Biometrics 62, 1099–1106
December 2006
DOI: 10.1111/j.1541-0420.2006.00616.x

Shrunken *p*-Values for Assessing Differential Expression with Applications to Genomic Data Analysis

Debashis Ghosh

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.
email: ghoshd@umich.edu

Summary. In many scientific problems involving high-throughput technology, inference must be made involving several hundreds or thousands of hypotheses. Recent attention has focused on how to address the multiple testing issue; much focus has been devoted toward the use of the false discovery rate. In this article, we consider an alternative estimation procedure titled shrunken *p*-values for assessing differential expression (SPADE). The estimators are motivated by risk considerations from decision theory and lead to a completely new method for adjustment in the multiple testing problem. In addition, the decision-theoretic framework can be used to derive a decision rule for controlling the number of false positive results. Some theoretical results are outlined. The proposed methodology is illustrated using simulation studies and with application to data from a prostate cancer gene expression profiling study.

Key words: Hypothesis testing; James–Stein estimator; Microarray; Multiple comparisons; Shrinkage estimators; Simultaneous inference.

1. Introduction

Because of technological developments in scientific fields such as genomics, it has become possible to simultaneously assay the biological activities of thousands of genes in parallel. Similarly, in neuroimaging, there is consideration of thousands of voxels as a global map of the human brain. A common problem in this setting is to determine which objects are differentially expressed between two conditions (genes in the microarray setting, voxels in the neuroimaging example). Consideration of all the hypotheses leads to a multiple comparisons problem.

Our work is motivated by a collaborative gene expression profiling study in prostate cancer (Dhanasekaran et al., 2001). The investigators have profiled tissue samples from various stages of prostate cancer (normal adjacent prostate, benign prostatic hyperplasia, localized prostate cancer, and advanced metastatic prostate cancer) using microarrays. In addition to the gene expression profiles for a sample, the investigators have access to several other clinical parameters. A key hypothesis made by investigators is that there exists a set of genes that distinguish aggressive prostate cancer from nonlethal prostate cancer. To begin with, a fairly standard analysis would be to determine which genes are differentially expressed between aggressive prostate cancer and nonaggressive prostate cancer. Here, we will focus on finding genes that are differentially expressed between metastatic prostate cancer (i.e., cancer that has spread to other organ sites) versus localized prostate cancer.

Methods for dealing with differential expression in the multiple testing setting have been the subject of much research interest in the recent statistical literature. Methods for controlling the familywise error rate (FWER) and related quantities have been proposed by Ge, Dudoit, and Speed (2003) and by van der Laan, Dudoit, and Pollard (2004). Several authors have argued that control of the FWER is too stringent and have advocated the use of the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995). Methods for both controlling the FDR as well as estimating it directly have appeared in the recent statistical literature (e.g., Benjamini and Yekutieli, 2001; Efron et al., 2001; Sarkar, 2002; Storey, 2002).

As noted by Storey (2002), there is an explicit mixture model for the distribution of marginal test statistics from which the positive FDR (Storey, 2003) and the FDR can be estimated. In this article, we consider an alternative statistical method that can be motivated from the same mixture model for dealing with multiple testing. The procedures proposed in this article, termed shrunken *p*-values for assessing differential expression (SPADE), have links to decision-theoretic considerations (Berger, 1985). In particular, we study shrinkage estimators for the *p*-value under both L_1 and L_2 loss functions. In addition, the decision-theoretic framework allows us to construct an optimal decision rule for the selection of differentially expressed genes that controls the number of false positives under L_1 loss, which we provide in Section 4.3.

The structure of the article is as follows. A definition of FDR in the multiple testing situation, along with previous work, is reviewed in Section 2. In Section 3, we describe some

Table 1
Outcomes of n tests of hypotheses

	Accept	Reject	Total
True null	U	V	n_0
True alternative	T	S	n_1
	W	Q	m

Q is the number of rejected hypotheses, and W = m − Q. U and V are the number of true null hypotheses that are not rejected and rejected, respectively. T and S are the number of true alternative hypotheses that are not rejected, respectively.

decision-theoretic results in the case of single and multiple hypotheses. In Section 4, we outline the SPADE method. A key notion here is shrinkage toward the two components of the mixture model. While such an idea has been pursued by George (1986) for data from a normal mixture model, he did not consider the case of a mixture model for p-values, nor did he give approaches for estimation with observed data. Practical implementation of the SPADE methodology is discussed and numerical comparisons with simulated and real data are made in Section 5. We conclude with a brief discussion in Section 6.

2. False Discovery Rate: Background

Our setup is that we have test statistics $T_1, \ldots, T_n$ for testing hypotheses H_{0i}, $i = 1, \ldots, n$. Of the n hypotheses, suppose that for n_0 of them, the null is true. Using the following 2×2 contingency table, we can categorize hypotheses by whether they are true or not and whether or not we reject or fail to reject them. This is shown in Table 1.

