Simultaneous estimation procedures and multiple testing: a decision-theoretic framework

Debashis Ghosh[*]

[*]University of Michigan, debashis.ghosh@ucdenver.edu

Simultaneous estimation procedures and multiple testing: a decision-theoretic framework

Debashis Ghosh

Abstract

There is recent tremendous interest in statistical methods regarding the false discovery rate (FDR). Two classes of literature on this topic exist. In the first, authors have proposed sequential testing procedures that control the false discovery rate. For the second, authors have studied the procedures involving FDR in a univariate mixture model setting. We consider a decision-theoretic approach to the assessment of FDR-based methods. In particular, we attempt to reconcile the current literature on false discovery rate procedures with more classical simultaneous estimation procedures. Formulation of the link will allow us to apply results from decision theory; we can then traverse between the two literatures. In particular, we propose double shrinkage estimators for the location parameter in the multiple testing problem for false discovery rates and provide conditions for obtaining minimaxity. We also describe a double shrinkage estimation procedure for p-values. Simulation studies are used to explore the risk properties of existing statistical methods and the potential gains of shrinkage. We then develop a procedure for calculating double shrinkage estimators from observed data. The procedures are applied to data from a gene expression profiling study in prostate cancer.

Simultaneous estimation procedures and multiple testing: a decision-theoretic framework

Debashis Ghosh

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Summary. There is recent tremendous interest in statistical methods regarding the false discovery rate (FDR). Two classes of literature on this topic exist. In the first, authors have proposed sequential testing procedures that control the false discovery rate. For the second, authors have studied the procedures involving FDR in a univariate mixture model setting. We consider a decision-theoretic approach to the assessment of FDR-based methods. In particular, we attempt to reconcile the current literature on false discovery rate procedures with more classical simultaneous estimation procedures. Formulation of the link will allow us to apply results from decision theory; we can then traverse between the two literatures. In particular, we propose double shrinkage estimators for the location parameter in the multiple testing problem for false discovery rates and provide conditions for obtaining minimaxity. We also describe a double shrinkage estimation procedure for p-values. Simulation studies are used to explore the risk properties of existing statistical methods and the potential gains of shrinkage. We then develop a procedure for calculating double shrinkage estimators from observed data. The procedures are applied to data from a gene expression profiling study in prostate cancer.

Keywords: Estimation Target; Hypothesis Testing; James-Stein Estimation; Multiple Comparisons; Simultaneous Inference.

1. Introduction

Because of technological developments in scientific fields (e.g., high-throughput genomics), experiments are now performed in which thousands of hypotheses are simultaneously tested. In problems dealing with multiple testing, the usual quantity that has been is the familywise error rate (FWER). One simple method for adjustment is Bonferroni's correction; many other methods are described by Westfall and Young (1993).

In the recent statistical literature, many authors have argued that control of the FWER is too stringent. One quantity that has been argued for instead is the false discovery rate (FDR), first proposed by Benjamini and Hochberg (1995).

The literature on false discovery rate procedures can be divided into two areas. For the first, the emphasis is on procedures that control FDR. A very simple procedure based on ordering the p-values of test statistics was proposed by Benjamini and Hochberg (1995), which we will describe in detail in Section 2. It was later shown in Benjamini and Yekutieli (2001) that the original Benjamini-Hochberg procedure controls FDR under a certain dependence. The Benjamini-Hochberg procedure is a step-down testing procedure; related testing procedures have been studied by Benjamini and Liu (1999), Benjamini and Yekutieli (2001) and Sarkar (2002). In much of this literature, the focus is on constructing a sequential testing procedure and demonstrating that it controls the false discovery rate.

The second class of false discovery rate procedures is based on estimation of the false discovery rate directly. This has been the approach adopted by Efron et al. (2001), Storey (2002, 2003), and Genovese and Wasserman (2002). These two classes of methods have been unified by Storey et al. (2004) and Genovese and Wasserman (2004), who proposed thresholding procedures based on the estimated distribution of the false discovery rate.

An attractive feature of the false discovery rate procedures mentioned is that it provides the data analyst a post-data assessment of the strength of evidence available in the dataset. To be concrete, based on the number of hypothesis the user rejects, the false discovery rate is interpretable as the expected number of hypotheses that have been falsely rejected. Given the number of hypotheses that are being tested in large-scale high-throughput scientific studies, it seems natural that post-data assessments are of interest to investigators so that they may determine which hypotheses should be followed up in further studies.

In this paper, we consider the use of decision theory to study the behavior of multiple testing procedures. Some work has been made in this area by Storey (2003), Bickel (2003) and Müller et al. (2004). In Storey (2003, Section 6), a link between an FDR-related quantity, the positive false discovery rate (pFDR), and classification theory is studied. The pFDR is shown to be a Bayes rule in this setting. The work of Bickel (2003) is to estimate another FDR-type quantity using principles of decision theory, while that of Müller et al. (2004) is to motivate sample size determination from a decision-theoretic point of view. The approach in this paper is quite different from these works. First, we seek to unify the current literature in false discovery rate estimation procedures with that of minimax estimation (Brown, 1971; Stein, 1981; Lehmann and Casella, 1997). Second, the decision-theoretic framework will allow to derive new procedures for testing multiple hypotheses that have desirable risk properties. It should be noted that while we will make use of prior and posterior distributions here, the decision-theoretic point of view taken in the paper is not Bayesian. Rather, the use of such distributions helps us to assess risk properties of various estimation procedures. The structure of the paper is as follows. Multiple testing concepts and false discovery rate procedures are reviewed in Section 2. In Section 3, we propose a two-stage model for estimation of the location parameter and relate it to false discovery rates. This link allows for the applications of various decision theoretic notions, such as minimaxity and Bayes rules, which are explored in Section 4. We also propose a construction of double shrinkage estimators and study the potential for gains using shrinkage. In Section 5, we propose an estimation procedure for calculating double shrinkage estimators using observed data. We illustrate this methodology using gene expression data from a microarray experiment in Section 6. Finally, we conclude with some discussion in Section 7.

2. Multiple Testing Background

Suppose we have test statistics $T_1, \ldots, T_n$ for testing hypotheses H_{0i}, $i = 1, \ldots$,n. First, we give a brief review of simultaneous hypothesis testing and the false discovery rate.

2.1. Multiple Testing Procedures

We wish to test a set of n hypotheses; of these n hypotheses, the number of true null hypotheses is m_0. Suppose we wish to cross-classify hypotheses based on whether or not it is a true null and whether or not it is rejected using a statistical test. This can be conceptualized using the following 2×2 contingency table:

[Note: Table 1 about here.]

Based on Table 1, the FWER is defined as $P(V \geq 1)$. Further dicussion for FWER-controlling procedures can be found in Ge et al. (2003) and in a collection of papers by Van der Laan and colleagues (Dudoit et al., 2004; van der Laan et al., 2004a,b).

