\mu, \Sigma)$ and $\lambda = \lambda(\mu, \Sigma)$ for the UMP unbiased and ODP so that each satisfy $\mathrm{EFP}(\mu, \Sigma) = q$.
4. Assess the power of each of the two procedures with $\mathrm{ETP}(\mu, \Sigma)$, the expected number of true positives.

We do not believe such comparisons necessarily place the two techniques on equal footing. The reason is that if trying to decide whether to use the UMP unbiased or ODP, one would not know the (μ, Σ) needed to form cutoffs $c(\mu, \Sigma)$ and $\lambda(\mu, \Sigma)$ making the Type I error rates equivalent. Rather, cutoffs c and λ chosen beforehand to define the two procedures might not provide equally sharp control at the (μ, Σ) distribution. If worst-case Type I error rates differ, we would not traditionally think of the two techniques as providing the same amount of protection against false positives. When confronted with a multiple testing problem, a more traditional approach is to decide upon a Type I error rate measure and consider procedures that ensure control no matter what distribution generates the data. In a typical example, Benjamini and Liu [6] compare the power of step-up and step-down methods, but only when both control the FDR at a given level under all possible data generating distributions (for independent tests). When the Type I error rate measure is the expected number of false positives, procedures meeting such a uniformity requirement should satisfy

$$\sup_{(\mu,\Sigma)} \mathrm{EFP}(\mu, \Sigma) \le q. \tag{3}$$

To compare the UMP unbiased and ODP, we consequently proceed as follows:

1. Decide upon a Type I error rate measure. As before, we can ensure the expected number of false positives does not exceed q.
2. Choose cutoffs c and λ so the UMP unbiased and ODP guarantee the expected number of false positives will not exceed q no matter what distribution generates the data.
3. Consider a specific data generating distribution, determined by (μ, Σ), and compare the power of the two procedures in terms of the expected number of true positives $\mathrm{ETP}(\mu, \Sigma)$.

The only implementation issue that may not be immediately clear is how to guarantee Type I error rate control. The following lemma specifies how the necessary cutoffs must be chosen for the two procedures.

Lemma 1 *The UMP unbiased procedure and ODP respectively guarantee control of the expected number of false positives at level q in the sense of (3) if and only if $c \ge Z_{1-q/(2m)}$ and $\lambda \ge m \exp(0.5\chi^2_{1-q/m,1})$, where Z_α and $\chi^2_{\alpha,1}$ are the α-quantiles of the standard normal distribution and chi-square distribution with one degree of freedom.*

Proof. Set c and λ to the values in the lemma, and consider arbitrary (μ, Σ). The UMP unbiased result is widely known. If the ith null hypothesis is true, then $P(|Z_i| \ge c) = q/m$. The expectation of $1(\mu_i = 0)1(|Z_i| \ge c)$, the ith term in $\mathrm{EFP}(\mu, \Sigma)$, is thus bounded above by q/m. Hence, $\mathrm{EFP}(\mu, \Sigma)$ is bounded above by q, yielding the "if" part of the lemma. Equality is achieved if all m null hypotheses are true, implying the "only if" part.

For the ODP, if the ith null hypothesis is true then $P(m \exp(0.5Z_i^2) \ge \lambda) = P(Z_i^2 \ge \chi^2_{1-q/m,1}) = q/m$. Further, as the quadratic $x \to -0.5x^2 + Z_i x$ is maximized at $x = Z_i$, it follows that $\exp(-0.5Z_j^2 + Z_i Z_j) \le \exp(0.5Z_i^2)$, so $T_i = \sum_{j=1}^m \exp(-0.5Z_j^2 + Z_j Z_i) \le m \exp(0.5Z_i^2)$. Therefore, the ith term of $\mathrm{EFP}(\mu, \Sigma)$ is the expectation of $1(\mu_i = 0)1(T_i \ge \lambda)$, which is bounded above by the expectation of $1(\mu_i = 0)1(m \exp(0.5Z_i^2) \ge \lambda)$, and we just showed this is bounded above by q/m. Thus, $\mathrm{EFP}(\mu, \Sigma)$ is bounded above by q. The argument shows equality is achieved by approaching the degenerate distribution where all m null hypotheses are true and $\Sigma_{i,j} = 1$, so $Z_1 = Z_2 = \ldots = Z_m$ with probability one. □

When both methods are required to ensure Type I error rate control at the same level q for all distributions, which will have higher power at the specific unknown distribution defined by (μ, Σ) that is generating the data? Our next result shows the UMP unbiased procedure should always be preferred.