Benjamini and Hochberg (1995) propose a method for controlling the so-called FDR, defined as

$$\text{FDR} \equiv E\left[\frac{V}{Q}\,\middle|\, Q > 0\right] P(Q > 0).$$

The conditioning on the event $[Q > 0]$ is needed because the fraction V/Q is not well defined when $Q = 0$. Several authors have developed single-step methods for controlling the FDR (Benjamini and Hochberg, 1995; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001; Sarkar, 2002).

An alternative approach that has been taken in the recent statistical literature is to fix a rejection region and to estimate FDR. Storey (2002, 2003) considers a mixture model for motivating the FDR. Define indicator variables $H_1, \ldots, H_n$, corresponding to $T_1, \ldots, T_n$, where $H_i = 0$ if the ith null hypothesis is true and $H_i = 1$ if the ith alternative hypothesis is true ($i = 1, \ldots, n$). $H_1, \ldots, H_n$ are a random sample from a Bernoulli distribution where $P(H_i = 0) = \pi_0$, $i = 1, \ldots, n$. If f_0 and f_1 correspond to the densities to $T_i \,|\, H_i = 0$ and $T_i \,|\, H_i = 1$ ($i = 1, \ldots, n$), respectively, the density corresponding to the marginal distribution of test statistics $T_1, \ldots, T_n$ is

$$f(t) \equiv \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \tag{1}$$

Methods for FDR estimation based on (1) have been developed by several authors (Efron et al., 2001; Storey, 2002; Pounds and Cheng, 2004; Dalmasso, Broët, and Moreau, 2005). While we assume here that the test statistics are independent, several authors (Genovese and Wasserman, 2004; Storey, Taylor, and Seigmund, 2004) have shown that estimates of FDR are fairly robust to various forms of dependence.

A related quantity to FDR is the positive false discovery rate (pFDR) (Storey, 2003), defined as $\text{pFDR} = E[\frac{V}{Q} \,|\, Q > 0]$. For the multiple testing context, an analogous quantity to p-values based on pFDR, proposed by Storey (2002), is the q-value. When inference is performed using p-values, rejection regions for the null hypothesis are intervals of the form $[0, c]$, where $0 < c < 1$. The q-value corresponding to a given p-value p is defined as

$$q(p) = \inf_{c \geq p} \text{pFDR}(c) = \inf_{c \geq p} \left\{ \frac{\pi_0 c}{F_P(c)} \right\}, \tag{2}$$

where F_P is the distribution function for the p-value. This corresponds to equation (21) in Storey (2002). q-values are tailored to the multiple comparisons problem, and their use is much like that of the p-value. Smaller q-values correspond to greater evidence against a null hypothesis.

3. Hypothesis Testing in the Decision-Theoretic Framework

We first start by considering the case of $n = 1$ hypothesis. In a decision theory framework, we use the following loss functions: for $k = 1, 2$,

$$L_k(\mu, d) = |\mu - d(T)|^k, \tag{3}$$

where μ is the population quantity to be estimated and d is an estimator or more generally, an element of the action space. Note that if μ and d can only take values 0 and 1, then for $k = 1$, (3) reduces to misclassification error.

For hypothesis testing, it is not clear what the parameter being estimated is in a loss-based framework. We follow the approach of Hwang et al. (1990) and take μ to be an indicator that the null hypothesis is true, that is, $\mu = I(H = 0)$ using the notation of Section 2. Thus, μ takes only two values, 0 and 1. Note that the Bayes rules for L_1 and L_2 loss functions are given by

$$d_1^{\text{opt}}(T) = I\{P(H = 0 \,|\, T) > 1/2\}$$

and

$$d_2^{\text{opt}}(T) = P(H = 0 \,|\, T).$$

Note that both procedures are based on the posterior probability $P(H = 0 \,|\, T)$ but that for L_1 loss, the probability is thresholded at $1/2$ while it is unthresholded for L_2 loss. Thus, the Bayes rule for L_1 loss takes two values, 0 and 1, while that for L_2 loss, its range is [0, 1].

We now go from one hypothesis to multiple hypotheses. In terms of a loss function, we now consider

$$L_k(\mathbf{H}, \mathbf{d}) = \sum_{i=1}^{n} |I(H_i = 0) - d_i(T_i)|^k, \tag{4}$$

where $\mathbf{H} = (H_1, \ldots, H_n)$, $k = 1, 2$, and d_i corresponds to d in (3). The individual component Bayes rules from (4) for $k = 1$ and $k = 2$ are given by $d_1^{\text{opt}}(T_i)$ and $d_2^{\text{opt}}(T_i)$, $i = 1, \ldots, n$. The idea behind SPADE is that by pooling information across genes, we can construct shrinkage estimators of $P(H_i = 0 \,|\, T_i)$ or equivalently, $P(H_i = 0 \,|\, T_i)$, that will

lead to reductions in risk behavior. This is known as the Stein phenomenon (Berger, 1985, p. 360). The target estimand is $P(H_i = 0 \,|\, T_i)$; this is the quantity that we will be constructing shrinkage estimators for in Section 4. We note in passing that shrinkage estimators of $I\{P(H_i = 0 \,|\, T_i) > 1/2\}$ will not lead to risk reductions because of the fact that only two values can be obtained.