The definition of false discovery rate (FDR) as put forward by Benjamini and Hochberg (1995) is

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

The conditioning on the event $[Q > 0]$ is needed because the fraction V/Q is not well-defined when $Q = 0$. Methods for controlling the false discovery rate have been proposed by several authors (Benjamini and Hochberg, 1995; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001, Sarkar, 2002).

Storey (2002) suggests use of the positive false discovery rate (pFDR), defined as

$$pFDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right].$$

Conditional on rejecting at least one hypothesis, the pFDR is defined to be the fraction of rejected hypotheses that are in truth null hypotheses. Note that pFDR and FDR are linked by the following relationship:

$$\text{pFDR} = \frac{\text{FDR}}{P(Q > 0)}.$$

2.2. Mixture Model Motivation of FDR

An alternative approach involving false discovery rate estimation procedures has been to estimate the false discovery rate directly. Define indicator variables $H_1, \ldots, H_n$ corresponding to $T_1, \ldots, T_n$, where $H_i = 0$ if the null hypothesis is true and $H_i = 1$ if the alternative hypothesis is true. Assume that $H_1, \ldots, H_n$ are a random sample from a Bernoulli distribution where $P(H_i = 0) = \pi_0$, $i = 1, \ldots, n$. We define the densities f_0 and f_1 corresponding to $T_i|H_i = 0$ and $T_i|H_i = 1$, $(i = 1, \ldots, n)$. The marginal density of the test statistics $T_1, \ldots, T_n$ is given by

$$f(t) \equiv \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \qquad (1)$$

The mixture model framework represented in (1) has been used by several authors to study the false discovery rate (Efron et al., 2001; Storey, 2002; Genovese and Wasserman, 2004; Storey et al., 2004; Cox and Wong, 2004). A result of Storey (2002) is the following:

$$\begin{aligned} pFDR(R) &= P(H = 0|T \in R) \\ &= \frac{\pi_0 P(T \in R|H = 0)}{P(T \in R)}. \end{aligned}$$

Provided one can estimate π_0, methods for false discovery rate estimation have been developed by several authors (Efron et al., 2001; Storey, 2002). In situations, for large n $P(Q > 0)$ will tend to one. Then the difference between pFDR and FDR should be asymptotically negligible.

While we assume here that the test statistics are independent, authors such as Storey et al. (2004) and Genovese and Wasserman (2004) have shown that the estimation procedure for the false discovery rate will be asymptotically unbiased under various forms of dependence. Intuitively, this makes sense because the the false discovery rate is a probability and hence a mean of an indicator function. Using probability tools such as ergodicity theory, estimates of means are fairly robust to various forms of dependence.

To embed the Bejamini-Hochberg procedure within this estimation framework, Storey et al. (2004) and Genovese and Wasserman (2004) consider a class of thresholding procedures. Define the following threshold function:

$$c_\alpha(F) = \sup\{0 \leq t \leq 1 : F(t) \leq \alpha\},$$

where F is a function. Based on an estimator of F, one obtains an estimator of the thresholding function. Storey et al. (2004) and Genovese and Wasserman (2004) consider estimates of F based on the estimated FDR. It can be shown that several methods of multiple testing correction, such as Bonferroni's method and the Benjamini-Hochberg method, can be expressed in terms of such a thresholding procedure. Asymptotic results for the thresholding procedures are obtained by Storey et al. (2004) and Genovese and Wasserman (2004).

3. Decision Theory and False Discovery Rates: A Connection

Decision theory is an area with a long history in statistics (Raiffa and Schlaifer, 1961; Ferguson, 1967). Much work has been focused on developing estimation procedures, or more generally decision procedures, that have desirable risk properties. It is crucial to think of estimators as estimating a population parameter; what decision theory allows for is evaluation of risk properties of such estimators. Generically, we let θ to be the population parameter to be estimated, d an estimator and $L(\theta, d)$ the loss function. The risk function is defined by

$$R(\theta, d) = E\{L(\theta, d)\},$$

where the expectation is taken with respect to the distribution of the data.

Let us start by considering the following two-stage model:

$$\begin{array}{rcl} T_i|\mu_i & \overset{ind}{\sim} & N(\mu_i, 1) \\ \mu_1, \ldots, \mu_n & \overset{iid}{\sim} & F, \end{array} \tag{2}$$

where μ_i is the mean of T_i, and F is some distribution function. Thus, model (2) specifies a two-stage model for the joint distribution of $(T_1, \ldots, T_n)$. Note that while the first stage of (2) specifies conditional independence of the T_i, marginally the joint distribution has an exchangeable correlation structure and hence are dependent. Also we are now viewing T_i as an estimator of μ_i, $i = 1, \ldots, n$.

Suppose we take F to be the cumulative distribution function for a degenerate point mass at μ. Then $T_1, \ldots, T_n$ represent a random sample from a normal distribution with mean μ and variance one. From a decision-theoretic point of view, this is a well-studied problem in statistics; for squared error loss, the sample mean is known to be the minimax estimator that is admissible (Lehmann and Casella, 1997, §5.1).

As an extension of the previous example, let us take F in (2) to be a normal distribution with mean μ and variance τ^2, where $\tau^2 > 0$ is known. This is the classical normal random effects model for T. The effect of the second of this model is to shrink the estimates of μ_i towards a common mean μ by borrowing strength from other observations. A classical estimator in this situation that performs well is the following estimator, proposed by Lindley and Smith (1972):

$$\tilde{\mu}_i = \left(1 - \frac{1}{1+\tau^2}\right) T_i + \frac{1}{1+\tau^2} \bar{T}_n,$$

where $\bar{T}_n = n^{-1}\sum_{i=1}^{n} T_i$. Note that there is shrinkage of the univariate statistic T_i ($i = 1, \ldots, n$) towards the population mean of the statistics, which is estimated based on the sample using the sample mean.

Let us now take F in (2) to be

$$F = \pi_0 F_{\mu_0} + (1 - \pi_0) F_{\mu_1},$$

where F_{μ_0} and F_{μ_1} are the cumulative distribution functions for the degenerate point mass distributions at μ_0 and μ_1. Plugging into (2), this implies that

$$T_i \overset{iid}{\sim} \pi_0 N(\mu_0, 1) + (1 - \pi_0) N(\mu_1, 1). \tag{3}$$

We have a special case of the mixture model for false discovery rate where f_0 and f_1 are densities for $N(\mu_0, 1)$ and $N(\mu_1, 1)$ random variables, respectively. This model was studied in some detail by Cox and Wong (2004).

In the multiple testing literature, model (3) would arise in a situation where we wished to test n hypotheses of the form $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, i.e. testing a simple null hypothesis versus a simple alternative, where the distribution of the test statistic is normal with mean μ and variance one. This type of structure might also arise in testing one-sided null hypotheses versus one-sided alternatives and simple null hypotheses versus one-sided alternatives in the situation where the distribution of the T_i exhibits the monotone likelihood ratio property (Lehmann, 1986, p. 78).

