The optimal discovery procedure: a new approach to simultaneous significance testing

John D. Storey

University of Washington, Seattle, USA

Summary. The Neyman–Pearson lemma provides a simple procedure for optimally testing a single hypothesis when the null and alternative distributions are known. This result has played a major role in the development of significance testing strategies that are used in practice. Most of the work extending single-testing strategies to multiple tests has focused on formulating and estimating new types of significance measures, such as the false discovery rate. These methods tend to be based on *p*-values that are calculated from each test individually, ignoring information from the other tests. I show here that one can improve the overall performance of multiple significance tests by borrowing information across all the tests when assessing the relative significance of each one, rather than calculating *p*-values for each test individually. The 'optimal discovery procedure' is introduced, which shows how to maximize the number of expected true positive results for each fixed number of expected false positive results. The optimality that is achieved by this procedure is shown to be closely related to optimality in terms of the false discovery rate. The optimal discovery procedure motivates a new approach to testing multiple hypotheses, especially when the tests are related. As a simple example, a new simultaneous procedure for testing several normal means is defined; this is surprisingly demonstrated to outperform the optimal single-test procedure, showing that a method which is optimal for single tests may no longer be optimal for multiple tests. Connections to other concepts in statistics are discussed, including Stein's paradox, shrinkage estimation and the Bayesian approach to hypothesis testing.

Keywords: Classification; False discovery rate; Multiple-hypothesis testing; Optimal discovery procedure; *q*-value; Single-thresholding procedure

1. Introduction

The well-known Neyman–Pearson (NP) lemma provides an optimal procedure for performing a single significance test when the null and alternative distributions are known (Neyman and Pearson, 1933). Given observed data, the optimal testing procedure is based on the likelihood ratio

$$\frac{\text{probability of data under alternative distribution}}{\text{probability of data under null distribution}}.$$

The null hypothesis is then rejected if the likelihood ratio exceeds some prechosen cut-off. This NP procedure is optimal because it is 'most powerful', meaning that for each fixed type I error rate there does not exist another rule that exceeds this one in power. The optimality follows intuitively from the fact that the strength of the alternative *versus* the null hypothesis is assessed by comparing their exact likelihoods.

Address for correspondence: John D. Storey, Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.
E-mail: jstorey@u.washington.edu

Although a single-hypothesis test involves forming a statistic, deriving a set of significance regions, and determining the type I error rate for each region, these components can conceptually be broken down into two major steps when testing multiple hypotheses:

(a) determining the order in which the tests should be called significant and
(b) choosing an appropriate significance cut-off somewhere along this ordering.

The first step can also be thought of as the process of ranking the tests from most to least significant. The ideas that Neyman and Pearson put forth are most related to this step. However, the field of 'multiple-hypothesis testing' has focused on the second step, where it is assumed that the first step has been carried out on a univariate, test-by-test, basis.

Multiple-testing methods are typically defined in terms of p-values, which are assumed to have been individually obtained from each significance test. The goal is then to estimate the appropriate cut-off to obtain a particular error rate, usually based on the familywise error rate or false discovery rate (Shaffer, 1995). The ordering of the p-values gives the ordering in which the tests are called significant, regardless of the multiple-testing error rate that is employed. Importantly, the p-values have individually been obtained from each significance test, using information from only that significance test. Therefore, no information *across tests* is employed when deciding the order in which to call the tests significant, ultimately affecting the quality of the entire procedure.

This paper is concerned with how to perform the first step optimally, given a certain significance framework for the second step. Specifically, the goal is to determine the order in which the tests are called significant so that we have maximized the expected number of true positive for each fixed expected number of false positive results, an optimality criterion that I show to be directly related to optimality in terms of false discovery rates. I define the 'optimal discovery procedure' (ODP) to be the procedure that achieves this optimization. I derive this ODP and prove its optimality. The procedure involves the formation of a statistic for each test that uses the relevant information from every other test, which is similar to shrinkage estimators that are now commonly used in simultaneous point estimation.

By introducing the ODP, I show that one can improve the performance of multiple-testing procedures by 'borrowing strength' across the tests when determining the order in which tests should be called significant. This is analogous to the well-known shrinkage estimation for multiple point estimation, which was introduced by Stein's paradox (Stein, 1956, 1981). By using a simple procedure which is well motivated by the ODP lemma, I show that the uniformly most powerful (UMP) unbiased test of a normal mean is no longer UMP in the multiple-testing setting, i.e. a procedure that is considered to be optimal for a single test may no longer be optimal for performing several tests simultaneously.

Analogous to the information that is required for the NP lemma, the ODP assumes that the true probability densities corresponding to each significance test are known. In practice, these probability densities will typically be unknown. Therefore, an approach is briefly illustrated to show that the ODP may be applied to derive practical multiple-testing procedures. An application of the ODP to genomics is fully developed and applied in a complementary paper (Storey *et al.*, 2005, 2006).

This paper is organized as follows: Section 2 develops the ODP theory; Section 3 applies this theory and develops a simple implementation of the ODP for the problem of testing multiple normal means for equality to 0; Section 4 proposes a general approach to estimating the ODP; Section 5 shows connections between the ODP and several other concepts in statistics, including false discovery rates, the Bayesian approach to hypothesis testing, Stein's paradox and shrinkage estimation.

2. Optimal discovery procedure

The core idea behind the ODP is that, in comparison with the NP statistic, the ODP uses all of the relevant information across tests when determining the relative significance of each test (i.e. the ordering in which the tests should be called significant). When performing multiple tests with the general goal of discovering several true positive without incurring too many false positive results, it is intuitively clear that we are better off calling tests significant with a common and persistent signal structure rather than a test with a rare signal structure. The ODP borrows strength across tests to capture these common patterns of signal in the optimal manner. This idea is summarized graphically in Fig. 1 of Storey *et al.* (2006); see also Storey (2005).

As is discussed below, both the NP and the ODP approaches satisfy some multiple-testing optimality criterion. However, the constraints under which the ODP is optimal are those which are usually encountered in practice, leading to the demonstration that methods based on the NP approach tend to underperform those based on the ODP approach.

A rigorous development of the optimal multiple-testing procedure involves several components:

(a) defining the optimality goal;
(b) properly constraining the set of procedures over which the optimality is to be found;
(c) deriving the procedure that achieves this optimality.

In the following subsections, I address each of these components in the order listed above.

2.1. Optimality goal

In forming an optimality goal, recall that each test has a potential type I error rate or power level. For any true null hypothesis, the type I error rate is also the expected number of false positive results from that test. Therefore, the sum of type I error rates across all true null hypotheses is the expected number of false positive results, EFP. Similarly, the sum of powers across all true alternative hypotheses is the expected number of true positive results, ETP. EFP and ETP for any given set of tests called significant are focused on the overall 'discovery rate', as motivated in the original paper introducing false discovery rates (Soric, 1989). Therefore, as a multivariate extension of maximizing power for each fixed type I error rate, the optimality goal that I propose is to maximize the expected number of true positive results, ETP, for each fixed expected number of false positive results, EFP.

Definition 1. A multiple-testing procedure is defined here to be optimal if it maximizes the expected number of true positive results, ETP, at each fixed level of expected number of false positive results, EFP. Both quantities are calculated among the true configuration of hypotheses, i.e. ETP is calculated among the subset of hypotheses with a true alternative, and EFP among the subset of hypotheses with a true null hypothesis.

A key property to the definition of EFP and ETP is that each is calculated under the *true* configuration of hypotheses. Therefore EFP is calculated among a subset of the hypotheses (those with a true null hypothesis) and ETP among the complementary subset. Another option would be to calculate EFP assuming that all null hypotheses are true and ETP assuming that all alternative hypotheses are true (Spjotvoll, 1972). This is problematic because it does not represent the underlying reality (no hypothesis can simultaneously have a true null and a true alternative hypothesis). It also prevents any connection being made to the false discovery rate FDR, which is calculated under the true configuration of hypotheses.

The proposed optimality criterion is directly related to optimality in terms of false discovery rates and misclassification rates. For false discovery rates the key observation is that

$$\text{FDR} \approx \frac{\text{EFP}}{\text{EFP} + \text{ETP}}, \tag{1}$$

where the approximate equality is sometimes an exact equality (Section 5.1). An exact equality exists for large numbers of tests with certain convergence properties (Storey *et al.*, 2004), under Bayesian mixture model assumptions (Storey, 2003) and under alternative definitions of FDR (Benjamini and Hochberg, 1995; Storey, 2003). The important point is that FDR may be interpreted and characterized in terms of EFP and ETP.

It is shown in Section 5.1 that, if ETP is maximized for each fixed EFP level, then the proportion of 'missed discoveries' (Genovese and Wasserman, 2002; Taylor *et al.*, 2005) is minimized for each fixed FDR level, i.e. achieving optimality in terms of EFP and ETP is equivalent to achieving optimality in terms of FDR. (This exact statement is sometimes an approximate one, depending on the relationship between the two quantities in equation (1).) Optimality in terms of misclassification error is also achieved under the goal of maximizing ETP for each fixed EFP level. Therefore, even though I derive the theory below in terms of EFP and ETP, it can also be applied in terms of these other error measures. Furthermore, it may be argued that EFP and ETP are the more fundamental units of a number of error measures, making them an appropriate choice for defining optimality.