4. SPADE: Proposed Methodology

4.1 *Shrinkage Estimation*

The starting point for our methodology is (1). Observe that (1) specifies a model for the test statistics, which we are assuming to be independent. In the original article by Storey (2002), the test statistics used for testing the hypotheses $H_1, \ldots, H_n$ were the p-values. The FDR was estimated on the basis of the p-values and estimating π_0, the proportion of true null hypotheses, using a permutation scheme.

If we let $p_1, \ldots, p_n$ denote the p-values for testing $H_{01}, \ldots, H_{0n}$, then the model induced by (1) is

$$p_1, \ldots, p_n \overset{\text{iid}}{\sim} \pi_0 F_U + (1 - \pi_0) F_V, \tag{5}$$

where F_U is the cumulative distribution function (c.d.f.) of $U \equiv \text{Uniform}(0,1)$ random variable and F_V is the cumulative distribution function of the p-values under the alternative hypothesis.

Our tack is to assume that each component of (5) specifies a value for the probability of the null being true. Following the arguments of George (1986), a James–Stein approach to constructing shrinkage estimators for $P(H_i = 0 \,|\, p_i)$ is to calculate for $i = 1, \ldots, n$,

$$p_i^{\text{JS}} = \pi_0(p_i) p_{0i}^{\text{JS}} + \{1 - \pi_0(p_i)\} p_{1i}^{\text{JS}}, \tag{6}$$

where

$$p_{0i}^{\text{JS}} = p_i - \left[1 \wedge \frac{(n-1)}{12 \sum_{i=1}^{n} (p_i - 1/2)^2} \right] (p_i - 1/2),$$

$$p_{1i}^{\text{JS}} = p_i - \left[1 \wedge \frac{(n-1)\sigma_1^2}{\sum_{i=1}^{n} (p_i - \mu_1)^2} \right] (p_i - \mu_1), \tag{7}$$

$$\pi_0(p) = \frac{\pi_0 p}{\pi_0 p + (1 - \pi_0) F_V(p)} = \frac{\pi_0 p}{F_P(p)}, \tag{8}$$

where μ_1 and σ_1^2 are the mean and variance corresponding to F_V. Note that the 1/2 and 1/12 refer to the mean and variance of a Uniform(0,1) distribution. These adjusted p-values are shrunken p-values that account for the multiple testing problem. This describes the essence of the SPADE methodology. Note that the mixture distribution of the p-values is providing two targets for shrinkage.

In fact, there are many choices for the definition of (8). We have defined it in terms of the c.d.f.'s for the two components of the mixture model. Suppose we consider an alternative definition for (8):

$$\tilde{\pi}_0(p) = \frac{\pi_0}{\pi_0 + (1 - \pi_0) f_V(p)}, \tag{9}$$

where f_V is the density function for V. Then (9) is precisely the local FDR (Efron et al., 2001) based on the p-value. We prefer the use of (8)–(9) because of variance issues. In particular (9) will have greater variance than (8) because density estimates tend to be much more variable than those based on the c.d.f.

The SPADE methodology adjusts the univariate p-values for the multiple testing problem by shrinkage in which the shrinkage weights (8) or (9) are data adaptive. Here and in the sequel, we consider (8). Suppose that a large fraction of null hypotheses are true, that is, $\pi_0 \approx 1$. Then based on (8) and (6), the p-value will be shrunk toward 1/2. By contrast, if a majority of the null hypotheses are false, then the shrunken p-value that adjusts for multiple testing will be weighted more toward the mean of the p-values under the alternative hypothesis. This shows how the SPADE methodology is data adaptive. In addition, one can view the adjusted p-values as empirical Bayes estimators of $P(H_i = 0 \,|\, p_i)$, $i = 1, \ldots, n$, similar to what has been done with the location parameter in normal probability models (Berger, 1985, Section 4.5). Note that the shrinkage factor presented here is an affine transformation of the original p-values so that the ranking of the original p-values is maintained.

It is interesting to note the relationship of the q-value method of Storey (2002) within this shrinkage framework. If we shrink the p-value to 1 under the null hypothesis and 0 under the alternative, then using (6), we get that p_i^{JS} $(i = 1, \ldots, n)$ equals $\pi_0(p_i)$, which is bounded from below by $q(p_i)$ from (2). Thus, there is a shrinkage aspect to the q-value methods, albeit much more extreme than that being proposed here. The shrinkage occurs because we are pooling information across genes.