Lemma 2 *Suppose the UMP unbiased procedure and ODP are defined by cutoffs $c = Z_{1-q/(2m)}$ and $\lambda = m \exp(0.5\chi^2_{1-q/m,1})$, to ensure control of the expected number of false positives as in Lemma 1. Then the UMP unbiased procedure dominates the ODP, in that the latter rejects hypothesis i only if the former also does so.*

Proof. From the argument in the proof of Lemma 1, we note that $T_i = \sum_{j=1}^{m} \exp(-0.5Z_j^2 + Z_jZ_i) \leq m \exp(0.5Z_i^2)$. Further, if the UMP unbiased procedure fails to reject hypothesis i, then $|Z_i| < c$ so $Z_i^2 < c^2 = \chi^2_{1-q/m,1}$. Hence, $T_i = \sum_{j=1}^{m} \exp(-0.5Z_j^2 + Z_jZ_i) \leq m \exp(0.5Z_i^2) < m \exp(0.5\chi^2_{1-q/m,1}) = \lambda$, so the ODP also fails to reject the ith null hypothesis. □

When the UMP unbiased procedure and ODP are each required to guarantee the same level of control for the expected number of false positives, the power loss from using the ODP can be substantial. We illustrate this below in Figure 1 by comparing the procedures on two microarray examples. In our first dataset, expression levels were available for 52 prostate tumors and 50 non-prostate tumor samples, for 6,033 genes. In the second dataset, there were expression measurements for 40 colon tumor and 22 normal colon tissues, for 2,000 genes. The prostate and colon cancer data were respectively discussed in Singh et al.[7] and Alon et al.[8]. We downloaded the datasets from a website maintained at http://stat.ethz.ch/~dettling/bagboost.html, and a description of the preprocessing can be found in Dettling and Bühlmann [9]. When the data for the ith gene in n tumor samples and l normal samples were $(X_{i,1}, \ldots, X_{i,n})$ and $(Y_{i,1}, \ldots, Y_{i,l})$, we formed a test statistic

$$Z_i = \frac{n^{-1}\sum_{j=1}^{n} X_{i,j} - l^{-1}\sum_{k=1}^{l} Y_{i,k}}{\sqrt{n^{-1}\sum_{j=1}^{n}(X_{i,j} - \bar{X}_i)^2 + l^{-1}\sum_{k=1}^{l}(Y_{i,k} - \bar{Y}_i)^2}}.$$

With m genes and scores $(Z_1, \ldots, Z_m)$ we considered this differential expression problem as an approximation to our earlier formulation of testing Gaussian means. For each dataset we applied both the UMP unbiased procedure and ODP, ensuring the expected number of false positives was controlled at levels between $q = 0.05$ and $q = 10$ by choosing cutoffs c and λ as specified in the previous lemmas. As shown in Figure 1, the ODP led to many fewer significant genes.