2.2. Single-thresholding procedures

To characterize the various classes of procedures over which this proposed optimality may be obtained, it is necessary to introduce some rigorous notation. Assume that m multiple-hypothesis tests are performed based on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. In general it should be assumed that the data sets are random vectors defined on a common probability space. For simplicity we can think of the data sets as being composed of n observations each, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^{\mathrm{T}}$ from $\mathbb{R}^n$.

Testing procedures are usually defined in terms of a critical function (Lehmann, 1986), which is a mapping from the data to the probability that the null hypothesis is rejected. Since randomized rejections are uncommon, we can think of the critical function as being an indicator function of whether the test is significant or not. At a finer level of detail, a 'significance thresholding function' may be defined as follows, which unravels the indicator critical function into an interpretable function of the data.

Definition 2. A significance thresholding function is defined to be a function $\mathcal{S} : \mathbb{R}^n \to [0, \infty)$ such that the null hypothesis is rejected if and only if $\mathcal{S}(\mathbf{x}) \geqslant \lambda$ for some λ chosen to satisfy an acceptable level of significance.

In this set-up, the critical function is now easily defined as the indicator function $\mathbf{1}\{\mathcal{S}(\mathbf{x}) \geqslant \lambda\}$. When multiple tests are performed, it may be that a separate significance thresholding function is defined for each test. In such a case, the λ-threshold would be applied to $\mathcal{S}_i(\mathbf{x}_i)$, where $\mathcal{S}_i$ is the test-specific thresholding function. A 'single-thresholding procedure' (STP) is defined to be a multiple-testing procedure where a single significance thresholding function is applied to all tests.

Definition 3. An STP is defined to be a multiple-testing procedure equivalent to applying a single significance thresholding function $\mathcal{S}$ and cut-off λ to every test, i.e. each test i is significant if and only if $\mathcal{S}(\mathbf{x}_i) \geqslant \lambda$ for a given $\mathcal{S}$ and λ.

The advantages of defining and employing significance thresholding functions rather than critical functions should now be clear. In Section 1, I discussed two steps that are involved in any multiple-testing procedure:

(a) determining the order in which the tests should be called significant and
(b) choosing an appropriate significance cut-off somewhere along this ordering.

Thresholding functions allow these two steps to be clearly separated. The evaluated significance thresholding functions (e.g. $\mathcal{S}_i(\mathbf{x}_i)$ or $\mathcal{S}(\mathbf{x}_i)$) produce a positive score for each test, which can be used to rank the tests from most significant to least. The value of λ is chosen to control a particular error rate.

As an example of an STP, suppose that a standard two-sided t-test is applied to each $\mathbf{x}_i$. The statistic is

$$\mathcal{S}(\mathbf{x}_i) = \left| \frac{\bar{x}_i}{s_i/\sqrt{n}} \right|,$$

where $\bar{x}_i$ is the sample mean and s_i is the sample standard deviation of $\mathbf{x}_i$. Since the sample mean $\bar{x}_i$ and standard error $s_i/\sqrt{n}$ are functions of $x_{i1}, \ldots, x_{in}$, it follows that $\mathcal{S}(\mathbf{x}_i)$ is also a function of these data. Also, each test has the same null distribution. Therefore, test i is called significant if and only if $\mathcal{S}(\mathbf{x}_i) \geqslant \lambda$, making the standard two-sided t-test an STP for multiple tests. Even if there are different numbers of observations for each test, then the multiple tests can still be written as an STP. Specifically, we can apply the significance thresholding function giving the p-value for each test:

$$\mathcal{S}(\mathbf{x}_i) = 1 - 2T_{n_i-1}^{-1}\left(-\left| \frac{\bar{x}_i}{s_i/\sqrt{n_i}} \right| \right),$$

where n_i is the number of observations for test i and T_{n_i-1} is the cumulative t-distribution function with $n_i - 1$ degrees of freedom.

A large class of multiple-testing procedures—those that are usually encountered in practice—can be written as an STP. Suppose that a multiple-testing procedure is performed where

(a) the data for each test follow a common family of distributions, where all possible differences in parameters are unknown,
(b) the null and alternative hypotheses are identically defined for each test and
(c) the same procedure is applied to each test in estimating unknown parameter values.

Any differences between test statistics are then due to differences in their data, not prior knowledge about the significance tests. In this case, the relative significance among the tests is based only on the data from each test, implying that any such testing procedure can be written as an STP. This characterization applies to generalized likelihood ratio statistics, t-statistics, F-statistics, p-values obtained from a common function (as above), etc.

This observation can be formalized and generalized in the following lemma, which basically states that any multiple-testing procedure that is invariant to the labelling of the tests is equivalent to an STP. The proof of this lemma is straightforward, so it is omitted.

Lemma 1. Suppose that m significance tests are performed, based on respective observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. For a given multiple-testing procedure, let ψ_i be the outcome of test i, where $\psi_i = 1$ if test i is significant and $\psi_i = 0$ otherwise. Moreover, let $r : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, m\}$ be any bijective function. If performing the multiple-testing procedure on data sets $\mathbf{x}_{r(1)}, \mathbf{x}_{r(2)}, \ldots, \mathbf{x}_{r(m)}$ results in outcomes $\psi_{r(1)}, \psi_{r(2)}, \ldots, \psi_{r(m)}$ for all possible bijective functions r (i.e. the procedure is invariant to relabelling of the test indices), then this multiple-testing procedure is equivalent to an STP.

Even though it is possible to consider *in theory* applying a different significance thresholding function to each test, an STP is typically the only one available option *in practice*. This brings

us to a key point of the paper. Most multiple-testing methods are performed on the basis of *p*-values that have been calculated from procedures aimed at approximating the NP approach or one of its variants. However, the NP procedure yields an optimality under conditions that are not possible in practice (i.e. that each test is treated differently on the basis of prior knowledge about each test's particular densities). This begs the question of what is optimal under the constraints that are actually present in practice, namely that we are constrained to perform an STP. This is the main motivation for the formulation of the ODP in terms of an STP. This is further motivated by the fact that an *estimate* of an ODP may substantially outperform an *estimate* of the NP procedure, which also turns out to be an STP (Section 3.3 and Storey *et al.* (2006)).

2.3.   Definition and derivation of the optimal discovery procedure

Motivated by these points, I define the ODP to be the procedure that maximizes ETP for each fixed EFP among all STPs.

Definition 4. The ODP is defined to be the multiple-testing procedure that maximizes ETP for each fixed EFP among all STPs.

Given these definitions, it is now straightforward to derive the ODP significance thresholding function and to prove its optimality.

Lemma 2 (ODP). Suppose that m significance tests are performed on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, where significance test i has null density f_i and alternative density g_i, for $i = 1, \ldots, m$. Without loss of generality suppose that the null hypothesis is true for tests $i = 1, 2, \ldots, m_0$, and the alternative is true for $i = m_0 + 1, \ldots, m$. The following significance thresholding function defines the ODP:

$$\mathcal{S}_{\mathrm{ODP}}(\mathbf{x}) = \frac{g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \ldots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_{m_0}(\mathbf{x})}. \tag{2}$$

Null hypothesis i is rejected if and only if $\mathcal{S}(\mathbf{x}_i) \geqslant \lambda$ for some $0 \leqslant \lambda < \infty$. For each fixed λ, this procedure yields the maximum number of expected true positive results ETP among all simultaneous thresholding procedures that have an equal or greater number of expected false positive results EFP.

Proof. Let $\Gamma_\lambda = \{\mathbf{x} : \mathcal{S}_{\mathrm{ODP}}(\mathbf{x}) \geqslant \lambda\}$ be the significance region for the ODP applied at cut-off λ. EFP and ETP of any general significance region Γ are

$$\mathrm{EFP}(\Gamma) = \int_\Gamma f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} + \ldots + \int_\Gamma f_{m_0}(\mathbf{x})\, \mathrm{d}\mathbf{x},$$

$$\mathrm{ETP}(\Gamma) = \int_\Gamma g_{m_0+1}(\mathbf{x})\, \mathrm{d}\mathbf{x} + \ldots + \int_\Gamma g_m(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