Another interpretation of (6) is as a doubly shrunken p-value, shrunk toward each component of the mixture. This idea was originally proposed by George (1986) in the context of a normal probability model. There are several differences between his work and ours. First, we are considering a mixture model for the p-values, which is fundamentally different from the normal model considered by George (1986). In addition, note that there are unknown population quantities in (7) and (8) that need to be estimated. George (1986) provides no estimation procedure from observed data. Estimation methodologies will be dealt with in Section 4.2.

4.2 *Practical Implementation*

The major issues in implementing SPADE are twofold. First, π_0 needs to be estimated. This is the proportion of hypotheses estimated to be truly null. Second, the c.d.f. for the p-values under the alternative hypothesis also needs to be calculated. This will then provide estimates of μ_1 and σ_1^2 in (7). Observe that (5) implies the following result for the cumulative distribution of the p-values:

$$F_P(p) = \pi_0 p + (1 - \pi_0) F_V(p). \tag{10}$$

Simple algebraic manipulation of (10) yields

$$F_V(p) = \frac{F_P(p) - \pi_0 p}{1 - \pi_0}. \tag{11}$$

We can estimate F_P in (11) using the empirical distribution function of the observed p-values. Provided we have an

estimator of π_0, we can then estimate F_V and subsequently the mean and variance in (7). We use the empirical c.d.f. of F_P in (8). Thus, the outstanding issue becomes one of estimating π_0. We consider three approaches to do this.

The first is the algorithm by Storey and Tibshirani (2003), which has proven quite popular in the analysis of microarray data. It is summarized as follows. First, we order the n p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$. Then, we construct a grid of $L\lambda$ values, $\lambda_1, \ldots, \lambda_L$ and calculate

$$\hat{\pi}_0(\lambda_l) = \frac{\#\{p_j > \lambda\}}{n(1-\lambda)},$$

$l = 1, \ldots, L$. A cubic smoothing spline is then fit to the values $\{\lambda_l, \hat{\pi}_0(\lambda_l)\}, l = 1, \ldots, L$. Consequently, π_0 is estimated by the interpolated value at $\lambda = 1$.

The second is the spacings loess histogram (SPLOSH) algorithm of Pounds and Cheng (2004). Their algorithm proceeds by ordering the p-values and computing a local regression, where the response variable is a transformed slope of the empirical distribution function of the p-values and the independent variable is a midpoint of the distribution function. Based on the nonparametric regression fit, we get an estimator of π_0. Pounds and Cheng (2004) argue that their method is better than the q-value method of Storey and Tibshirani (2003) because while the estimator of π_0 from the latter method uses information to the left of λ, SPLOSH uses information from both directions. This implies that the SPLOSH estimator should be more stable than the q-value estimator.

The last method for π_0 is based on the location-based estimator (LBE) algorithm of Dalmasso et al. (2005). The motivation of this method is based on the asymptotic normality of estimators of π_0 using the central limit theorem. Dalmasso et al. (2005) consider an approach in which the p-values are transformed, and an estimator of π_0 is calculated as a back-transformation from the empirical distribution of the transformed p-values. There are many potential transformations that satisfy the necessary technical conditions in Dalmasso et al. (2005); they consider the following estimator of π_0:

$$\tilde{\pi}_0 = \frac{n^{-1}\sum_{i=1}^{n}[-\log(1-p_i)]^m}{m!}, \tag{12}$$

where m is an integer that needs to be estimated. The role of m is similar to that of a bandwidth in nonparametric regression. Larger values of m correspond to decreasing bias in the estimate of π_0, while smaller values of m lead to decreased variance in the estimate of π_0. Thus, we see a bias-variance tradeoff based on the choice of m. Dalmasso et al. (2005) suggest the following rule for the choice of m:

$$m = \max\left[1, \max\left\{m : \frac{\binom{2m}{m} - 1}{n} \leq l\right\}\right],$$

where l is a postulated value for the variance of π_0. The performance of SPADE with π_0 estimated from the three algorithms is assessed in a simulation study in Section 5.1.

4.3 *Misclassification Error and Terminal Stopping Rules*

So far we have focused on estimation of strength of evidence measures in the multiple testing context. The class of estimators has been motivated by decision-theoretic ideas. While not the focus of the article, the decision theory framework can be used to find classes of procedures that appropriately control the number of false positives in the multiple comparisons setting.