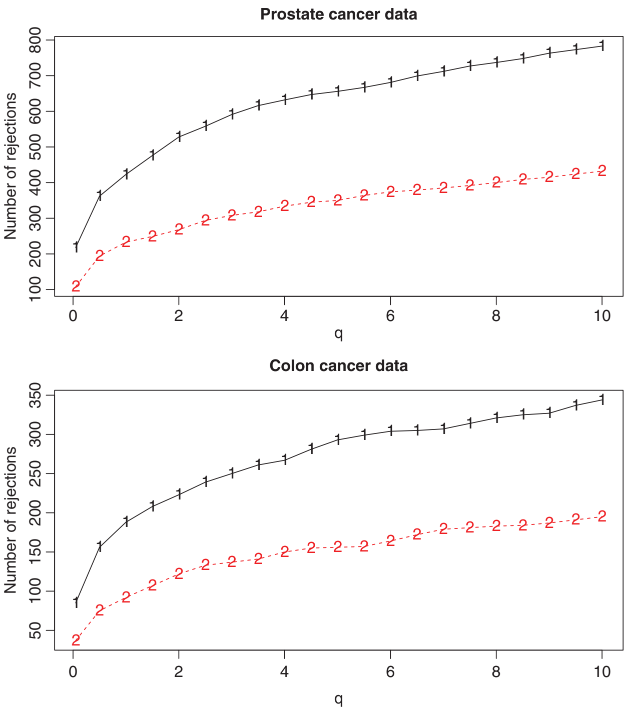

Figure 1: Number of rejections for the prostate and colon cancer problems discussed in Section 3, when ensuring multiple testing procedures must control the expected number of false positives at level q for all distributions. The number of rejections for the UMP unbiased procedure and ODP are represented by 1 and 2.

4 Evaluating power under FDR control

We next examine the power of the ODP when aiming to control the false discovery rate (FDR). As previously noted, we assume independence of z-scores $Z_1, \ldots, Z_m$ when considering FDR control. Storey [1] (section 5.1) discusses the connection between the ODP and false discovery rates, and formulates an approximate optimality property of the testing method for this Type I error rate measure.

The benchmark we use for comparisons is the well-known step-up method of Benjamini and Hochberg [3]. For $\Phi(\cdot)$ denoting the cumulative distribution function, this technique defines p-values $P_i = 2(1 - \Phi(|Z_i|))$ and the corresponding sorted values $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ and sorted null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(m.)}$. The procedure rejects all $H_{(i)}$ for $i \leq k$, where k is the largest value such that $P_{(k)} \leq k \cdot q/m$. Benjamini and Hochberg prove that with independent tests this method guarantees FDR control at level q.

Letting $\tilde{\mu} = (0, \ldots, 0)$ denote the zero vector with m components, it is simple to select the threshold for the ODP via Monte Carlo such that $\text{FDR}(\mu) = q$. When we cannot a priori exclude the mean vector λ as being responsible for generating the data, and seek to control the FDR, the significance cutoff λ clearly must be large enough to ensure FDR control under this distribution. Hence, selecting the threshold λ at this value allows a fair power comparison between the Benjamini-Hochberg method and ODP, because the former provides worst-case FDR control that is at least as strong as the latter:

$$\sup_{\mu} \text{FDR.BH}(\mu) \leq q = \text{FDR.ODP}(\tilde{\mu}) \leq \sup_{\mu} \text{FDR.ODP}(\mu). \tag{4}$$

We compared the power of the two multiple testing techniques under four configurations considered by Storey. The distributions were as follows, where when several alternative means are listed there were equal numbers among true alternatives: (i) 48 tests, 24 true nulls, alternative means −1, 1, 2, and 3; (ii) 48 tests, 12 true nulls, alternative means −1, 1, 2, and 3; (iii) 48 tests, 24 true nulls, alternative means 1, 2, and 3; (iv) 48 tests, 24 true nulls, alternative means −2, −1, 1, and 2.

In each of the four configurations we varied the level of FDR control between $q = 0.01$ and $q = 0.10$. From Figure 2 we observe the Benjamini-Hochberg procedure will generally be more powerful than the ODP if requiring both methods to guarantee control of the FDR.

5 Optimal discovery procedure rankings

A possible criticism of our approach in the preceding sections is that we evaluated the power of the ODP only after selecting the rejection threshold λ large enough to uniformly control a Type I error measure over multiple distributions, including distributions unlikely to arise in practice. Less conservative or more adaptive choices of the rejection threshold might allow increased power while only inflating a Type I error rate measure above a prescribed level in extreme circumstances, such as when all null hypotheses are true. However, this line of argument is not specific to our formulation and would critically apply to much of the multiple testing literature.