The goal is then to show that, for any Γ' such that $\mathrm{EFP}(\Gamma') \geqslant \mathrm{EFP}(\Gamma_\lambda)$, it is the case that $\mathrm{ETP}(\Gamma') \leqslant \mathrm{ETP}(\Gamma_\lambda)$. This is equivalent to showing that $\mathrm{EFP}(\Gamma')/m_0 \geqslant \mathrm{EFP}(\Gamma_\lambda)/m_0$ implies $\mathrm{ETP}(\Gamma')/(m - m_0) \leqslant \mathrm{ETP}(\Gamma_\lambda)/(m - m_0)$. For this, define $\bar{f} = [\Sigma_{i=1}^{m_0} f_i]/m_0$ and $\bar{g} = [\Sigma_{i=m_0+1}^{m} g_i]/(m - m_0)$; it is easily verified that these functions each integrate to 1. It then follows that

$$\frac{\text{EFP}(\Gamma)}{m_0} = \frac{\int_\Gamma f_1(\mathbf{x})\,\mathrm{d}\mathbf{x} + \ldots + \int_\Gamma f_{m_0}(\mathbf{x})\,\mathrm{d}\mathbf{x}}{m_0}$$

$$= \int_\Gamma \frac{f_1(\mathbf{x}) + \ldots + f_{m_0}(\mathbf{x})}{m_0}\,\mathrm{d}\mathbf{x}$$

$$= \int_\Gamma \bar{f}(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

$$\frac{\text{ETP}(\Gamma)}{m - m_0} = \int_\Gamma \bar{g}(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Since these functions each integrate to 1, the NP lemma can be invoked as a mathematical optimization result among this class of functions. (It should be pointed out that no frequentist probabilistic interpetation of $\bar{f}$ and $\bar{g}$ is used in this proof.) According to the NP lemma, the significance regions based on $\bar{g}(\mathbf{x})/\bar{f}(\mathbf{x})$ are optimal for maximizing $\int_\Gamma \bar{g}(\mathbf{x})\,\mathrm{d}\mathbf{x}$ for each fixed level of $\int_\Gamma \bar{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}$. However,

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{(m - m_0)\,\bar{g}(\mathbf{x})}{m_0\,\bar{f}(\mathbf{x})},$$

making the two thresholding functions equivalent. Therefore, the ODP maximizes ETP for each fixed EFP. $\qquad\square$

Note that no restrictions are placed on the probabilistic dependence between the tests, i.e. the lemma holds under arbitrary dependence. Below, I extend the ODP to the case where the status of each significance test is random, producing a rule that uses both the null and the alternative densities of every test. Exceptions to the STP constraint occur in practice when prior information about the tests is known that distinguishes them. In such a case, we can nevertheless arrange the tests into the largest groups possible where an STP is necessary and apply the ODP to each group separately.

It should also be noted that it is possible to give each test a relative weight, where it is desired that each test does not make an equal contribution to EFP or ETP. For example, if it were known that a set of tests were related, then these could be weighted so that they essentially count as a single false positive or a single true positive result. As another example, if prior information on certain tests indicates that these are more important, then they can be given higher weight. In general, if test i is given relative weight w_i, then the ODP can be generalized to maximize this weighted version of ETP in terms of the weighted version of EFP. The only difference in the formula above is that each f_i or g_i is multiplied by w_i. A positive result from test i then contributes w_i to the EFP or ETP. The proof of this easily follows from the proof of lemma 2.

The ODP as presented above requires one to know the true distribution corresponding to each significance test, but these will not be known in practice. However, these may be estimated because the data that are observed for each test do come from their true distribution, whether it be a null or alternative distribution. Therefore, we do not necessarily need to know the status of each test to estimate the ODP effectively.

2.4. Previous work on optimal multiple testing

Several earlier references have considered the optimality of familywise error rate controlling procedures in terms of p-values. See Shaffer (1995) for a review of optimality in terms of the familywise error rate, as well as the more recent Lehmann *et al.* (2005). Spjotvoll (1972) considered the problem of maximizing the expected number of true positive results—assuming that

all alternative hypotheses are true—with respect to a fixed expected number of false positive results—assuming that *all* null hypotheses are true. This is a potentially problematic optimality because either the null or the alternative hypothesis is true for each test, but not both. Furthermore, he allowed each test to employ a different critical function (or significance thresholding function in our terminology). Therefore, Spjotvoll (1972) proposed both a different optimality criterion and a different class of procedures from what I propose here. In the Spjotvoll (1972) set-up, the optimal procedure turns out to be directly thresholding the traditional NP statistic.

It should be noted that any implementation of Spjotvoll's (1972) procedure reduces to estimating the NP statistic for each test, which is not different from what is already done in practice. The maximum likelihood implementation of the procedure is exactly equivalent to the standard generalized likelihood ratio test. For example, when performing multiple tests on the mean of a normal distribution (with unknown variance), the maximum likelihood implementation of Spjotvoll's (1972) procedure algebraically reduces to directly thresholding the absolute *t*-statistics, which is the standard practice. It is shown in Section 3.3 and Storey *et al.* (2006) that the generalized likelihood ratio test substantially underperforms a maximum likelihood implementation of the ODP.

3.   Testing several normal means for equality to zero

In this section, I derive the exact ODP as well as a method for estimating it, when testing several normal means for equality to 0. This example is purposely kept simple to elucidate some of the operating characteristics of the ODP. In following sections, I describe an implementation of the ODP for testing genes for differential gene expression, which reflects a more typical situation.

3.1.   Testing several normal means

Suppose that we observe $z_i \sim N(\mu_i, 1)$ for eight tests ($i = 1, \ldots, 8$), where significance test i is $\mu_i = 0$ *versus* $\mu_i \neq 0$. (In the notation of the previous section, it follows that $\mathbf{x}_i \equiv z_i$, where here $n = 1$ and $m = 8$.) The conventional approach to testing multiple hypotheses is to apply the best procedure for single tests to each one individually. The UMP unbiased test, the generalized likelihood ratio test, the maximum likelihood implementation of Spjotvoll (1972) and the maximum likelihood implementation of the NP approach are all equivalent to fixing a cut-off c and rejecting all null hypotheses with $|z_i| \geqslant c$. Since each test has the same null distribution, this is equivalent to calculating a *p*-value for each test and then forming the equivalent *p*-value threshold. Even though the true NP rule allows the significance rule to differ between tests (i.e. $z_i \leqslant c$ or $z_i \geqslant c$, depending on the sign of the alternative parameter value), the estimated implementable version leads to applying the same significance thresholding function to each test (i.e. an STP). As lemma 1 shows, STPs are virtually ubiquitous for actual implementations of multiple-testing procedures, which is the case here.

3.2.   Optimal discovery procedure for testing several normal means

The ODP, which is the optimal STP, is straightforward to derive in this simple example. First, assume that we know the *true* values of the means $\mu_1, \mu_2, \ldots, \mu_8$. This implies that we know whether each null hypothesis is true or false and, if it is false, we know the alternative distribution. Table 1 provides this information for the eight significance tests. Let

$$\phi(z; \mu) = \frac{1}{\sqrt{(2\pi)}} \exp\left\{\frac{-(z - \mu)^2}{2}\right\}$$

be the density of an $N(\mu, 1)$ random variable. According to lemma 2 above, the ODP for this case is based on the significance thresholding function

Table 1. Example of eight significance tests performed on the mean of a normal distribution, each based on a single observation z_i from the $N(\mu_i, 1)$ distribution, $i = 1, 2, \ldots, 8$†

Significance test i	Alternative value of μ_i	True value of μ_i	Observed datum z_i	ODP significance rank	Estimated ODP significance rank	UMP unbiased significance rank
1	−3	0	1.0	4	4	4
2	−2	−2	−2.3	3	3	2
3	−2	0	−0.02	6	6	8
4	−1	0	−0.4	8	8	6
5	1	1	0.5	5	5	5
6	2	2	2.2	2	2	3
7	2	0	−0.1	7	7	7
8	3	3	3.4	1	1	1

†For each test, the null hypothesis is $\mu_i = 0$. The second column gives the value of μ_i under the alternative hypothesis if it were known. The third column is the *true* value of μ_i. The fourth column is the observation for each test, z_i. The fifth column gives the ranking of the tests in terms of their significance according to the ODP. The sixth column gives the ranking based on the estimated ODP, which tests the alternative hypothesis $\mu_i \neq 0$. The seventh is the significance ranking based on the univariate UMP unbiased test against the alternative $\mu_i \neq 0$, which uses the significance thresholding rule $|z_i| \geqslant c$.

$$\mathcal{S}_{\mathrm{ODP}}(z) = \frac{\phi(z; \mu_2) + \phi(z; \mu_5) + \phi(z; \mu_6) + \phi(z; \mu_8)}{\phi(z; \mu_1) + \phi(z; \mu_3) + \phi(z; \mu_4) + \phi(z; \mu_7)}. \tag{3}$$

With the actual values of μ_i plugged in, this function is

$$\mathcal{S}_{\mathrm{ODP}}(z) = \frac{\phi(z; -2) + \phi(z; 1) + \phi(z; 2) + \phi(z; 3)}{\phi(z; 0) + \phi(z; 0) + \phi(z; 0) + \phi(z; 0)}. \tag{4}$$

For a fixed λ, null hypothesis i is rejected if $\mathcal{S}_{\mathrm{ODP}}(z_i) \geqslant \lambda$, $i = 1, 2, \ldots, 8$. The threshold λ would be chosen to obtain a certain EFP level (or perhaps an FDR level). The key is that this single significance thresholding rule is applied to every test, and it involves the true densities for every test.