We consider the case where the loss function corresponds to misclassification error. Thus, for n hypotheses, either the null or alternative is true, and we decide in favor of one of them. Thus, for a single hypothesis, the decision is $d = 1$ (reject the null) or $d = 0$ (fail to reject the null). For an individual hypothesis, the loss is given by

$$L_i(H_i, d_i) = \begin{cases} 0, & \text{reject } H_{0i} \text{ when } H_i = 1 \text{ or fail to reject} \\ & \quad H_{0i} \text{ when } H_i = 0; \\ 1, & \text{reject } H_{0i} \text{ when } H_i = 1; \\ c_0, & \text{fail to reject } H_{0i} \text{ when } H_i = 1. \end{cases} \tag{13}$$

Note that in (13), c_0 can be viewed as the relative cost of a type II error relative to a type I error for a single hypothesis. Over the n hypotheses, the loss function is given by $L(\theta, d) = \sum_{i=1}^{n} L_i(H_i, d_i)$.

Based on this loss function, the number of false discoveries is given by $\text{FD} \equiv \sum_{i=1}^{n} L_i(0,1)$, while the number of false nondiscoveries equals $\text{FN} \equiv \sum_{i=1}^{n} L_i(1,0)$. The FDR and false nondiscovery rates (FNR) are given by

$$\text{FDR} = \frac{\text{FD}}{\sum_{i=1}^{n} d_i + a}$$

and

$$\text{FNR} = \frac{\text{FN}}{n - \sum_{i=1}^{n} d_i + a},$$

where a is a constant. Note that we have included a in the definition of FDR and FNR to allow for the situation where 0 or n hypotheses are rejected.

In multiple testing situations, we desire to control the expected proportion of false discoveries. This can be through control of either FDR or FD. Suppose we consider the following two-dimensional criterion functions:

$$L^1(\mathbf{H}, \mathbf{d}) = \{\text{E(FD)}, \text{E(FN)}\}$$

and

$$L^2(\mathbf{H}, \mathbf{d}) = \{\text{E(FDR)}, \text{E(FNR)}\}.$$

When dealing with multidimensional optimization criterion functions to minimize, a standard approach is to minimize one component subject to constraints on the others (Keeney, Raiffa, and Meyer, 1976). Thus, we may either find a decision rule that minimizes E(FN) subject to $\text{E(FD)} \leq \alpha_1$ or that minimizes E(FNR) subject to $\text{E(FDR)} \leq \alpha_2$. Interestingly, under the misclassification loss function, the optimal rules have very similar forms. The following result can be derived using arguments similar to those in the proof of Theorem 1 of Müller et al. (2004):

LEMMA 1: *Under loss functions L^1 and L^2, the optimal decision takes the following forms, respectively*:

$$d_i^{\text{opt},1} = I\{P(H = 1 \mid T_i) \geq t_1^*\},$$

where $t_1^* = \min\{s : \text{E(FD)} \leq \alpha_1\}$ *and*

$$d_i^{\text{opt},2} = I\{P(H = 1 \mid T_i) \geq t_2^*\},$$

where $t_2^* = \min\{s : \text{E(FDR)} \leq \alpha_2\}$, $i = 1, \ldots, n$.

Based on Lemma 1, we have a simple algorithm for selecting hypotheses in a way that controls false positives. We will not discuss implementation of the procedure here and will leave it as a topic for future research.

5. Numerical Examples

5.1 *Simulation Studies*

To evaluate the proposed procedures, we performed several simulation studies. In these numerical experiments, p-values were generated from model (5) with F_U being the c.d.f. for a uniform[0,1] distribution and F_V being the c.d.f. for a beta distribution under three scenarios, which we refer to as small, medium, and large. The adjectives refer to the discrepancy of the p-value from the null hypothesis:

(1) Small: beta distribution with parameters $\alpha = 3$ and $\beta = 4$. This choice of parameters gives a mean of 3/7 and a variance of 3/98 for the distribution of p-values under the alternative hypothesis.
(2) Medium: beta distribution with parameters $\alpha = 3$ and $\beta = 12$. This choice of parameters gives a mean of 3/15 and a variance of 1/100 for the distribution of p-values under the alternative hypothesis.
(3) Large: beta distribution with parameters $\alpha = 3$ and $\beta = 50$. This choice of parameters gives a mean of 3/53 and a variance of approximately 0.001 for the distribution of p-values under the alternative hypothesis.

For each simulation setting, we generated 1000 data sets. We considered sample sizes $n = 10{,}000$ and π_0 values of 0.2, 0.5, and 0.8. Our results did not change substantially with sample sizes $n = 2000$ and $n = 5000$; we do not report those results here. The q-value estimation procedures proposed by Storey (2002), Pounds and Cheng (2004), and Dalmasso et al. (2005) were used; we refer to them as Q-value1, Q-value2, and Q-value3. The shrunken p-value procedures based on the three algorithms described in Section 4.2 were also studied. The algorithm of Pounds and Cheng (2004) uses a local regression; the span used is the default value of 0.75. For the Dalmasso et al. (2005) method, we chose the default value of $m = 1$. The results using mean square error and $L_1 \equiv n^{-1}\sum_{i=1}^{n} |I(H_i = 0) - d_i|$ are presented in Table 2. We also studied performance using a misclassification error criterion; for this situation, we thresholded the q-value or the shrunken p-value at 1/2; the results are provided in Table 3. Because the methods are virtually indistinguishable in Table 2, we focus our attention on Table 2.