We can generalize (3) to allow for unequal variances:

$$T_i \overset{iid}{\sim} \pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_1, \sigma_1^2). \tag{4}$$

To extend the decision theoretic viewpoint here, what the mixture distribution for F as manifested in (3) and (4) does is to provide two targets for shrinkage: μ_0 and μ_1. Thus, the multiple testing framework considered in the FDR literature provides a natural statistical model for which to develop shrinkage estimators and to study their risk properties. Note that a complication of this multiple testing framework is that we must shrink towards two targets.

4. Simultaneous Estimation Procedures: Theoretical Results

An alternative to the FDR procedures that would address the multiple testing issue in (4) is to construct shrinkage estimators that target the two distributions. We now demonstrate how to do this using (4); note that we are considering T_i ($i = 1, \ldots, n$) to be estimators of the location parameter μ. Assume that μ_0 and μ_1 are known and that the variances are unknown. No shrinkage gains are possible in (3) because of the equal variances for the two component distribution of the mixture model. Note that in order to apply the decision theoretic framework, we have to view the statistics as estimating a population quantity. We will first start by considering an estimation of the location parameter. Another point of departure from previous work on FDR estimation is that we will assume that μ_0, μ_1, π_0 and π_1 are known. We will study a procedure for data analysis where π_0 is estimated in Section 6.

To construct a shrinkage estimator of μ in model (4), we calculate it relative to each of the component of the mixture distribution and then mix the estimators with weights. With respect to the first component, a shrinkage estimator is given by

$$T_{0i}^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^n (T_i - \mu_0)^2}\right](T_i - \mu_0), \tag{5}$$

while for the second component, a shrinkage estimator is given by

$$T_{1i}^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^n (T_i - \mu_1)^2}\right](T_i - \mu_1) \tag{6}$$

for $i = 1, \ldots, n$. A shrinkage estimator combining (5) and (6) is then given by $T_i^{JS} = \pi_0(T_i)T_{0i}^{JS} + \pi_1(T_i)T_{1i}^{JS}$, $i = 1, \ldots, n$, where

$$\pi_k(T_i) = \frac{\pi_k f_k(T_i)}{\pi_0 f_0(T_i) + \pi_1 f_1(T_i)}. \tag{7}$$

and f_0 and f_1 refer to the marginal densities of the distribution of the test statistics under the null and alternative hypotheses from (1). We will refer to this as a double shrinkage estimator. Note that T_i is shrunk towards both μ_0 and μ_1 by construction. Let us consider (7) further with $k = 0$. If m_0 and m_1 equal f_0 and f_1 in (1), then for $k = 0$, (7) represents the local false discovery rate (Efron and Tibshirani, 2002). There is thus an intimate connection between the double shrinkage estimators with the false discovery rate; in particular, the shrinkage weights are based on the local false discovery rate. This interpretation does not exist when f_0 and f_1 are not density functions corresponding to the null and alternative hypotheses.

Efron et al. (2001) considered the following test statistic in the microarray setting:

$$\tilde{T}_i = \frac{\hat{T}_i^{num}}{\hat{T}_i^{den} + a_0}, (i = 1, \ldots, n) \tag{8}$$

where T_i^{num} and T_i^{den} represent the numerator and denominator of the ith statistic, and a_0 is a percentile of the empirical distribution of $T_1^{den}, \ldots, T_n^{den}$. We can view $\tilde{T}_1, \ldots, \tilde{T}_n$ as estimators of $\mu_1, \ldots, \mu_n$. The adjustment in the denominator in the test statistics $\tilde{T}_1, \ldots, T_n$ achieves shrinkage for the multiple testing situation; however, the "fudge factor" in (8) is not based on a formal probabilistic model. By contrast, the statistical framework described for the construction of $T_1^{JS}, \ldots, T_n^{JS}$ leads to shrinkage in a more principled manner.

To guide us in the construction of double shrinkage estimators, it is necessary to think of available optimality properties. The first optimality property to consider is Bayes rules. Bayes rules or Bayes estimators minimize the posterior expected loss, conditional on data, with respect to some prior distribution. Note that the definition of Bayes rules is contingent on the choice of loss function used. For squared error loss of a functional of parameter μ, $g(\mu)$, the Bayes rule is given by the posterior mean $E[g(\mu)|\mathbf{T}]$, while for L_1 error, the Bayes rule is given by the posterior median of $g(\mu)$, conditional on data.

The mixture model for false discovery rates is (2) with F being a mixture of two distributions, i.e. $F = \pi_0 F_0 + \pi_1 F_1$. We can then construct a Bayes estimator of $g(\mu)$ by finding the Bayes estimator with respect to F_0 and F_1 and convolving the result. Let $\tilde{g}_0(\mu)$ and $\tilde{g}_1(\mu)$ denote the Bayes estimators of $g(\mu)$ with respect to F_0 and F_1. Then the Bayes estimator of $g(\mu)$ with respect to F is given by

$$\tilde{g}(\mu) = \pi_0\{\tilde{g}_0(\mu)\}\tilde{g}_0(\mu) + (1 - \pi_0\{\tilde{g}_0(\mu)\})\tilde{g}_1(\mu).$$

For example, if we take F_0 to be the cdf for a normal random variable with mean θ_0 and variance η_0^2 and F_1 that for a normal random variable with mean θ_0 and variance η_1^2, then under both squared error loss and L_1 loss, the Bayes estimator of μ is

$$\pi_0\left(\sum_{i=1}^n T_i\right)\frac{\sum_{i=1}^n T_i+\mu_0/\eta_0^2}{n+1/\eta_0^2}+\left\{1-\pi_0\left(\sum_{i=1}^n T_i\right)\right\}\frac{\sum_{i=1}^n T_i+\mu_1/\eta_1^2}{n+1/\eta_1^2}.$$

Another optimality property that can be used to evaluate estimators is minimaxity. As described in Lehmann and Casella (2002, §5.1, p. 309), an estimator δ^M of $g(\mu)$ which minimizes the maximum risk (expected loss), i.e. which satisfies

$$\inf_\delta \sup_\mu R\{g(\mu),\delta\} = \sup_\mu R\{g(\mu),\delta^M\},$$

is said to be minimax. We focus on the situation where g is the identity function and the loss function is quadratic and seek to characterize the class of minimax estimators. To do this, we need some more mathematical background. Define m to be a function from R^n to R. We define the differential operator $\nabla m \equiv (\nabla_1 m,\ldots,\nabla_n m)$ to be the function from R^n to R such that for all $z \in R^n$,

$$m(t+z)-m(t)=\int_0^1 t'\nabla m(t+yz)dy.$$

A function is superharmonic if

$$\nabla^2 m(t)=\sum_{i=1}^n \nabla_i^2 m(t) \leq 0.$$

We consider estimators of the following form:

$$\hat{T}_{ki} = T_i + \nabla \log m_k(T_i), \qquad i=1,\ldots,n; k=0,1 \tag{9}$$

where m_0 and m_1 are functions that are twice differentiable. As shown in Brown (1971), estimators of the form (9) generate a wide class of rules, as it contains Bayes and admissible rules. A generalized double shrinkage estimator is given by

$$\hat{T}_i = \sum_{k=0}^1 c_k(T_i)\hat{T}_{ki} \tag{10}$$

where

$$c_k(T_i) = \frac{\pi_k m_k(T_i)}{\pi_0 m_0(T_i)+\pi_1 m_1(T_i)}, \tag{11}$$

for $i=1,\ldots,n$ and $k=0,1$.