The intuition behind this procedure is that the significance of each observed z_i is made relatively higher as evidence builds up that there are multiple true positive results that are likely to have similar values. Supposing that z_i corresponds to a true alternative hypothesis, it should be the case that z_i is close to μ_i with high probability, making the contribution from its density

$$\frac{1}{\sqrt{(2\pi)}} \exp\left\{\frac{-(z_i - \mu_i)^2}{2}\right\}$$

to $\mathcal{S}_{\mathrm{ODP}}(z_i)$ substantial. However, if there are other true alternatives with $\mu_j \approx \mu_i$, then the likelihood of z_i under

$$\frac{1}{\sqrt{(2\pi)}} \exp\left\{\frac{-(z_i - \mu_j)^2}{2}\right\}$$

will also make a substantial contribution to $\mathcal{S}_{\mathrm{ODP}}(\mu_j)$. Since the goal is to maximize ETP for each fixed EFP, it makes sense to increase the relative significance of a particular test if there are other hypotheses with similar signal that also are well distinguished from the true null hypotheses. If one particular true alternative has mean μ_i which is very different from the others, then $\mathcal{S}_{\mathrm{ODP}}(z_i)$ will behave like its NP statistic because the contribution from the other densities will be negligible.
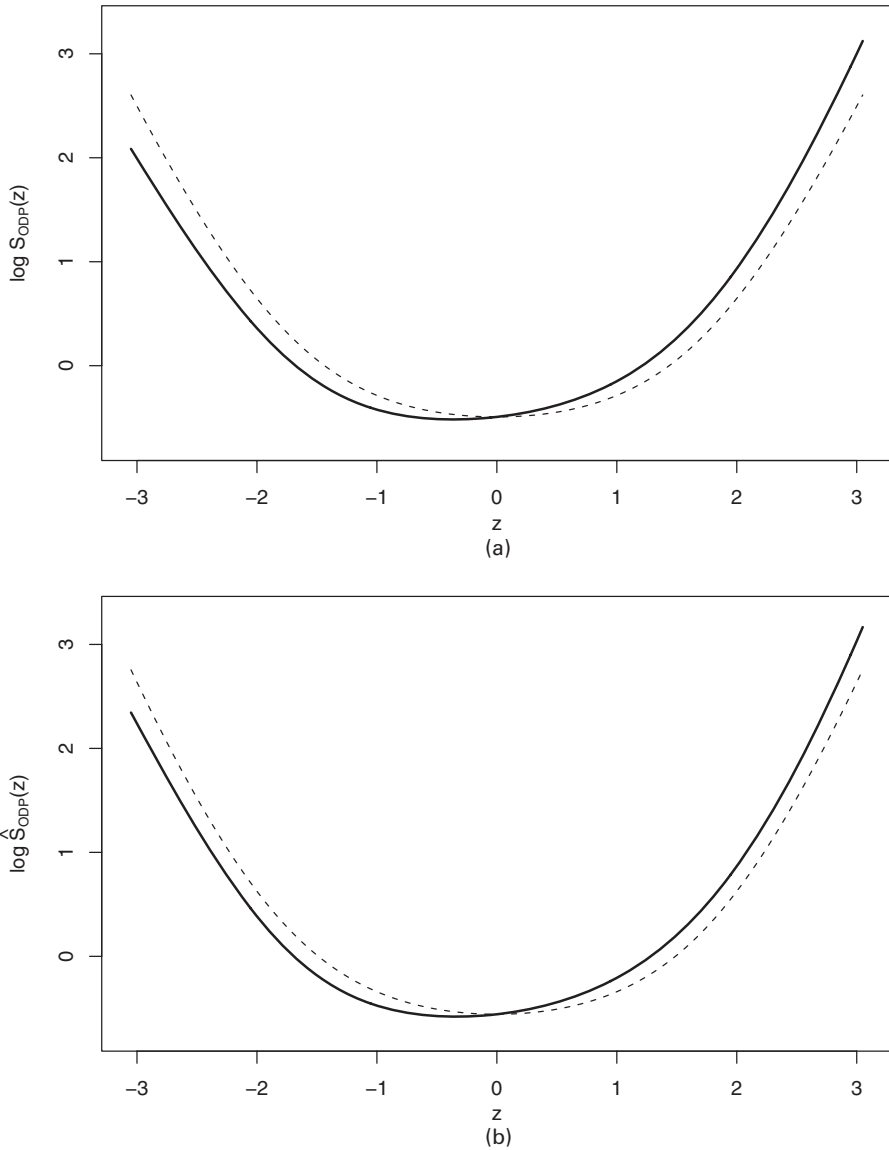
Fig. 1. True and estimated ODP significance thresholding functions for the multiple normal means that are summarized in Table 1: (a) significance thresholding function $\mathcal{S}_{\mathrm{ODP}}(z)$ *versus* observed statistic z for the true ODP (———) and a symmetric thresholding function (-------), which is equivalent to the optimal procedure when performing a single test; (b) analogous plot to (a) except the estimated ODP significance thresholding function as defined in equation (6) is shown

Fig. 1(a) shows a plot of $\mathcal{S}_{\mathrm{ODP}}(z)$ over a range of z-values for this particular example. It can be seen that $\mathcal{S}_{\mathrm{ODP}}(z)$ captures the asymmetry in the signal among the true alternative hypotheses. The true alternative mean values are -2, 1, 2 and 3, so the relative significance is increased among the positive values of z. Table 1 gives an example of realized values of $z_1, z_2, \ldots, z_8$. From Table 1 it can be seen that the statistic for test 6 is greater than that for test 2, i.e. $\mathcal{S}_{\mathrm{ODP}}(2.2) > \mathcal{S}_{\mathrm{ODP}}(-2.3)$. The UMP unbiased procedure uses a symmetric significance rule, which is also shown in Fig. 1(a). Under this rule, test 2 with $|z_2| = 2.3$ would be considered

more significant than test 6 with $|z_6| = 2.2$. The ODP significance rule ranks test 6 higher than test 2 because the true alternative means 1, 2 and 3 all contribute substantially to $\mathcal{S}_{\text{ODP}}(2.2)$.

3.3. Estimated optimal discovery procedure for testing several normal means

It seems paradoxical to define the ODP under the assumption that the truth about each hypothesis is known. However, this theoretical result allows for a straightforward approach to estimating the ODP, requiring no estimation beyond what is required for estimating the NP rule for a single significance test. The goal is essentially to estimate the significance thresholding function that is given in equation (3). In practice, every μ_i would be unknown, and the fact that there are four true null hypotheses in the denominator would also be unknown. However, as it turns out, all we need to do is to estimate the true μ_i for each test; it is not necessary to distinguish the true and false null hypotheses.

Some rearranging of the function in equation (3) yields a form that is estimable without requiring any explicit separation of true and false null hypotheses. Note that the threshold $\mathcal{S}_{\text{ODP}}(z) \geqslant \lambda$ is exactly equivalent to $4\{\mathcal{S}_{\text{ODP}}(z) + 1\} \geqslant 4(\lambda + 1)$. It can easily be calculated that

$$4\{\mathcal{S}_{\text{ODP}}(z) + 1\} = \sum_{i=1}^{8} \phi(z; \mu_i) / \phi(z; 0),$$

where the denominator follows from the fact that all true null hypotheses follow the $N(0, 1)$ distribution. Therefore, the ODP can equivalently be performed with the thresholding function,

$$\mathcal{S}_{\text{ODP}}^*(z) = \sum_{i=1}^{8} \phi(z; \mu_i) / \phi(z; 0). \tag{5}$$

Different thresholds would have to be applied to equations (3) and (5) to obtain the same results, but this is irrelevant for actual applications where the threshold would be chosen empirically (see Section 4). Even though the derivation in this example uses the fact that $\mu_1 = \mu_3 = \mu_4 = \mu_7 = 0$, the thresholding function in equation (5) will result for any combination of true and false null hypotheses.