The simulation sheds light as to the decision-theoretic performance of the q-value methods as well as the proposed methods. In Table 2, we see that the L_1 error for the q-value methods is lower than for the other three methods; this is based on the fact that the shrinkage of the q-values is toward 0 and 1 in the double shrinkage framework outlined earlier. In Table 2, this is not always the case. In particular, when π_0 is small, then the proposed methods are quite competitive with the q-value. In fact, for smaller values of π_0, the q-value estimator of π_0 becomes quite unstable. The instability in the q-value-based estimator of π_0 has also been noticed by Pounds and Cheng (2004). If the groups being compared in the differential expression analysis represent grossly different phenotypes, then we would expect π_0 to be small. For more subtle phenotypes, the value of π_0 is larger; this is precisely where the q-value method will perform at its best. In addition, we find

Table 2
Estimated risk from simulation studies

Loss	Effect	π_0	Q-value1	Q-value2	Q-value3	SPADE1	SPADE2	SPADE3
L_2	Small	0.2	0.179	0.171	0.200	0.181	0.174	0.1582
		0.5	0.264	0.290	0.340	0.333	0.358	0.302
		0.8	0.165	0.190	0.240	0.333	0.380	0.31
	Medium	0.2	0.161	0.170	0.200	0.176	0.182	0.176
		0.5	0.251	0.212	0.500	0.348	0.326	0.348
		0.8	0.162	0.143	0.230	0.385	0.384	0.386
	Large	0.2	0.161	0.170	0.200	0.177	0.182	0.176
		0.5	0.252	0.212	0.500	0.348	0.330	0.348
		0.8	0.161	0.143	0.210	0.385	0.384	0.386
L_1	Small	0.2	0.235	0.253	0.200	0.223	0.235	0.282
		0.5	0.391	0.382	0.491	0.433	0.424	0.433
		0.8	0.365	0.424	0.535	0.564	0.596	0.549
	Medium	0.2	0.240	0.198	0.200	0.195	0.196	0.195
		0.5	0.391	0.373	0.495	0.429	0.428	0.430
		0.8	0.370	0.354	0.602	0.572	0.571	0.573
	Large	0.2	0.198	0.196	0.200	0.195	0.196	0.195
		0.5	0.391	0.375	0.500	0.433	0.424	0.433
		0.8	0.370	0.353	0.620	0.573	0.571	0.573

Q-value1, Q-value2, and Q-value3 refer to the methods of Storey and Tibshirani (2003), Pounds and Cheng (2004), and Dalmasso et al. (2005). SPADE1 is the SPADE methodology, where π_0 is estimated using the algorithm of Storey and Tibshirani (2003); SPADE2 is based on the Pounds and Cheng (2004) method for estimation of π_0; SPADE3 is based on the Dalmasso et al. (2005) method for estimation of π_0.

Table 3
Estimated misclassification errors from simulation studies

Effect	π_0	Q-value1	Q-value2	Q-value3	SPADE1	SPADE2	SPADE3
Small	0.2	0.200	0.201	0.201	0.200	0.200	0.196
	0.5	0.500	0.498	0.499	0.500	0.499	0.500
	0.8	0.201	0.200	0.200	0.199	0.201	0.202
Medium	0.2	0.202	0.197	0.198	0.201	0.198	0.201
	0.5	0.497	0.502	0.499	0.498	0.499	0.497
	0.8	0.201	0.200	0.196	0.201	0.202	0.200
Large	0.2	0.198	0.202	0.200	0.201	0.199	0.198
	0.5	0.499	0.497	0.499	0.501	0.500	0.499
	0.8	0.202	0.201	0.200	0.199	0.198	0.201

See the footnote in Table 2.

that the difference between the SPADE procedures with the q-value method diminishes for larger numbers of tests. This also suggests that shrinkage will be powerful in high-dimensional situations. We also note that collectively, the SPADE estimators tend to have more stable behavior than the q-value methods. This suggests that the SPADE methods are not as sensitive to the π_0 estimator as are the q-value methods.

One point of note is that (6) is constructed using squared error loss, which might not be the most appropriate scale for p-values. We also studied a modification of the procedure in Sections 3 and 4.2 in which transformed p-values based on twice the negative log p-values were calculated, and double shrinkage estimators were calculated on the transformed scale and backtransformed. In simulation experiments not reported here, this approach tended to have much worse performance than the procedure described here.