The next step is to characterize the class of minimax double shrinkage estimators. We have the following theorem from George (1986):

Theorem 1: *Define $\hat{T}_{ki}$ as in (9). If m_k and ∇m_k are differentiable, m_k is superharmonic and satisfies the following conditions:*

$$E|\nabla_i^2 m_k(\mathbf{T})/m(\mathbf{T})| < \infty, \quad i=1,\ldots,n \tag{i}$$

and

$$E\|\nabla \log m_k(\mathbf{T})\|^2 < \infty \qquad (ii),$$

then for a fixed k, $\hat{T}_{ki}$ $(i = 1, \ldots, n)$ *is minimax.*

Theorem 1 provides sufficiency conditions to check for the minimaxity of T_{ki} $(i = 1, \ldots, n)$ for a given k. The conditions that m_k $(k = 0, 1)$ must satisfy in Theorem 1 are very similar to the regularity conditions that densities must satisfy for the usual asymptotic results for maximum likelihood estimation procedures. Recall, however, that the mixture model for false discovery rate consists of two components and that we want to perform shrinkage in two directions, corresponding to each component of the mixture. The following lemma is immediate from Theorem 1.

Lemma 1: If $\hat{T}_{ki}$ satisfy the conditions of Theorem 1, and m_0 and m_1 are superharmonic, then $\hat{T}_i$, defined in (10) is minimax.

Proof: Note that if m_0 and m_1 are superharmonic, then $\sum_{k=0}^{1} \pi_k m_k$ will also be superharmonic. By Theorem 1, (10) will be minimax.

Based on the results of Theorem 1 and Lemma 1, we have constructed a class of estimators that are optimal in the sense of minimizing the worst-case risk scenario. This can be viewed as providing some robustness to the estimation procedure.

To study the potential gains of shrinkage, we performed a simulation study. We considered estimation of the location parameter. The two-group problem was studied in which measurements on $m \equiv 10$ individuals for each group was generated from a normal distribution; the distribution for group i is normal with mean η_i and variance one, $i = 0, 1$. Note that the target estimand in this setting is $\mu \equiv \eta_1 - \eta_0$. In this setting, we took $\pi_0 = 0.1, 0.5$ and 0.8. We considered three situations:

- **Small:** We take μ to be 0.25 with probability 0.75 and 0.5 with probability 0.25.
- **Medium:** We take μ to be 0.25 with probability 0.5 and 0.5 with probability 0.5.
- **Large:** We take μ to be 0.5 with probability one.

The proposed method of Efron et al. (2001) was used, along with the double shrinkage estimators. However, the true value of π_0 was used for the for the weights, i.e. $\pi_0(t) = \pi_0$ instead of (7). Thus, we are not incorporating the data-adaptive nature of the weights at this stage; how to estimate this is considered further in Section 5. With regards to the target in the double shrinkage estimators, we considered two situations. The first is where the true target is used and the second where the target is misspecified. The misspecified target is taken to be one. The mean-squared error results are shown in Table 1. Based on the true target results, there is a major increase in risk by using the Efron et al. (2001) statistics. Even when the target is misspecified, the double shrinkage estimator leads to a risk reduction relative to the Efron et al. statistic. Note that the risk reduction occurs even when using non-data-adaptive weights. This suggests that shrinkage towards two targets offers advantages relative to shrinkage towards one in this multiple testing framework.

5. Double shrinkage estimation

We now seek to construct a double shrinkage estimator using the observed data. While George (1986) discusses theoretical aspects of multiple shrinkage estimators, he does not give methods for their calculation using data. Note that estimation of a double shrinkage estimator requires that the distribution of the test statistic under the null and alternative hypotheses be estimated, as well as π_0. One method would be to assume parametric forms for the component densities in (1) and to fit a finite mixture model to $T_1, \ldots, T_n$. However, we seek more flexible models that attempt to use as much of the data distribution for the test statistics as possible. To do this, we utilize a density estimation method proposed by Efron (2004). He was not addressing the problem of construction of shrinkage estimators for multiple testing but rather the issue of estimating effect sizes in a false discovery rate framework.

Note from (1) that we have

$$\pi_1 f_1(t) = f(t) - \pi_0 f_0(t).$$

We can estimate $f(t)$ by applying density estimation methods to the "data" $T_1, \ldots, T_n$. For estimation of $\pi_0 f_0(t)$, the zero assumption in Efron (2004) is utilized. What this means is that most test statistics with a value near zero comes from the null distribution component. The assumption implies that one can use a normal-based moments matching technique as described in Efron (2004) to obtain an estimate of π_0 and $f_0(t)$. Given the estimate of $f(t)$ and $\pi_0 f_0(t)$, we then obtain an estimate of π_1 and $f_1(t)$ by simple subtraction. In the example presented in the next section, we utilized this procedure. The interested reader is referred to Efron (2004) for numerical details.

Based on the estimates of $f_0(t)$, $f_1(t)$ and π_0, we can estimate $T_1^{JS}, \ldots, T_n^{JS}$ by

$$\hat{T}_i^{JS} = \hat{\pi}_0(T_i)\hat{T}_{0i}^{JS} + \{1 - \hat{\pi}_0(T_i)\}\hat{T}_{0i}^{JS}, \tag{12}$$

where

$$\hat{T}_{0i}^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^n (T_i - \hat{\mu}_0)^2}\right](T_i - \hat{\mu}_0),$$

$$\hat{T}_{1i}^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^n (T_i - \hat{\mu}_1)^2}\right](T_i - \hat{\mu}_1),$$

$$\hat{\pi}_k(t) = \frac{\hat{\pi}_k \hat{f}_k(t)}{\hat{\pi}_0 \hat{f}_0(t) + (1 - \hat{\pi}_0)\hat{f}_1(t)},$$

$\hat{\mu}_0 = \int t d\hat{F}_0(t)$ and $\hat{\mu}_1 = \int t d\hat{F}_1(t)$. Thus, we have provided a method for practical implementation of double shrinkage estimators for test statistics. We reiterate that in George (1986), no algorithms for estimating (7) or μ_0 and μ_1 are given. In addition, we have linked up the double shrinkage estimators to the multiple testing problem.

6. P-values: Theoretical Considerations and Double Shrinkage Estimation

In the original paper by Storey (2002), the test statistics used for testing the hypotheses $H_1, \ldots, H_n$ were the p-values. The false discovery rate was estimated on the basis of the p-values and estimating π_0, the proportion of true null hypotheses, using a permutation scheme.