We may estimate $\Sigma_{i=1}^{8} \phi(z; \mu_i)$ by estimating each individual μ_i and then forming a plug-in estimate of the total sum. Each true mean can be estimated by the data that are observed for its respective test. For example, we can set $\hat{\mu}_j = z_j$ and substitute these into $\mathcal{S}_{\text{ODP}}^*(z)$, producing an estimated version of the ODP rule:

$$\hat{\mathcal{S}}_{\text{ODP}}^*(z) = \sum_{i=1}^{m} \phi(z; \hat{\mu}_i) / \phi(z; 0). \tag{6}$$

Therefore, for a fixed λ, test j is called significant if

$$\sum_{i=1}^{m} \phi(z_j; \hat{\mu}_i) / \phi(z_j; 0) \geqslant \lambda.$$

Table 1 gives a simulated example of realized values of $z_1, z_2, \ldots, z_8$. Substituting these values into the estimate, we obtain

$$\hat{\mathcal{S}}_{\text{ODP}}^*(z) = \{\phi(z; 1.0) + \phi(z; -2.3) + \phi(z; -0.02) + \phi(z; -0.4) + \phi(z; 0.5) + \phi(z; 2.2)$$
$$+ \phi(z; -0.1) + \phi(z; 3.4)\} / \phi(z; 0).$$

Fig. 1(b) shows a plot of $\hat{\mathcal{S}}_{\text{ODP}}^*(z)$ *versus* z as well as a plot of the symmetric thresholding function, which is equivalent to the single-test UMP unbiased rule. It can be seen that the asymmetry in the distribution of the μ_is is captured in this estimated significance thresholding function, and that it is similar to the true ODP significance thresholding function that is shown in Fig. 1(a).
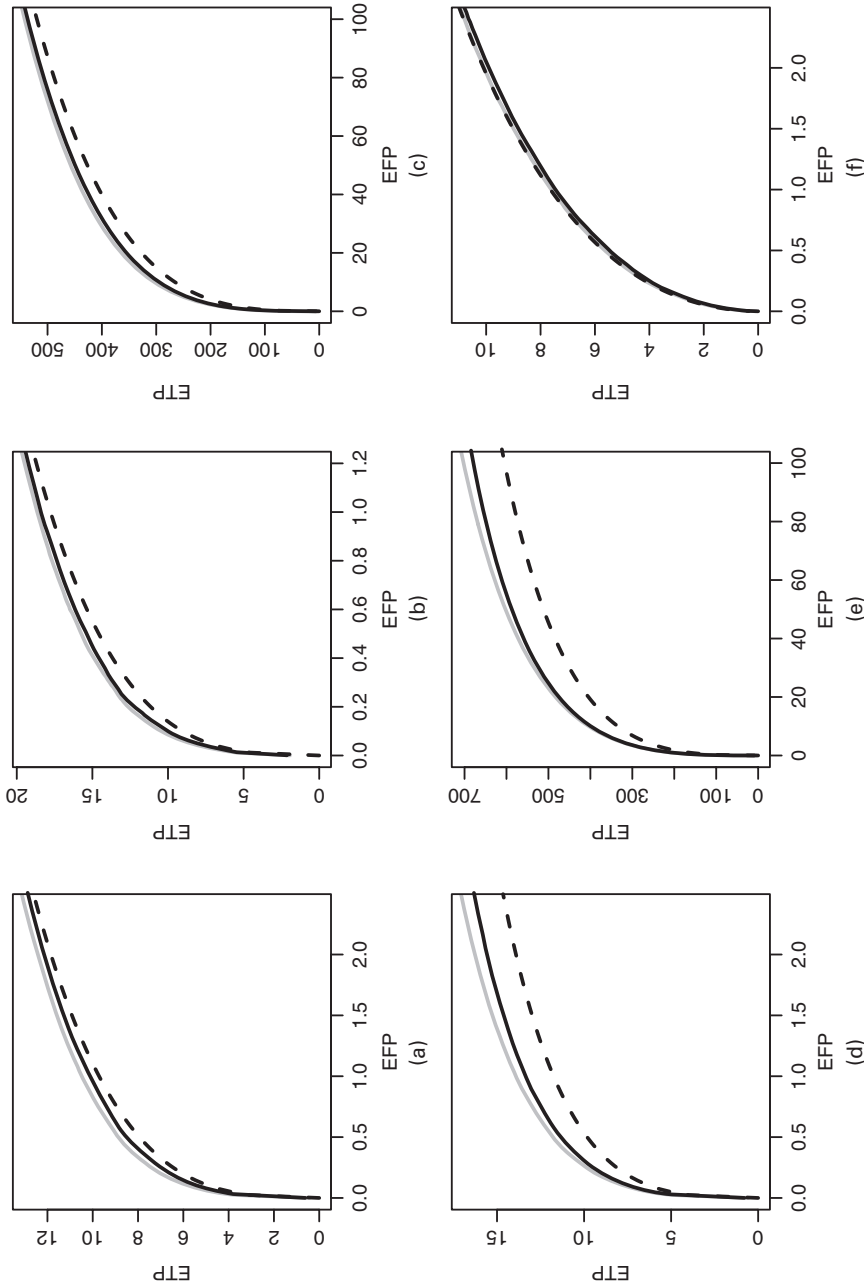
Fig. 2. Comparison in terms of ETP *versus* EFP between the true ODP (———), estimated ODP (———), and conventional procedure (– – –) (in this case, the conventional procedure is both the UMP unbiased procedure and the generalized likelihood ratio test; it can be seen that the estimated ODP outperforms the conventional procedure in all cases except where the signal is perfectly symmetric, in which case the true ODP is equivalent to the conventional procedure; for each panel, multiple significance tests of $\mu_i = 0$ *versus* $\mu_i \neq 0$ were performed based on a single observation $z_i \sim N(\mu_i, 1)$; for each test a number of alternative means were assigned, and there are equal proportions of each one among the false null hypotheses): (a) 48 tests, 24 true nulls, alternative means -1, 1, 2 and 3; (b) 48 tests, 12 true nulls, alternative means -1, 1, 2 and 3; (c) 2000 tests, 1000 true nulls, alternative means -1, 1, 2 and 3; (d) 48 tests, 24 true nulls, alternative means 1, 2 and 3; (e) 2000 tests, 1000 true nulls, alternative means 1, 2 and 3; (f) 48 tests, 24 true nulls, alternative means -2, -1, 1 and 2

The performance of this ODP estimate was compared with the 'conventional procedure', which rejects null hypotheses with $|z_i| \geqslant c$ for a given threshold c and is equivalent to the UMP unbiased test, generalized likelihood ratio test, the maximum likelihood implementation of Spjotvoll (1972) and the maximum likelihood implementation of the NP statistic. Fig. 2 shows a comparison between the true ODP, the estimated ODP and the conventional procedure in terms of ETP for each fixed EFP level. The EFP- and ETP-values were calculated via simulation, where the number of iterations was sufficiently large that the Monte Carlo error is negligible.

The comparisons were made over a variety of configurations of true alternative means and numbers of tests performed. It can be seen that the true ODP is always the best performer, as the theory implies it should be. The larger the number of tests, and the larger the proportion of true alternative to true null hypotheses, the closer the estimated ODP is to the true ODP in terms of performance. The estimated ODP rule outperforms the conventional procedure in all cases where the alternative means are not arranged in a perfectly symmetric fashion around zero. What is meant by perfect symmetry is that, if there is a true alternative mean of value μ_i, then there is another true alternative mean of value $-\mu_i$. When there is perfect symmetry, it can be shown that the true ODP and the conventional procedure are equivalent; in this case, the estimated ODP is slightly inferior to the other two because it is a noisy version of the perfect symmetry rule.

4. Estimating the optimal discovery procedure in a general setting

Two simplifying properties of the above ODP estimate are that

   (a) every test has the same null distribution and
   (b) there are no nuisance parameters.

It is possible to estimate the ODP in more general scenarios, which I derive in this section. This involves formulating an estimate when every test may have a different probability distribution, including the possibility that every null distribution is different. The approach is related to that above, although some additional complications must be properly handled. Nevertheless, this shows that the theory that was developed above can be applied to substantially more complicated scenarios. Storey *et al.* (2006) fully developed a method for estimating the ODP and apply it to the problem of identifying differentially expressed genes in deoxyribonucleic acid (DNA) microarray experiments.

4.1. An equivalent optimal discovery procedure significance thresholding function

Recall the notation of Section 2 and assumptions of lemma 2. The ODP significance thresholding function is

$$\mathcal{S}_{\mathrm{ODP}}(\mathbf{x}) = \frac{g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \ldots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_{m_0}(\mathbf{x})}.$$

By defining $\mathcal{S}_{\mathrm{ODP}}^* = 1 + \mathcal{S}_{\mathrm{ODP}}$, we have

$$\mathcal{S}_{\mathrm{ODP}}^*(\mathbf{x}) = \frac{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_{m_0}(\mathbf{x}) + g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \ldots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_{m_0}(\mathbf{x})}. \tag{7}$$

Since $\mathcal{S}_{\mathrm{ODP}}^*(\mathbf{x}) = 1 + \mathcal{S}_{\mathrm{ODP}}(\mathbf{x})$, these produce the exact same STP, where a threshold of λ applied to $\mathcal{S}_{\mathrm{ODP}}(\mathbf{x})$ is equivalent to a threshold of $1 + \lambda$ applied to $\mathcal{S}_{\mathrm{ODP}}^*(\mathbf{x})$.

Corollary 1 (ODP equivalent significance thresholding function). Under the assumptions of lemma 2, applying the significance thresholding function $\mathcal{S}_{\mathrm{ODP}}^*$ of equation (7) as an STP is

equivalent to applying $\mathcal{S}_{\text{ODP}}$ of equation (2) in lemma 2. Therefore, $\mathcal{S}_{\text{ODP}}^{*}$ equivalently defines the ODP.

The significance thresholding function $\mathcal{S}_{\text{ODP}}^{*}$ is estimated rather than $\mathcal{S}_{\text{ODP}}$ because of the particular strategy that is proposed.