5.2 *Microarray Example*

We now return to the microarray data described in the Introduction. Measurements were made on $n = 9984$ genes for 79 individuals. There are 59 localized prostate cancer samples and 20 metastatic prostate cancer samples. Before analyzing the data, we took the following preprocessing steps:

(1) Genes that were reported as missing in more than 10% of samples were filtered out; and
(2) Genes that had a sample variation greater than 0.15 across all samples were retained.

This left a total of $n = 5241$ genes available for analysis.

We first calculated t-tests comparing gene expression in localized versus metastatic prostate cancer samples; we assumed unequal variances between the two groups. For the purposes of illustration, we used a normal approximation to calculate the p-values. The estimated fraction of null hypotheses using the Q-value, SPLOSH, and LBE methods are 0.308, 0.460, and 0.331, respectively. The corresponding values of (μ_1, σ_1) from the method of moments procedure are (0.04, 0.15), (0.018, 0.21), and (0.024, 0.15), respectively. The SPLOSH method gives a larger value for the proportion of nondifferentially expressed genes. However, the mean and variance for the distribution of p-values under the alternative hypothesis appear to be fairly concordant between the three methods.

We compared the adjustment in p-values using SPADE to the q-value estimates provided by other procedures (Q-value, SPLOSH, LBE). These are given in Figures 1–3; each q-value and corresponding SPADE method are plotted. Based on the plots, we find that there is high shrinkage of p-values using SPADE relative to all three methods. With the method of Storey (2002) and SPLOSH, the relationship between the shrunken p-values and q-values is monotone but nonlinear. With the LBE-estimated q-values, the shrunken p-values tend to show a more linear relationship.

6. Discussion

In this article, we have argued for a reinterpretation of the mixture model for multiple testing that allows for the consideration of shrinkage procedures. The work of Efron et al. (2001) and Storey (2002) for the estimation of FDRs represents one method of pooling information across test statistics. We have constructed an alternative procedure based on a James–Stein construction for the p-values. The resulting adjusted p-values from the SPADE procedure represent another method for addressing the multiple testing issue. Our framework is quite general and, in fact, the q-value methods proposed in the literature fit into it quite nicely.

While we have proposed new methods for multiple testing, the simulation studies also showed the situations in which the q-value (Storey and Tibshirani, 2003) performs relatively well. Namely, if the proportion of true null hypotheses is large, then the q-value will perform well. If the proportion is small, then the estimate of π_0 will be unstable, which will lead to poor performance of the q-value.

The multiple testing procedure proposed in the article is based on shrinkage estimation. This is also a common element in Bayesian testing procedures. It has been noted that Bayesian adjustment to the multiple testing problem leads to well-calibrated and more conservative inference procedures than non-Bayesian methods (Gelman and Tuerlinckx, 2000). Based on the results in the real data example, that appears to be the case here as well.

The decision-theoretic framework in which the FDR procedures have been studied complements the work of Storey (2003), Müller et al. (2004), and Bickel (2004). While we have assumed here that the test statistics are independent, we expect that the risk behavior of the estimated shrunken p-values will be robust to dependence such as exchangeable correlation due to the empirical Bayes construction. We are also currently studying estimators for the decision rules in Section 4.3.

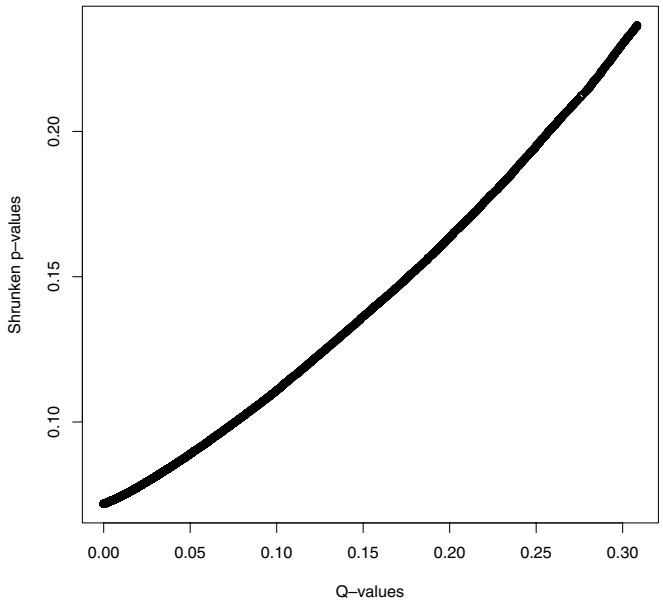

Figure 1. Plot of q-values using Storey (2002) method (horizontal axis) versus shrunken p-values from SPADE, where π_0 is estimated using the method of Storey and Tibshirani (2003).