It is a bit more problematic to come up with a population "parameter" estimated by a p-value. We follow the approach of Hwang et al. (1990) and consider the p-value to be an estimator of the probability of the null hypothesis. Equivalently, we can consider estimators of the expected value for the indicator function corresponding to the null hypothesis being true. If we let $p_1, \ldots, p_n$ denote the p-values for testing $H_{01}, \ldots, H_{0n}$, then the model induced by (3) is

$$p_1, \ldots, p_n \overset{iid}{\sim} \pi_0 F_U + (1 - \pi_0) F_V, \tag{13}$$

where F_U is the cdf of a $U \equiv \text{Uniform}(0, 1)$ random variable and F_V is that of a random variable stochastically smaller than U.

To cast these statistics into a decision theoretic framework, we consider a multivariate generalization of the work of Hwang et al. (1990), who considered a decision-theoretic approach to the hypothesis testing problem. Suppose that the null and alternative hypotheses can be phrased as

$$H_0 : \mu \in \Theta \text{ versus } H_1 : \mu \in \Theta^c.$$

Our target parameter to estimate is $I(\mu \in \Theta)$, i.e. the indicator function for the parameter being in the set described by the null hypothesis. A loss function that is a natural multivariate generalization of that of Hwang et al. (1990) is

$$L(\mu, d) = \sum_{i=1}^{n} |I(\mu_i \in \Theta) - d(T_i)|^k \tag{14}$$

for $k = 1, 2$; the choices of k correspond to L_1 and L_2 error loss functions. If were to look for Bayes rules with respect to (14) under the assumption that the μ_i $(i = 1, \ldots, n)$ are independent but not identically distributed, then the Bayes rules are the usual componentwise Bayes rules. For L_2 loss, the Bayes rule is $P(\mu_i \in \Theta | T_i)$ and for L_1 loss, it is $I\{P(\mu_i \in \Theta | T_i) \geq 1/2\}$, $i = 1, \ldots, n$. If we were to construct Bayes rules with the fact that $\mu_i \equiv \mu$ and that the T_i are iid, then the Bayes rules under L_1 and L_2 error are $P(\mu \in \Theta | \mathbf{T})$ and $I\{P(\mu \in \Theta | \mathbf{T}) \geq 1/2\}$. Storey (2003) derived this type of result for L_1 error with a slightly different conditioning event. Note that the Bayes rules for L_2 error is simply the local false discovery rate (Efron and Tibshirani, 2002). The Bayes rules for L_1 error loss have implicitly assumed that the cost of a Type I error and Type II error are equal; generalization to the situation of unequal costs is easy. It is interesting to note, however, that the L_1 loss function is not a proper scoring loss function (Schervish, 1989); we thus focus attention on L_2 loss.

Mimicking what was done in the earlier sections, an approach to constructing double shrinkage estimators for $I(\mu_i \in \Theta)$ is to calculate for $i = 1, \ldots, n$,

$$p_i^{JS} = \pi_0(p_i) p_{0i}^{JS} + \{1 - \pi_0(p_i)\} p_{1i}^{JS}, \tag{15}$$

where

$$p_{0i}^{JS} = p_i - \left[1 \wedge \frac{(n-1)}{12 \sum_{i=1}^{n} (p_i - 1/2)^2}\right] (p_i - 1/2) \tag{16}$$

$$p_{1i}^{JS} = p_i - \left[1 \wedge \frac{(n-1)\sigma_V^2}{\sum_{i=1}^{n} (p_i - d)^2}\right] (p_i - \mu_V), \tag{17}$$

$$\pi_0(p) = \frac{\pi_0 p}{\pi_0 p + (1 - \pi_0) F_V(p)}$$

$(p_1, \ldots, p_n)$ are the n p-values, and μ_V and σ_V^2 are the mean and variance of the p-values under the alternative hypothesis. 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. The mixture distribution of the p-values is providing two targets for shrinkage.

We sought to study the potential gains in risk from adopting a double shrinkage estimation framework using simulation studies. Here, p-values were generated from model (3) with F_0 being the cdf for a uniform [0,1] distribution and F_1 being the cdf for a Beta distribution. As in the previous paragraph, for each simulation setting, we generated 1000 datasets; each simulated dataset consisted of $n = 10000$ values. Again, both true and misspecified targets were used. The misspecified target was taken to be a mean of 0.2 with a variance of 0.01. We let $\pi_0 = 0.2, 0.5$ and 0.8. In this setting, we took $\pi_0 = 0.2, 0.5$ and 0.8. We considered three situations:

- **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. We term this a Beta$(3, 4)$ distribution.
- **Medium:** A mixture of a Beta$(3, 4)$ and Beta$(3, 15)$ distribution with mixing proportions 0.5 and 0.5, respectively.
- **Large:** A Beta$(3, 15)$ distribution. This gives a mean of $3/18$ and a variance of f for the distribution of p-values under the alternative hypothesis.

The q-value estimation procedure proposed by Storey (2002) was used; the shrunken p-value (15) was also used with the true π_0. The mean-squared error results are shown in Table 3. We found the q-value method to be very competitive with the double shrinkage estimator for the p-value. Given that the q-value method is based on the pFDR, which is a Bayes rule under L_1 error loss (Storey, 2003), its performance is not surprising. The q-value appears to work well when π_0 is relatively large. However, it appears that for π_0 being small, there are potential gains to be had by using the double shrinkage adjusted p-value. These results seem quite consistent across the various effect size situations. Note that an adaptive weight for π_0 was not used; this suggests that there will be major risk gains by using a data-dependent estimator of $\pi_0()$.

For implementing the shrunken p-value procedure with observed data, estimators of π_0 and $F_V(p)$ are needed. To estimate these quantities, observe the following relationship

$$F_P(p) = \pi_0 p + (1 - \pi_0)F_V(p),$$

where $F_P(p)$ is the cumulative distribution function for $p_1, \ldots, p_n$. Solving for F_V, we get

$$F_V(p) = \frac{F_P(p) - \pi_0 p}{1 - \pi_0}. \tag{18}$$

We can estimate F_P using the empirical distribution function of the observed p-values. Provided we can estimate π_0, we can estimate the cdf of V. There exist many choices of estimators for π_0 in the literature. Due to its popularity in the area of genomic data analysis, we choose the Q-VALUE algorithm by Storey and Tibshirani (2003). It is summarized as follows:

(a) Order the G p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(G)}$.
(b) Construct a grid of L λ values, $\lambda_1, \ldots, \lambda_L$ and calculate

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

$l = 1, \ldots, L$.

(c) Fit a cubic smoothing spline to the values $\{\lambda_l, \hat{\pi}_0(\lambda_l)\}$, $l = 1, \ldots, L$.

(d) Estimate π_0 by the interpolated value at $\lambda = 1$.