4.2. A maximum likelihood implementation of the optimal discovery procedure

A parametric approach can be taken to estimate the ODP significance thresholding function, motivated by the generalized likelihood ratio test for single significance tests. In general, the density functions f_i and g_i will be indexed by a set of parameters (e.g. the mean and variance of a normal distribution). For each test $i = 1, \ldots, m$, let $\hat{f}_i$ and $\hat{g}_i$ be well-behaved estimates of f_i and g_i respectively. For example, we can define $\hat{f}_i$ as the version of f_i with all the unknown parameters replaced by their maximum likelihood estimates under the constraints of the null hypothesis, and $\hat{g}_i$ as the analogous estimate given by the unconstrained maximum likelihood estimates.

In single-hypothesis testing, the NP procedure for test i is based on $g_i(\mathbf{x}_i)/f_i(\mathbf{x}_i)$, and it can be estimated by the generalized likelihood ratio statistic $\hat{g}_i(\mathbf{x}_i)/\hat{f}_i(\mathbf{x}_i)$ (Lehmann, 1986). The proposed approach for estimating the ODP significance thresholding function builds on this strategy.

Since g_i is estimated in an unconstrained fashion, the $\hat{g}_i$ will serve as an estimate for the true density, whether it be f_i or g_i. For *true* null hypotheses $i = 1, \ldots, m_0$, the maximum likelihood parameters defining $\hat{f}_i$ and $\hat{g}_i$ are both consistent estimates of the actual values of f_i as the number of observations n grows to ∞. Likewise, $\hat{g}_i$ is composed of consistent parameter estimates of g_i for false null hypotheses $i = m_0 + 1, \ldots, m$. Therefore, $\hat{g}_1 + \ldots + \hat{g}_m$ can be used to estimate the numerator of equation (7), where it is now unnecessary to be able to distinguish between true and false null hypotheses.

We call the following a 'canonical plug-in estimate' of the ODP significance thresholding function:

$$\hat{\mathcal{S}}_{\text{ODP}}^{*}(\mathbf{x}) = \frac{\hat{g}_1(\mathbf{x}) + \ldots + \hat{g}_{m_0}(\mathbf{x}) + \hat{g}_{m_0+1}(\mathbf{x}) + \ldots + \hat{g}_m(\mathbf{x})}{\hat{f}_1(\mathbf{x}) + \ldots + \hat{f}_{m_0}(\mathbf{x})}. \tag{8}$$

However, the denominator of this statistic still requires specification of the true null hypotheses. One general approach is to approximate the canonical plug-in estimate by estimating which null densities should be included in the denominator of the statistic. Let $\hat{w}_i = 1$ if $\hat{f}_i$ is to be included in the denominator, and $\hat{w}_i = 0$ otherwise. The estimate of the ODP statistic is then

$$\hat{\mathcal{S}}_{\text{ODP}}^{*}(\mathbf{x}) = \sum_{i=1}^{m} \hat{g}_i(\mathbf{x}) \Big/ \sum_{i=1}^{m} \hat{w}_i \, \hat{f}_i(\mathbf{x}). \tag{9}$$

Storey *et al.* (2006) proposed some strategies for forming the $\hat{w}_i$-estimates. One of these is to set the $\hat{w}_i$ to 0 or 1 by employing a Kruskal–Wallis statistic. The tests are ranked from most to least significant according to this conventional statistic, where the bottom $\hat{m}_0$ tests have $\hat{w}_i = 1$ and $\hat{w}_i = 0$ otherwise. The quantity $\hat{m}_0$ is the estimated number of true null hypotheses from the methodology that was proposed in Storey (2002) and Storey and Tibshirani (2003).

There are of course many approaches that we could take to estimating the ODP significance thresholding function. For example, one modification of equation (9) that is immediately apparent is

$$\hat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}) = \sum_{i=1}^{m} (1 - \hat{w}_i) \, \hat{g}_i(\mathbf{x}) \Big/ \sum_{i=1}^{m} \hat{w}_i \, \hat{f}_i(\mathbf{x}),$$

where this would serve as a direct estimate of the significance thresholding function $\mathcal{S}_{\text{ODP}}$ that is defined in equation (2) of lemma 2.

When applying the ODP in practice, it is important to consider the effect of ancillary information on the procedure. Specifically, it is desirable to obtain a 'nuisance parameter invariance' property. Storey *et al.* (2006) laid out some general properties that yield the nuisance parameter invariance in practice as well as some specific steps to take to achieve this when the probability densities follow the normal distribution.

4.3. Assessing significance

The following proposal is the basic idea for assessing significance in this general setting. The estimated ODP significance thresholding function is evaluated on each test's data set, producing a set of observed statistics. Then the bootstrap data resampling technique is used to generate sets of null ODP statistics. For any given threshold λ applied to the observed statistics, EFP, ETP and FDR are all straightforwardly estimated by using the techniques that were proposed in Storey (2002) and Storey and Tibshirani (2003). The formulae are written slightly differently though, because the estimates are based on thresholding the statistics directly rather than forming p-values first. However, these formulae are implicitly carried out in the methodology that was proposed by Storey and Tibshirani (2003), and we showed the exact equivalence in Storey *et al.* (2006).

Let $\mathbf{x}_i^{0b}$ be the bootstrap null data set for test i and bootstrap iteration b, $b = 1, 2, \ldots, B$. For a threshold λ define

$$\widehat{\mathrm{EFP}}(\lambda) = \frac{\hat{m}_0 \sum\limits_{b=1}^{B} \sum\limits_{i=1}^{m} \mathbf{1}\{\hat{\mathcal{S}}_{\mathrm{ODP}}^*(\mathbf{x}_i^{0b}) \geqslant \lambda\}}{mB}$$

$$\widehat{\mathrm{ETP}}(\lambda) = \sum\limits_{i=1}^{m} \mathbf{1}\{\hat{\mathcal{S}}_{\mathrm{ODP}}^*(\mathbf{x}_i) \geqslant \lambda\} - \widehat{\mathrm{EFP}}(\lambda)$$

where $\hat{m}_0$ is estimated as in Storey and Tibshirani (2003). The FDR estimate can then be written in terms of these estimates, further showing the direct connection between EFP, ETP and FDR:

$$\widehat{\mathrm{FDR}}(\lambda) = \frac{\widehat{\mathrm{EFP}}(\lambda)}{\widehat{\mathrm{EFP}}(\lambda) + \widehat{\mathrm{ETP}}(\lambda)}.$$

These estimates have previously been proposed and theorems proved about them in Storey (2002), Storey *et al.* (2004) and Storey and Tibshirani (2003). The formulae look slightly different here, but they are equivalent to those based on certain p-values that have been calculated from an STP.

5. Extensions and connections to other concepts

The formulation and optimality of the ODP can be connected to several other well-known concepts in statistics. Here I discuss connections to FDR, the Bayesian approach to hypothesis testing, Stein's paradox and shrinkage estimation.

5.1. False discovery rate optimality by the optimal discovery procedure

The optimality that is achieved by the ODP is described in terms of maximizing ETP for each fixed EFP-level. However, it is straightforward to show that this is related to optimality in terms of FDR, which is currently a popular multiple-testing error measure. In what follows I show that minimizing the 'missed discovery rate' for each fixed FDR is approximately (and sometimes exactly) equivalent to the optimization that the ODP achieves.

Let FP = false positives, TP = true positives, FN = false negatives and TN = true negatives for some particular significance threshold. These are the four types of outcomes that occur when applying a significance threshold to multiple significance tests. FDR is the proportion of false positive results among all tests called significant (Soric, 1989; Benjamini and Hochberg, 1995):

$$\text{FDR} \equiv E\left[\frac{\text{FP}}{\text{FP} + \text{TP}}\right],$$

where the denominator of the ratio is set to 0 when no null hypotheses are rejected. This quantity can be written as a trade-off between EFP and ETP:

$$\text{FDR} \approx \frac{\text{EFP}}{\text{EFP} + \text{ETP}},$$

where the approximate equality is sometimes an exact equality. The approximation applies when testing a large number of hypotheses that are at most weakly dependent so that

$$\lim_{m \to \infty} \left|\text{FDR} - \frac{\text{EFP}}{\text{EFP} + \text{ETP}}\right| = 0.$$

The exact conditions for this convergence have been defined and studied elsewhere (Storey *et al.*, 2004). A variation of FDR, called the positive FDR, pFDR, is the proportion of false positive results only in cases where at least one test is called significant. Under a Bayesian mixture model assumption (which is similar to lemma 4 below), it has been shown that pFDR = EFP/(EFP + ETP), where this is an *exact* equality (Storey, 2003). Finally, we can consider another variation on FDR, which I call the marginal FDR, mFDR, that is simply defined to be the ratio of EFP to the total number of tests expected to be significant: mFDR ≡ EFP/(EFP + ETP). The important point is that in all these cases FDR can be written and understood in terms of EFP and ETP.