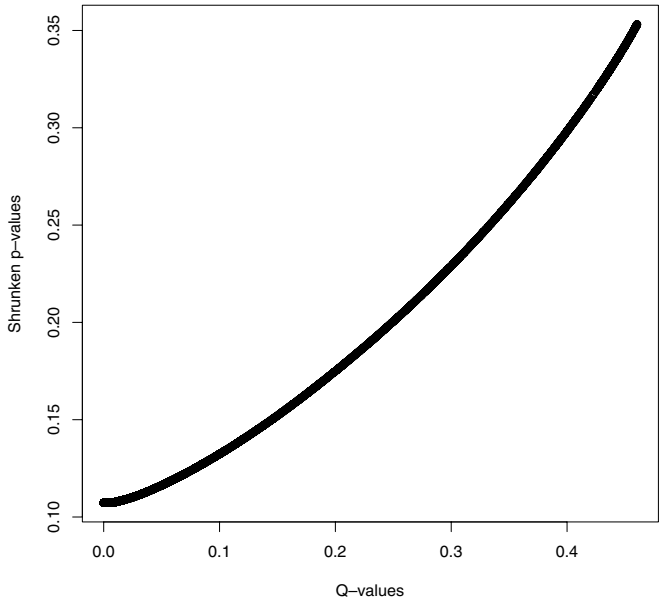

Figure 2. Plot of q-values using Pounds and Cheng (2004) method (horizontal axis) versus shrunken p-values from SPADE, where π_0 is estimated using the method of Pounds and Cheng (2004).

Because of the availability of software for FDR estimation procedures (Storey, 2002; Pounds and Cheng, 2004; Dalmasso et al., 2005), implementation of the SPADE methodology is very straightforward. `R` functions implementing the proposed procedures can be obtained from the author.

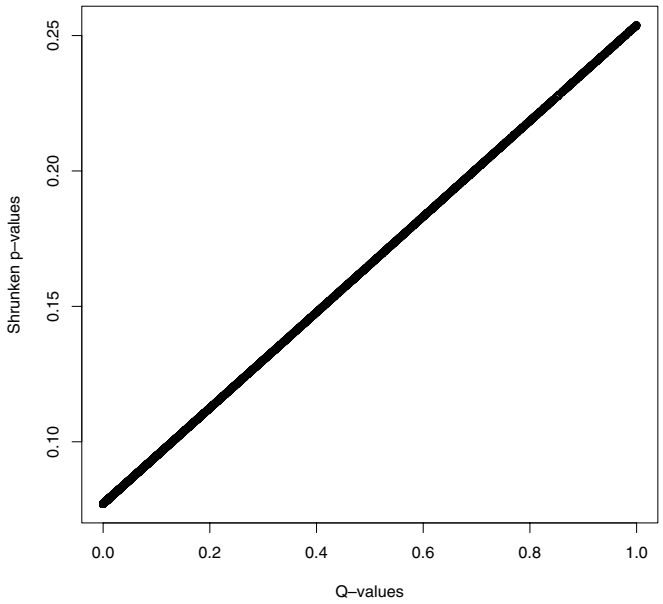

Figure 3. Plot of q-values using Dalmasso et al. (2005) method (horizontal axis) versus shrunken p-values from SPADE, where π_0 is estimated using the method of Dalmasso et al. (2005).

Acknowledgements

The author would like to thank Tom Nichols and Trivellore Raghunathan for helpful discussions. He would like to thank the associate editor and a referee whose comments substantially improved the article. This research is supported by grant GM72007 from the Joint DMS/DBS/NIGMS Biological Mathematics Program.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57,** 289–300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* **82,** 163–170.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29,** 1165–1188.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Bickel, D. R. (2004). Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? *Statistical Applications in Genetics and Molecular Biology* **3,** 8.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83,** 596–610.

Dalmasso, C., Broët, P., and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics* **21,** 660–668.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412,** 822–826.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96,** 1151–1160.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis (with discussion). *Test* **12,** 1–77.

Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15,** 373–390.

Genovese, C. and Wasserman, L. (2004). A stochastic approach to false discovery control. *Annals of Statistics* **32,** 1035–1061.

George, E. I. (1986). Minimax multiple shrinkage estimation. *Annals of Statistics* **14,** 188–205.

Hwang, J. T., Casella, G., Robert, C., Wells, M. T., and Farrell, R. H. (1990). Estimation of accuracy in testing. *Annals of Statistics* **20,** 490–509.

Keeney, R. L., Raiffa, H. A., and Meyer, R. F. C. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.

Müller, P., Parmigiani, G., Robert, C. P., and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association* **468,** 990–1001.

Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics* **20,** 1737–1745.

Sarkar, S. K. (2002). Some results on false discovery rates in multiple testing procedures. *Annals of Statistics* **30,** 239–257.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64,** 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31,** 2013–2035.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 9440–9445.

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. *Journal of the Royal Statistical Society, Series B* **66,** 187–205.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* **3,** 15.

Received April 2005. *Revised March* 2006.
Accepted March 2006.