Based on the resultant estimator of π_0, $\hat{\pi}_0$, we can estimate $F_V(p)$ by

$$\hat{F}_V(p) = \frac{\hat{F}_P(p) - \hat{\pi}_0 p}{1 - \hat{\pi}_0}.$$

Then an estimator of the shrunken p-value is given by

$$\hat{p}_i^{JS} = \hat{\pi}_0(p_i) p_{0i}^{JS} + \{1 - \hat{\pi}_0(p_i)\} \hat{p}_{1i}^{JS}, \tag{19}$$

where

$$\hat{p}_{1i}^{JS} = p_i - \left[1 \wedge \frac{(n-1)\hat{\sigma}_{p1}^2}{\sum_{i=1}^n (p_i - \hat{\mu}_V)^2} \right] (p_i - \hat{\mu}_V),$$

$$\hat{\pi}_0(p) = \frac{\hat{\pi}_0 p}{\hat{\pi}_0 p + (1 - \hat{\pi}_0)\hat{F}_V(p)} \tag{20}$$

and $\hat{\mu}_V$ and $\hat{\sigma}_{p1}^2$) are the estimated mean and variance of V. Thus, we have constructed a double shrinkage estimator for p-values. One difference for the double shrinkage estimator in this situation relative to that for the test statistics is that the distribution of the p-values under the null hypothesis is completely independent of parameters. Thus, only the mean and variance of the p-values under the alternative hypothesis needs to be estimated here. If we took the density functions instead of the cdf of the component distributions in (20), then we would have the local false discovery rate using the p-value.

Note that we are implicitly assuming a squared error loss function here. While this is natural for estimators of the mean, it might not be as appropriate for p-values. If the shrunken p-value is less than zero, we will threshold it at zero. One tempting alternative is to try a transformation of the p-values that unconstrains the range (e.g. log(-log(1-pvalue))), calculate the double shrinkage estimator for the transformed p-value and then backtransform. The problem with this approach is that the uniform(0,1) distribution of the p-value under the null hypothesis is lost, and the null distribution of the transformed p-value will not in generally be analytically tractable. This remains a topic for further study.

7. Microarray example

We now apply the proposed methodology to a microarray profiling experiment in prostate cancer (Dhanasekaran et al., 2001; Varambally et al., 2002), Using 10K cDNA microarrays, the investigators have profiled tissue samples from various stages of prostate cancer (normal adjacent prostate, benign prostatic hyperplasia, localized prostate cancer, advanced metastatic prostate cancer). In addition to the gene expression profiles for a sample, the investigators have access to several other clinical parameters, such as Gleason score, survival time and status, and time to PSA recurrence. Throughout the profiling studies, one of the hypotheses made by investigators is that there exists a set of genes that distinguish aggressive prostate cancer from non-lethal prostate cancer. To begin to address this, a fairly standard analysis would be to determine which genes are differentially expressed between aggressive prostate cancer from nonaggressive prostate cancer. While various definitions of aggressiveness could be considered, 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.

Measurements were made on $n = 9984$ genes for 79 individuals. There are 59 localized prostate cancers and 20 metastatic prostate cancer samples. Before analyzing the data, we took the following preprocessing steps:

(a) Genes that were reported as missing in more than 10% of samples were filtered out.
(b) Genes that had a sample variation less than 0.05 across all samples were filtered out.

This left a total of $n = 6040$ genes available for analysis.

We first calculated t-statistic numerators comparing gene expression in localized versus metastatic prostate cancer samples. Note that this corresponds to estimating the population quantity of the mean difference in expression between two groups. We then applied the Efron (2004) procedure for estimating the distribution of the null density $f_0(t)$; the results are given in Figure 1. The red line represents the empirical null density of Efron (2004), which we take to be our estimate of $f_0(t)$. The blue line is the density of the observed t-statistic numerators, which is our estimate of $f(t)$. We also estimate π_0 to be 0.995. Based on the estimates of f_0, f, and π_0, we can calculate $f_1(t)$ in (3) by subtraction. The resulting shrunken statistics, compared to the original statistics, are given in Figure 2. Note that the shrinkage estimation works in this example like shrinkage towards one target because the estimate of π_0 is large.

The methodology using shrunken p-values is illustrated next. 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 estimate of π_0 using the QVALUE algorithm of Storey and Tibshirani is approximately 0.30. Based on this, we estimate F_V; the mean and variance are given by 0.17 and 0.07. We can then construct shrunken p-value estimators using (19). Histograms of the original and the shrunken p-values are provided in Figure 3. Note that we have some estimated negative p-values, so we treat them as zero. Because of the shrinkage, many p-values that were initially nonsignificant now become significant.

8. Discussion

In this article, we have provided a reinterpretation of the multiple testing problem in terms of estimation targets that allows for consideration of a decision-theoretic framework. This framework also motivates the double shrinkage methods proposed in this article. In particular, shrunken t-statistics and shrunken p-values, which are analogs of James-Stein estimators, are considered. It is shown that shrinkage towards the two targets that comprise the mixture distribution potentially leads to better risk behavior than existing procedures. While shrinkage towards multiple targets was studied from a risk point of view by George (1986), in this paper, we extend that view to the study of p-values and to actual computation using observed data.

With the explosion of high-dimensional hypothesis testing problems, we find that there is a great opportunity for pooling information across hypotheses using the mixture model framework described here. The shrinkage provides a natural method for adjusting for the multiple testing problem. While we have focused primarily on using t-statistics in this paper, the methodology is fairly flexible and could work with any Wald-type statistic.

We also find in our examination that there is a natural connection between the false discovery rate with weight functions for the shrinkage estimators. This gives a natural intuition as to why shrinkage of estimators will work for the multiple testing problem considered here.

Finally, this study also provides further justification of the optimality of posterior probabilities of hypotheses, conditional on data. The results in this paper complement those of Storey (2003) and Müller et al. (2004).

Acknowledgments

The author would like to thank Tom Nichols and Trivellore Raghunathan for helpful discussions. 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. *J. R. Statist. Soc. 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. *J. Statist. Plan. Inf.*, **82**, 163 – 170.

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

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.

Brown, L. D. (1971) Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Ann. Math. Stat.*, **42**, 855 – 903.

Cox, D. R. and Wong, M. Y. (2004) A simple procedure for the selection of significant effects. *J. R. Statist. Soc. B*, **66**, 395 – 402.

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.

Dudoit, S., van der Laan, M. J. and Pollard, K. S. (2004) Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 13.

Efron, B. (2004) Selection and estimation for large-scale simultaneous inference. *J. Am. Statist. Ass.*, **96**, 96 – 104.

Efron, B. and Tibshirani, R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epid.*, **23**, 70 – 86.

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.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Boston: Academic Press.

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

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

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

George, E. I. (1986). Minimax multiple shrinkage estimation. *Ann. Statist.* **14**, 188 – 205.

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

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berk. Symp. Math. Statist. Prob. 1*, pp. 361 - 380. Berkeley: Univ. California Press.

Lehmann, E. L. (1986) *Testing Statistical Hypotheses, 2nd edition.* New York: Springer.

Lehmann, E. L and Casella, G. (2002) *Theory of Point Estimation, 2nd Edition.* New York: Springer.