Storey [1] notes that multiple testing procedures can conceptually be broken into the two major steps of "(a) determining the order in which the tests should be called significant and (b) choosing an appropriate significance cut-off somewhere along this ordering." Much of the novelty of the ODP lies in the ordering of the initial step, as the ranking for one hypothesis depends on z-scores for others. In contrast, our preceding results revolved around the significance threshold in the second step.

In this section we attempt to directly assess the ODP rankings, and decouple this assessment from our discussion of the significance threshold. If the ODP often dramatically improves rankings by borrowing strength across z-scores, its behavior under extreme data generating distributions might be less practically relevant. Conversely, if rankings from the ODP are consistently worse than the standard sorting of absolute z-scores, the procedure might be adding noise rather than borrowing strength.

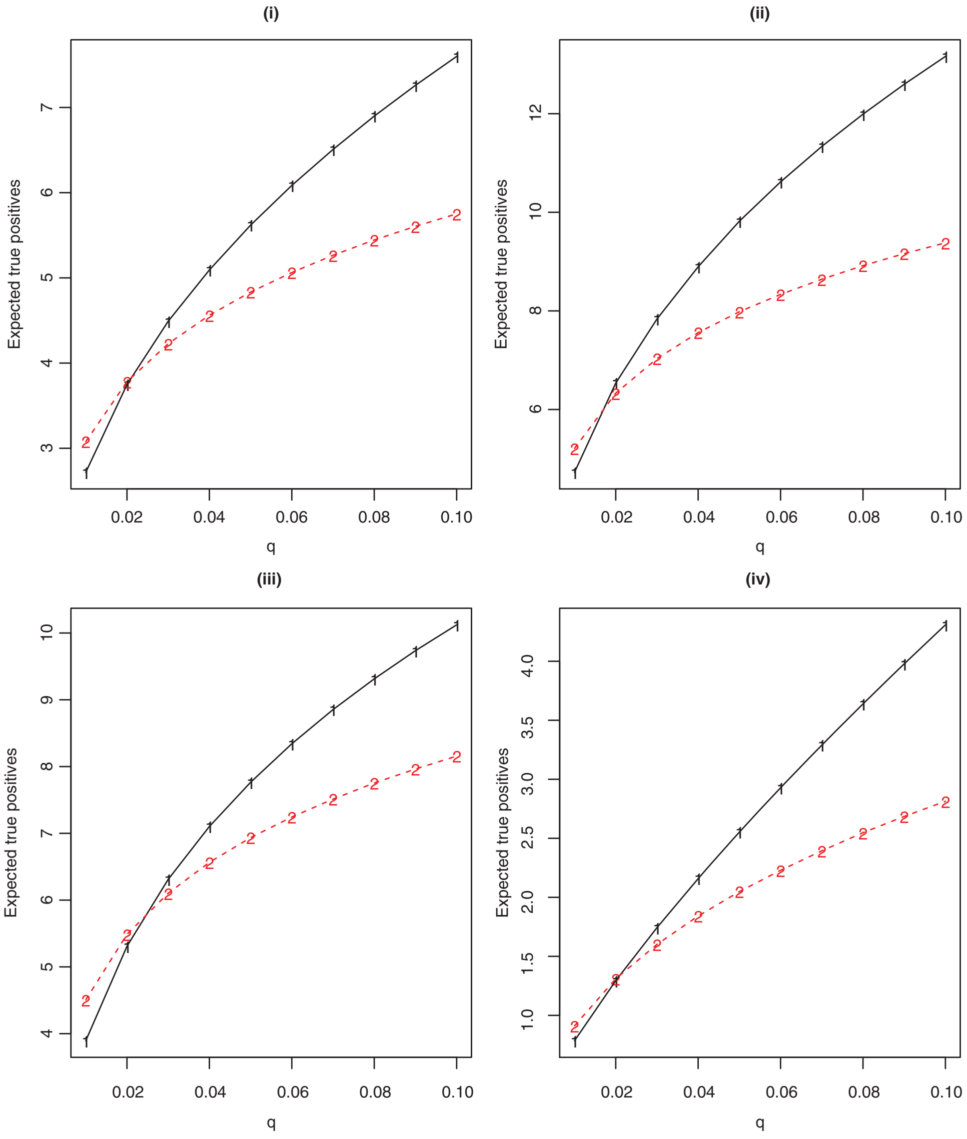

Figure 2: Power comparisons for FDR-controlling procedures for the four distributions discussed in Section 4. The *x*-axis displays the level of FDR control between $q = 0.01$ and $q = 0.10$. The *y*-axis displays the expected true positives (ETP), which for the Benjamini-Hochberg method and ODP are represented by 1 and 2.