It has recently been suggested that FDR optimality should be defined in terms of the proportion of true alternatives among the tests that are not called significant (Genovese and Wasserman, 2002). This quantity has been called the 'false non-discovery rate' (Genovese and Wasserman, 2002) and the 'miss rate' (Taylor *et al.*, 2005); to find a common name, I call it the 'missed discovery rate' MDR. Specifically, a procedure is considered to be optimal if, for each fixed FDR-level, MDR is minimized. The above formulations of FDR can easily be extended to MDR. It can be shown that

$$\text{MDR} \equiv E\left[\frac{\text{FN}}{\text{FN} + \text{TN}}\right] \approx \frac{\text{EFN}}{\text{EFN} + \text{ETN}},$$

where EFN is the expected number of false negative and ETN is the expected number of true negative results. Again, the approximate equality is sometimes an exact equality. This shows that MDR can be understood and written in terms of EFN and ETN. However, EFN and ETN do not provide any additional information beyond EFP and ETP:

$$\frac{\text{EFN}}{\text{EFN} + \text{ETN}} = \frac{m - m_0 - \text{ETP}}{m - \text{ETP} - \text{EFP}}.$$

Therefore, the trade-off between FDR and MDR can be understood in terms of EFP and ETP, which is stated precisely in the following lemma.

Lemma 3 (FDR optimality by the ODP). Suppose that we represent FDR and MDR by EFP/(EFP + ETP) and EFN/(EFN + ETN) respectively. If ETP is maximized for each fixed EFP level, then MDR is minimized for each fixed FDR level (i.e. achieving optimality in terms of EFP and ETP is equivalent to achieving optimality in terms of FDR and MDR), implying that the ODP also achieves FDR optimality under this representation.

Therefore, the optimality that is achieved by the ODP is closely related, and sometimes exactly related, to FDR optimality. It can also be argued that EFP and ETP are more fundamental components than FDR, so perhaps the ODP optimality is more relevant in general.

5.2. Optimal discovery procedure under randomized null hypotheses

The ODP was formulated in Section 2 under the assumption that the status (i.e. truth or falsehood) of each null hypothesis is fixed. Because of this, the ODP is defined in terms of the *true* distribution for each significance test. As it turns out, the status of each test must be modelled as *random* in order for the null and alternative densities of every test to be present in the ODP. The following lemma derives the ODP when the status of each test is randomized.

Lemma 4 (randomized null ODP). Suppose that m significance tests are performed on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, where significance test i has null density f_i and alternative density g_i, for $i = 1, \ldots, m$. Suppose that each null hypothesis is true with probability π_0. The following significance thresholding function defines the ODP in this case:

$$\mathcal{S}_{\mathrm{ODP}}(\mathbf{x}) = \frac{g_1(\mathbf{x}) + g_2(\mathbf{x}) + \ldots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_m(\mathbf{x})}.$$

Null hypothesis i is rejected if and only if $\mathcal{S}(\mathbf{x}_i) \geqslant \lambda$ for some $0 \leqslant \lambda < \infty$. For each fixed λ, this procedure yields the maximum number of expected true positive results (ETP) among all simultaneous thresholding procedures that have an equal or greater number of expected false positive results (EFP).

Proof. The proof is similar to that for lemma 2. Let $\pi_1 = 1 - \pi_0$, where π_0 is the prior probability that any null hypothesis is true. EFP and ETP for any general significance region Γ are

$$\mathrm{EFP}(\Gamma) = \int_{\Gamma} \pi_0 \, f_1(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \ldots + \int_{\Gamma} \pi_0 \, f_m(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

$$\mathrm{ETP}(\Gamma) = \int_{\Gamma} \pi_1 \, g_1(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \ldots + \int_{\Gamma} \pi_1 \, g_m(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

The proof follows exactly as above, except that $\bar{f} = [\Sigma_{i=1}^m f_i]/m$ and $\bar{g} = [\Sigma_{i=1}^m g_i]/m$. □

This lemma and proof are easily extended to accommodate both

(a) priors that differ from test to test and
(b) versions of EFP and ETP where each test may be weighted differently.

No assumptions are made about independence between tests, either among the observed data for each test or the randomized status of the null hypothesis; the lemma holds under arbitrary levels of dependence. This result provides a bridge between the purely frequentist ODP and the Bayesian approach to classification of hypotheses, and it allows us to explore the ODP as a shrunken version of the likelihood ratio test, both discussed below.

5.3. Bayesian hypothesis testing

Besides randomizing the status of the null hypotheses, a prior distribution can be placed on the f_i and the g_i making the null and alternative densities random for each test. Classical Bayesian classification theory begins with these assumptions, which can then be used to find the rule that classifies each test as 'true null' or 'true alternative' so that the misclassification rate is minimized. It is easily shown that the goal of optimizing ETP for each fixed EFP level is equivalent to minimizing the misclassification rate, where the relative weights of the two types of misclassifications is determined by the level at which EFP is fixed. The optimal Bayesian classifier

(called the Bayes rule) has a simple form that can be connected to the NP and ODP approaches. Letting f_B be the expected null density and g_B the expected alternative density, the significance thresholding function can be written as

$$S(\mathbf{x}) = \frac{g_B(\mathbf{x})}{f_B(\mathbf{x})},$$

where the alternative hypothesis i is classified as true if and only if $S(\mathbf{x}_i) \geqslant \lambda$. The choice of λ is driven by the prior distributions and the loss that is placed on each type of error. This is not the usual way for writing the Bayes rule classifier, although it is algebraically equivalent.

From one perspective, this rule is more similar to the NP approach than to the ODP approach. The reason is that the high dimensional information across tests is averaged out in forming f_B and g_B, so we are essentially back into the NP setting. As it turns out, the ODP can be seen as a more specialized version of the Bayes rule, where one instead conditions on the actual realized densities for each test when forming an optimal testing procedure. Since the ODP minimizes the misclassification rate for each realized set of densities, it continues to do so when averaging over all densities. Therefore, the ODP and the Bayes rule both obtain the lowest misclassification rate when averaging over randomized hypotheses and probability density functions.

In the multiple normal means example of Section 3, we could put a prior distribution on the μ_i. Suppose that μ_i equals 0 (the null case) with probability π_0, and μ_i is drawn from, say, an $N(\theta, 1)$ distribution with probability $1 - \pi_0$. The Bayes rule that minimizes the misclassification rate can be written in terms of the significance thresholding function

$$S(z) = \frac{\{1/\surd(4\pi)\} \exp\{-(z-\theta)^2/4\}}{\{1/\surd(2\pi)\} \exp(-z^2/2)},$$

where we threshold it as above. With no prior knowledge about any asymmetry in the true alternative values, a reasonable value for θ would be $\theta = 0$. In this case, the Bayesian approach reduces to thresholding tests based on $|z_i|$, as is the case for the conventional procedures that we compared the ODP with in Section 3.3 and Fig. 2. Examples were shown there where the ODP outperforms these procedures.

Even though the Bayes rule is more transparently related to the NP approach, there is also a direct connection to the ODP. The following is an empirical Bayes interpretation of our approach. Suppose that there is a uniform prior on the unknown null densities $f_1, f_2, \ldots, f_m$, as well as on the alternative densities $g_1, g_2, \ldots, g_m$. Further, each null hypothesis is true with some prior probability π_0. It easily follows on the basis of lemma 4 that the ODP in this scenario is defined by the significance thresholding function

$$\frac{g_1(\mathbf{x}) + g_2(\mathbf{x}) + \ldots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_m(\mathbf{x})}. \tag{10}$$

This rule is equivalent to thresholding tests for significance according to the quantity

$$\frac{(1-\pi_0) \sum_{i=1}^{m} g_i(\mathbf{x})}{\pi_0 \sum_{i=1}^{m} f_i(\mathbf{x}) + (1-\pi_0) \sum_{i=1}^{m} g_i(\mathbf{x})},$$

which is exactly equal to the posterior probability that a test with data $\mathbf{x}$ is a true alternative. It follows by theorems 1 and 6 of Storey (2003) that this rule is also optimal in terms of both FDR and the misclassification rate in this Bayesian setting.

If we set all $\hat{w}_i = 1$ in the proposed ODP estimate in equation (9) of Section 4.2, we obtain the significance thresholding rule

$$\frac{\hat{g}_1(\mathbf{x}) + \hat{g}_2(\mathbf{x}) + \ldots + \hat{g}_m(\mathbf{x})}{\hat{f}_1(\mathbf{x}) + \hat{f}_2(\mathbf{x}) + \ldots + \hat{f}_m(\mathbf{x})}. \tag{11}$$

This statistic is simply equation (10) with each f_i and the g_i replaced with an estimate based on the data, making it interpretable as an empirical Bayes estimate. Moreover, the estimate of π_0 that is formed when estimating the q-values can be included to form conservative empirical Bayes estimates of the probability that a test with data $\mathbf{x}$ is a true alternative:

$$\widehat{\Pr}(\text{true alternative}|\mathbf{x}) = \frac{(1 - \hat{\pi}_0) \sum\limits_{i=1}^{m} \hat{g}_i(\mathbf{x})}{\hat{\pi}_0 \sum\limits_{i=1}^{m} \hat{f}_i(\mathbf{x}) + (1 - \hat{\pi}_0) \sum\limits_{i=1}^{m} \hat{g}_i(\mathbf{x})}.$$

Note that this probability estimate yields the same thresholding rule as the above statistic.