Lindley, D. V. and Smith, A. F. (1972) Bayes estimates for the linear model. *J. R. Statist. Soc. Ser. B*, **34**, 1 – 41.

Müller, P., Parmigiani, G., Robert, C. P. and Rousseau, J. (2004) Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays. *J. Am. Statist. Ass.*, **468**, 990 – 1001.

Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory.* Cambridge, MA: Harvard University Press.

Sarkar, S. K. (2002) Some results on false discovery rates in multiple testing procedures. *Ann. Statist.*, **30**, 239 - 257.

Schervish, M. (1989) A general method for comparing probability assessors *Ann. Statist.*, **17**, 1856 – 1879.

Stein, C. M. (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., 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.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 14.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004b). 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**, Article 15.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624 – 629.

Table 1. Outcomes of n tests of hypotheses

	Accept	Reject	Total
True Null	U	V	m_0
True Alternative	T	S	m_1
	W	Q	n

Table 2. Simulation results for location estimators

		True		Misspecified	
Effect	π_0	Efron	DSE	Efron	DSE
Small	0.1	0.250	0.001	0.258	0.011
	0.5	0.256	0.002	0.257	0.015
	0.8	0.254	0.001	0.254	0.018
Medium	0.1	0.253	0.002	0.274	0.062
	0.5	0.260	0.002	0.260	0.142
	0.8	0.253	0.001	0.256	0.164
Large	0.1	0.304	0.000	0.250	0.200
	0.5	0.274	0.003	0.253	0.195
	0.8	0.260	0.006	0.252	0.197

Table 3. Simulation results for strength of evidence methods

		True		Misspecified	
Effect	π_0	Q-value	DSE	Q-value	DSE
Small	0.2	0.303	0.173	0.402	0.181
	0.5	0.264	0.253	0.350	0.302
	0.8	0.164	0.266	0.280	0.311
Medium	0.2	0.341	0.179	0.398	0.186
	0.5	0.266	0.252	0.346	0.289
	0.8	0.166	0.253	0.279	0.308
Large	0.2	0.302	0.166	0.173	0.164
	0.5	0.266	0.254	0.275	0.296
	0.8	0.166	0.266	0.310	0.312

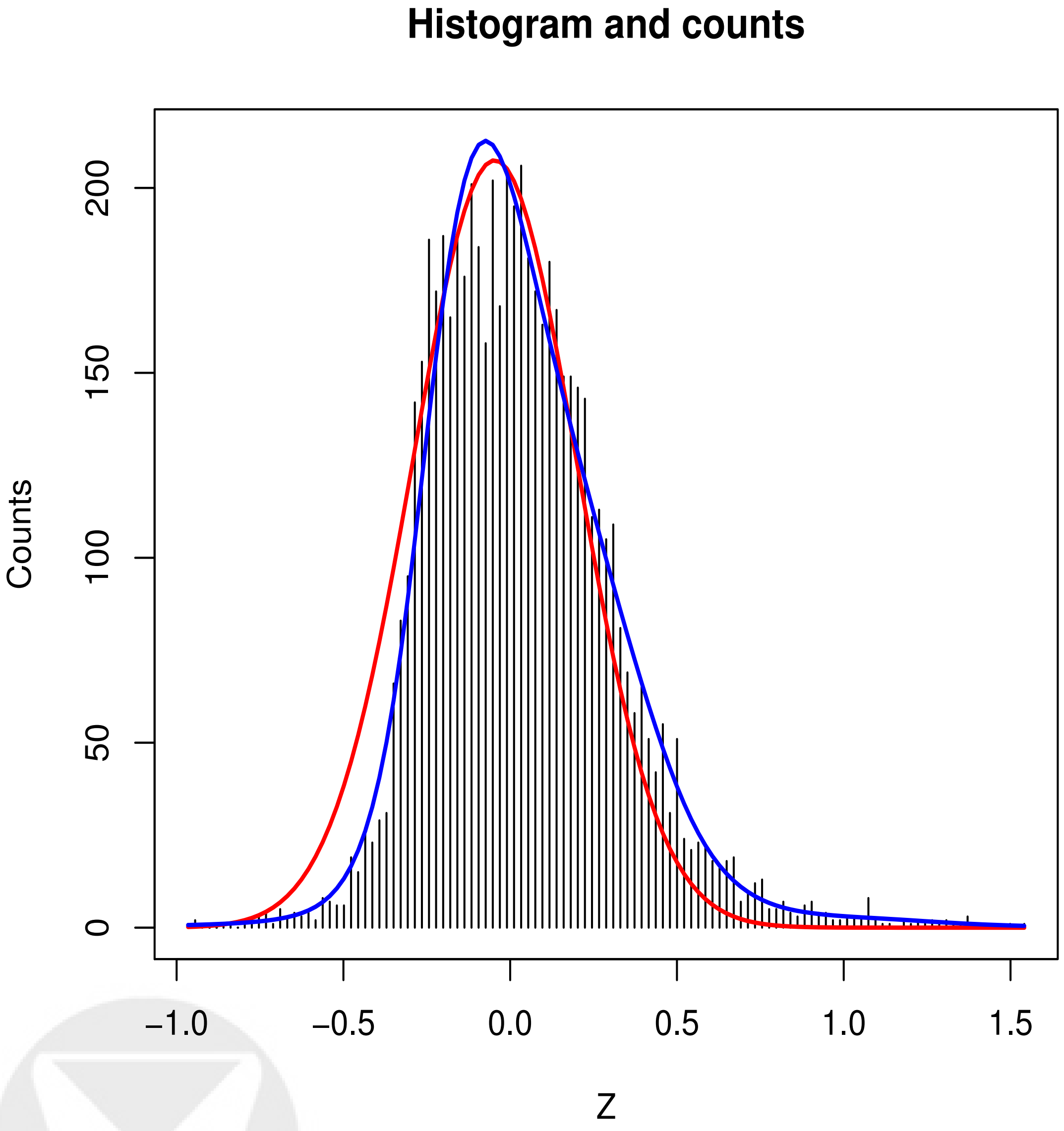

Fig. 1. Plot of empirical (blue line) and null (red line) densities using local false discovery rate estimation method of Efron (2004).