We compared rankings obtained by ordering the ODP test statistics $T_1, \ldots, T_m$ to the rankings obtained by ordering the absolute values of *z*-scores $|Z_1|, \ldots, |Z_m|$. The latter was equivalent to the usual ordering based on sorting univariate *p*-values from smallest to largest.

We considered classification rules of the form “call the top *k* ranked test statistics significant.” Such a rule made a correct classification when the test statistic for a true alternative hypothesis was among the *k* largest, or when the test statistic for a true null hypothesis was not among the *k* largest. After fixing a value of $k \in \{0, 1, \ldots, m\}$, and fixing a specific data generating distribution, we compared the ODP rankings to the standard rankings through their average rates of correct classification.

We note that such an ordering-based classification might occur in practice, for instance, if time or resource constraints only allow follow-up investigation of *k* hypotheses considered in an initial experiment. However, our purpose in considering such classification was to compare two sets of rankings on a level playing field, and this section is not meant to endorse the “top k” method of significance testing for fixed *k*.

As an illustrative example, we evaluated the two procedures under the configuration considered by Storey in which $\mu = (0, -2, 0, 0, 1, 2, 0, 3)$. Hence, there were four true null hypotheses, there were four true alternative hypotheses, and the nonzero means were not all of the same sign. With test statistics $Z_1, \ldots, Z_8$ and a fixed k we sought to order the hypotheses such that the top k contained as many true alternatives as possible and as few true nulls as possible.

When constraining both ranking procedures to classify k of the eight means as nonzero, we calculated the expected classification rate of the ODP rankings and the absolute z-score rankings through the Monte Carlo method, with enough number of replications to ensure that simulation error did not qualitatively affect our results.

Because we found that correlation impacted the quality of rankings, we repeated the entire simulation experiment four times with different values of $\text{Cor}_{i \neq j}(Z_i, Z_j) = \rho$. In our simulations we ranged the correlation parameter ρ over 0, 0.15, 0.3, and 0.45.

Figure 3 displays results. When the z-scores were uncorrelated the two procedures gave rankings that were of similar quality, although the expected rate of correct classifications rate was slightly lower for the