Even though empirical Bayes procedures for hypothesis testing have been proposed (Newton *et al.*, 2001; Efron *et al.*, 2001; Lönnstedt and Speed, 2002), the above ODP estimate is a novel empirical Bayes procedure. Employing a prior of m distinct probability density functions for the null and alternative cases has not been considered, and it is not obvious to do so until one formulates and derives the frequentist ODP.

We have found that our ODP estimate performs better when the $\hat{w}_i$ are estimated from the data and not all set to 1 (Storey *et al.*, 2006). The reason for this is that the estimates $\hat{f}_i$ can behave particularly badly for tests whose data show a strong alternative hypothesis signal. When the $\hat{w}_i$ are not trivially all set to 1, the procedure is based on the significance thresholding function that is given in equation (9) as opposed to equation (11) above. Here, the prior on the null hypothesis is taken from some $\Sigma_{i=1}^{m} \hat{w}_i$ distinct null probability density functions, which provides an even more non-obvious and novel empirical Bayes approach.

A question remains about whether it is better to apply the ODP or the Bayes rule when doing a large number of tests. Although this is quite difficult to answer in general, we have shown that an estimated ODP substantially outperforms the empirical Bayesian methods of Efron *et al.* (2001) and Lönnstedt and Speed (2002) when attempting to identify differentially expressed genes in DNA microarray experiments (Storey *et al.*, 2006).

5.4. Stein's paradox
Stein's paradox (Stein, 1956, 1981) greatly influenced notions about high dimensional point estimation when it was shown that estimating several normal means can be universally improved by shrinking the usual estimators (the sample means) towards a common value. The amount of shrinkage depends on the behaviour of the data as a whole. In other words, the shrunken estimates take into account all of the data at once. This estimator is usually referred to as the James–Stein estimator (James and Stein, 1961). This result has been a motivation for work done in empirical Bayes methodology (Efron and Morris, 1972, 1973) and wavelet estimation (Donoho and Johnstone, 1995). In general, it has played a substantial role in the way that statisticians think about point estimation in the high dimensional setting.

There does not appear to be any previously well-established analogue to Stein's paradox in the significance testing setting. However, the formulation of the ODP and the numerical comparisons that were shown above provide a first step towards this. In particular, in Section 3 I derived the ODP for the problem of testing multiple normal means for equality to 0. I also derived an estimate of the ODP that outperformed the UMP unbiased procedure for testing a normal mean for equality to 0. What is notable about the numerical results in Section 3.3 is that the UMP unbiased test is no longer optimal in the multiple-testing setting. In other words,

what performs optimally test by test may not continue to do so when considering multiple-test optimality.

The UMP unbiased test as presented in Section 3 is 'UMP' for a single significance test of $\mu = 0$ *versus* $\mu \neq 0$. This means that, for any alternative value of μ, there is no other unbiased testing procedure that exceeds this one in power—this is uniformly true among all type I error rates and alternative mean values. By taking the sum of the powers across all tests (i.e. ETP), I have shown several cases where the estimated ODP does exceed the UMP unbiased test in power, thereby showing that the UMP unbiased test is no longer optimal in the multiple-test setting. Furthermore, I have shown that the UMP unbiased test is equivalent to the theoretically optimal procedure only in cases where the ODP thresholding function is symmetric about zero.

5.5. Shrinkage estimation

The ODP also provides a multiple-testing analogue to the shrinkage estimator interpretation of the James–Stein estimate. As stated above, the James–Stein estimate shrinks each individual sample mean towards a central quantity. Under certain assumptions, the ODP can be written as shrinking the single-test likelihood ratio statistic towards a quantity involving information from the other tests. When the status of each test is made random (i.e. each null hypothesis is true with a certain probability), the ODP thresholding rule is defined by

$$\mathcal{S}_{\mathrm{ODP}}(\mathbf{x}) = \frac{g_1(\mathbf{x}) + g_2(\mathbf{x}) + \ldots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_m(\mathbf{x})}.$$

This result is explicitly stated above in lemma 4 of Section 5.2.

The overall ODP statistic can then be written as a weighted average of

(a) the individual test's NP likelihood ratio statistic and
(b) the ODP statistic applied to the remaining tests. For example, hypothesis test 1 has significance thresholding function

$$\mathcal{S}_{\mathrm{ODP}}(\mathbf{x}_1) = \gamma_1 \frac{g_1(\mathbf{x}_1)}{f_1(\mathbf{x}_1)} + (1 - \gamma_1) \frac{g_2(\mathbf{x}_1) + \ldots + g_m(\mathbf{x}_1)}{f_2(\mathbf{x}_1) + \ldots + f_m(\mathbf{x}_1)},$$

where the weight γ_1 is specific to test 1. The first term in the weighted sum is the NP statistic applied to test 1, and the second term is the ODP for tests 2–m applied to test 1. The formula for the weight is $\gamma_1 = f_1(\mathbf{x}_1) / \{f_1(\mathbf{x}_1) + f_2(\mathbf{x}_1) + \ldots + f_m(\mathbf{x}_1)\}$. The interpretation is that γ_1 quantifies how useful the information from the other tests is: the more similar the null distributions, then the more the ODP uses information from the other tests. This makes sense because it looks at the relative contribution to the EFP from the other tests relative to test 1. Note that, when an estimated form of the ODP can be written as above (e.g. the estimate of equation (9) with all $\hat{w}_i = 1$), then the interpretation is similar.

6. Discussion

A new statistical theory has been developed here that shows how optimally to perform multiple significance tests based on a simultaneous thresholding procedure. The ODP allows us to test multiple hypotheses simultaneously in such a way that the total number of expected true positive results is maximized for each fixed number of expected false positive results. This procedure can be viewed as a multiple-test extension of the NP procedure for testing a single hypothesis. The ODP has connections to several different areas of statistics, including FDRs, Bayesian classification, shrinkage estimation and Stein's paradox.

The recent explosion in research on multiple testing has consistently started with the assumption that *p*-values are obtained for each test individually. In contrast, the ODP is a high dimensional approach that uses all of the relevant information across tests when assessing the significance of each one. The ODP method does not merely modify existing single-test procedures, thus apparently differing from the current point of view.

I have briefly discussed a strategy for implementing the ODP in practice. However, this is not an easy task, and it will take some substantial developments to arrive at a generally applicable set of methods. We have applied and extended the ideas here to applications in genomics to produce an estimated version of the ODP for identifying genes that are differentially expressed in comparative microarray experiments (Storey *et al.*, 2005, 2006). This method shows surprisingly strong gains in power relative to several leading methods that are currently available. It is also pointed out there that the ODP strategy may be useful in a variety of high dimensional biological studies because there is often a strong and largely unknown structure among the significance tests, where the goal is to extract as much biologically meaningful signal as possible.

Acknowledgements

References

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

Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

Efron, B. and Morris, C. N. (1972) Empirical Bayes on vector observations: an empirical Bayes approach. *Biometrika*, **59**, 335–347.

Efron, B. and Morris, C. N. (1973) Stein's estimation rule and its competitors: an empirical Bayes approach. *J. Am. Statist. Ass.*, **68**, 117–130.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B, **64**, 499–517.

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.

Lehmann, E. L. (1986) *Testing Statistical Hypotheses*, 2nd edn. Berlin: Springer.

Lehmann, E. L., Romano, J. P. and Shaffer, J. P. (2005) On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.*, **33**, 1084–1108.

Lönnstedt, I. and Speed, T. (2002) Replicated microarray data. *Statist. Sin.*, **12**, 31–46.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. and Tsui, K. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Computnl Biol.*, **8**, 37–52.

Neyman, J. and Pearson, E. S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc.*, **231**, 289–337.

Shaffer, J. (1995) Multiple hypothesis testing. *A. Rev. Psychol.*, **46**, 561–584.

Soric, B. (1989) Statistical discoveries and effect-size estimation. *J. Am. Statist. Ass.*, **84**, 608–610.

Spjotvoll, E. (1972) Optimality of some multiple testing procedures. *Ann. Math. Statist.*, **43**, 398–411.

Stein, C. (1956) Inadmissability of the usual estimator for the mean of a multivariate distribution. In *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 197–206. Berkeley: University of California Press.

Stein, C. (1981) Estimation of the mean of a multivariate Normal distribution. *Ann. Statist.*, **9**, 1135–1151.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.

Storey, J. D. (2003) The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Statist.*, **31**, 2013–2035.

Storey, J. D. (2005) The optimal discovery procedure: a new approach to simultaneous significance testing. *Working Paper 259*. Department of Biostatistics, University of Washington, Seattle. (Available from `http://www.bepress.com/uwbiostat/paper259/`.)

Storey, J. D., Dai, J. Y. and Leek, J. T. (2005) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Working Paper 260.* Department of Biostatistics, University of Washington, Seattle. (Available from `http://www.bepress.com/uwbiostat/paper260/`.)

Storey, J. D., Dai, J. Y. and Leek, J. T. (2006) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, doi: 10.1093/biostatistics/kx1019.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.

Taylor, J., Tibshirani, R. and Efron, B. (2005) The miss rate for the analysis of gene expression data. *Biostatistics*, **6**, 111–117.