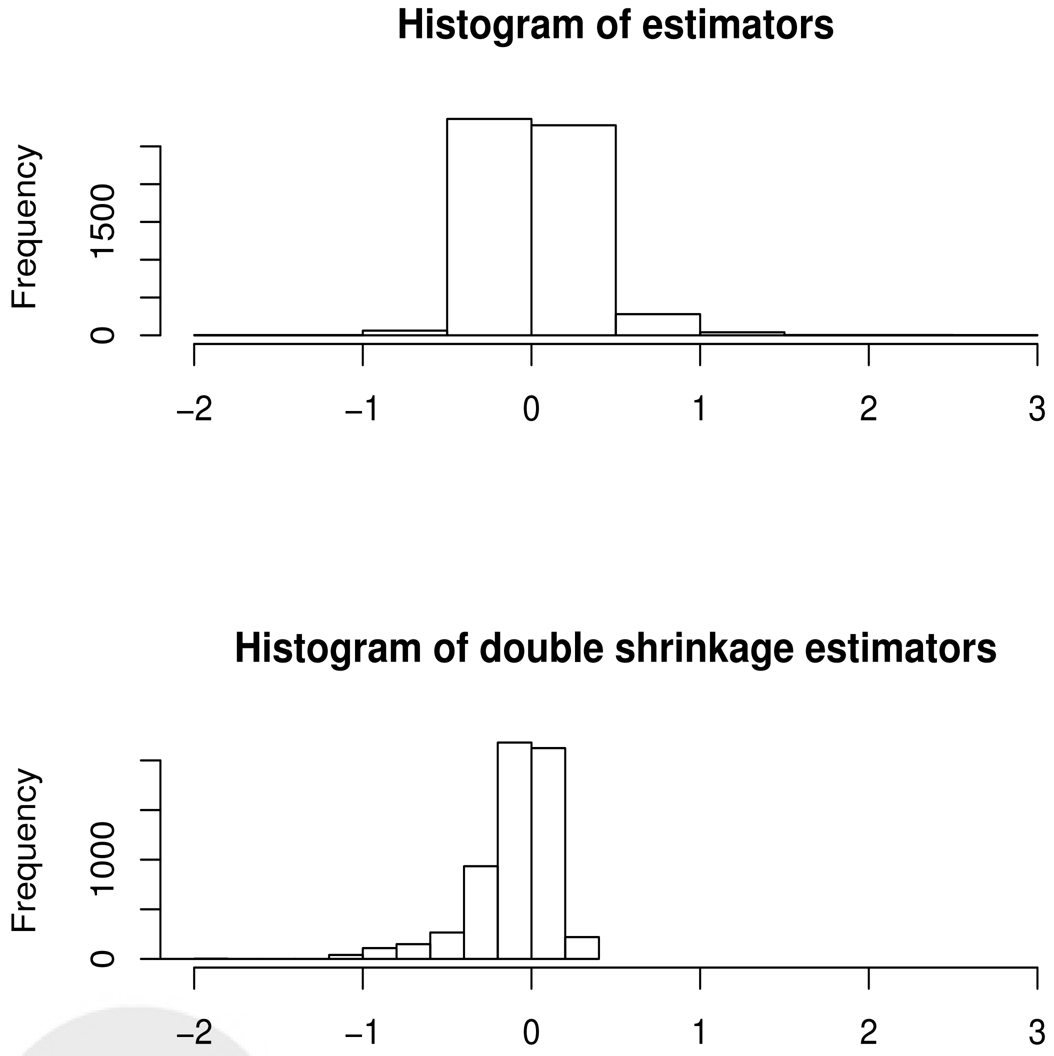

Fig. 2. Histogram of original estimators (numerator of t-statistics) in upper plot and double shrinkage estimators in lower plot.

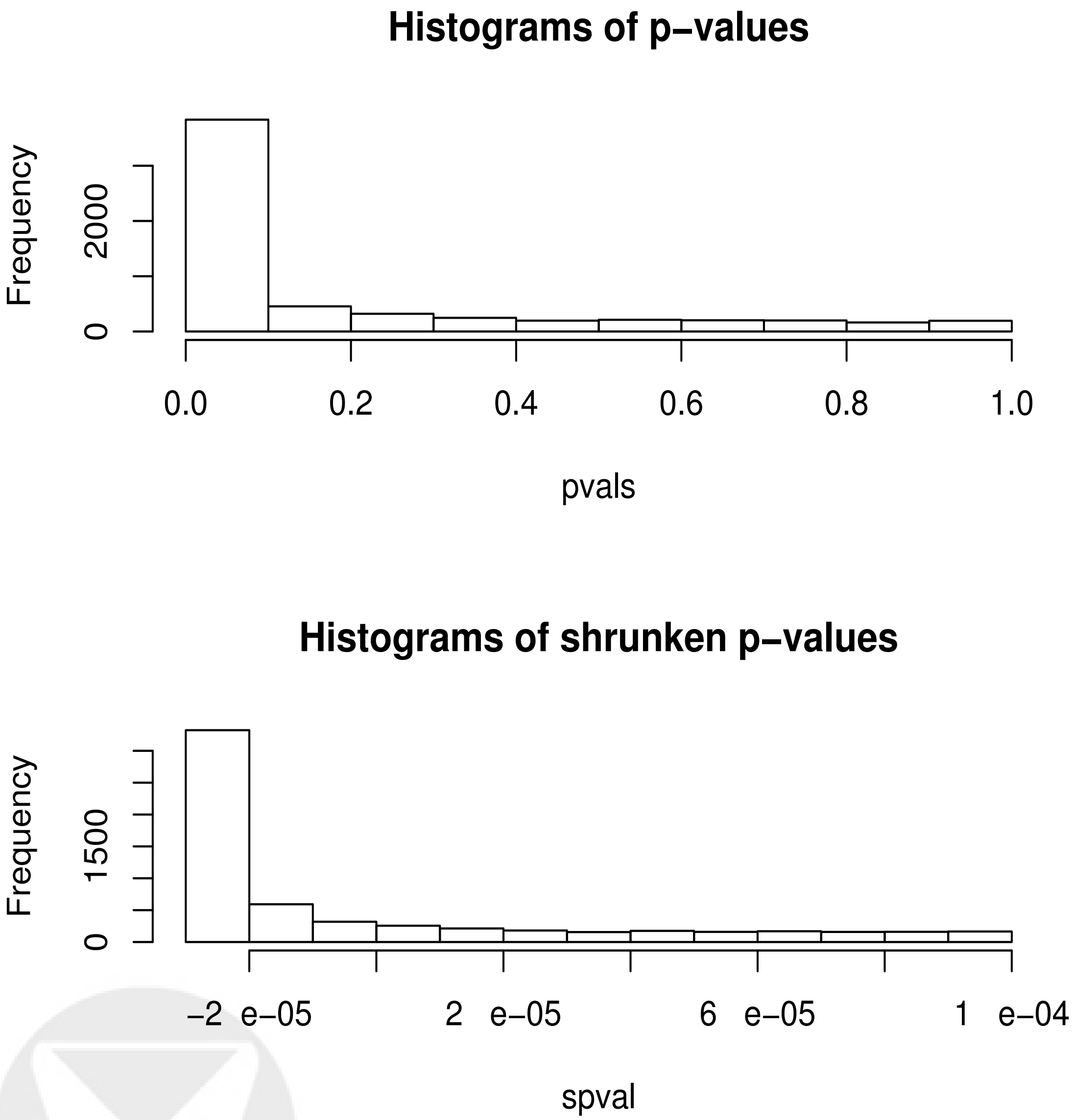

Fig. 3. Histogram of original p-values in upper plot and doubly shrunken p-values in lower plot.