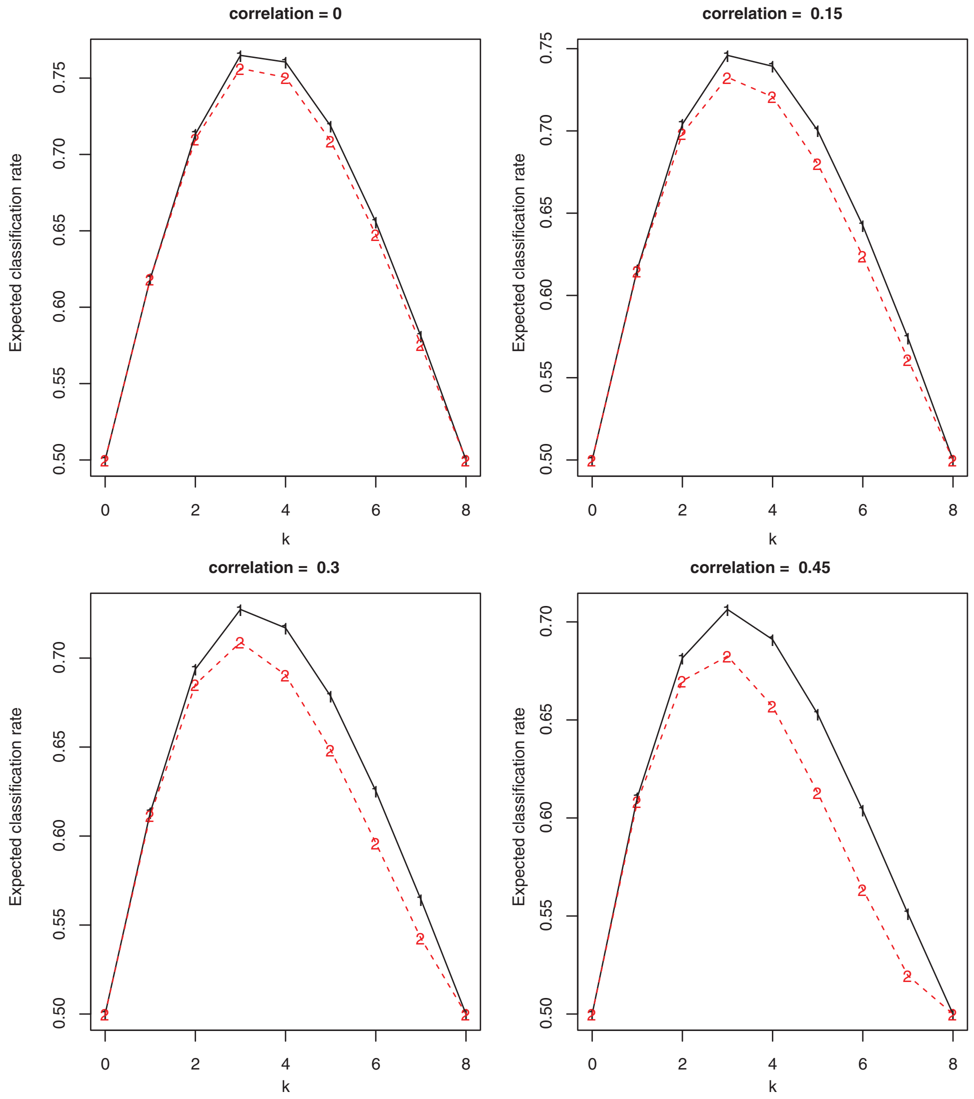

Figure 3: Expected classification rates for rules that classify means as zero or nonzero based on the top k absolute z-scores (represented by 1 in the figure) or the top k ODP test statistics (represented by 2 in the figure), for the distribution discussed in Section 5.

ODP than for the standard rankings obtained by ordering the absolute z-scores. As we increased correlation between z-scores, the ODP rankings deteriorated.

We concluded from this simple example that the usual univariate ranking of z-scores or p-values can sometimes outperform the rankings of the ODP. A more detailed assessment would be required to determine which of the two procedures more generally tends to produce a better ordering.

6 Discussion

In this work we have presented several cautionary results that should be kept in mind when applying the ODP in a simultaneous testing problem. First, if the significance cutoff is chosen to guarantee control of a multivariate Type I error rate over all possible data generating distributions, the ODP will tend to have less statistical power than competing methods providing similar guarantees. Second, for the examples we considered, the rankings obtained by ordering the ODP test statistics were outperformed by the standard ranking of univariate p-values. While our overall assessment is cautionary the evaluations in this work apply to the most basic version of ODP, and variants of the method may address limitations.

Disclaimer: This article reflects the views of the author and should not be construed to represent FDA's views or policies.

References

1. Storey JD. The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc Series B 2007;69:347–68.
2. James W, Stein C. Estimation with quadratic loss. Proc Fourth Berkeley Symp Math Statist Probl 1961;1:361–79.
3. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B 1995;57:289–300.
4. Storey JD, Dai JY, Leek JT. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostatistics 2007;8:414–32.
5. Lehman EL. Testing statistical hypotheses. New York, NY: John Wiley & Sons, 1959.
6. Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. J Stat Plan Inference 1999;82:163–70.
7. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203–9.
8. Alon U, Barkai N, Notterdam D, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 1999;96:6745–50.
9. Dettling M, Bühlmann P. Supervised clustering of genes. Genome Biol 2002;3:research0069-0069.15.