Statistica Sinica **19** (2009), 125-143

EMPIRICAL BAYES METHODS FOR ESTIMATION AND CONFIDENCE INTERVALS IN HIGH-DIMENSIONAL PROBLEMS

Debashis Ghosh

Penn State University

Abstract: There is much recent interest in statistical methods regarding the false discovery rate (FDR). The literature on this topic has two themes. In the first, authors propose sequential testing procedures that control the false discovery rate. In the second, authors study the procedures involving FDR in a univariate mixture model setting. While this work is useful for the selection of hypotheses, there is interest in estimation as well. We take an Empirical Bayes approach and propose estimators and associated confidence intervals in the multiple testing setting. Our framework is general; the proposed methodology is applied to data from a genome scan in Alzheimer's disease.

Key words and phrases: Estimation target, hypothesis testing, James-Stein estimation, multiple comparisons, simultaneous inference.

1. Introduction

Because of technological developments in scientific fields (e.g., neuroimaging and high-throughput genomics), experiments are now performed in which thousands of hypotheses are tested simultaneously. In problems dealing with multiple testing, the usual quantity that has been controlled 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).

Many authors have argued that control of the FWER is too stringent. An alternative to it 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. The first is on procedures that control the FDR. A very simple procedure based on ordering the p-values of test statistics was proposed by Benjamini and Hochberg (1995); it was later shown in Benjamini and Yekutieli (2001) that the original Benjamini-Hochberg procedure controls FDR under a certain dependence structure. 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 sequential procedures and demonstrating that they control the false discovery rate.

The second class of false discovery rate procedures is based on direct estimation of the false discovery rate. This is the approach adopted by Efron, Tibshirani, Storey and Tusher (2001), Storey (2002) and Genovese and Wasserman (2002). These two classes of methods have been unified by Storey, Taylor and Siegmund (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 they provide the data analyst with a post-data assessment of the strength of evidence available in the dataset. To be concrete, based on the number of hypotheses 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 might determine which hypotheses should be followed up in further studies.

Most of the literature described above has focused on the issue of selection, that is, determining which null hypotheses are false. While this is useful for many high-dimensional problems, it is also clear that in many cases a null hypothesis might not be clearly defined. Thus, interest might focus instead on constructing estimators and associated confidence intervals. This topic has been explored less. One quantity from the multiple testing literature that has an estimation quality is the local false discovery rate (Efron et al. (2001)), but its main use so far has been for selection of hypotheses. van der Laan, Dudoit and Pollard (2004a,b) propose various error-controlling procedures and methods for simultaneous confidence intervals for the multiple testing problem. A recent proposal by Benjamini and Yekutieli (2005) developed a method of confidence interval construction that is analogous to the idea of the false discovery rate. However, its use was criticized by several authors in the discussion, among them Westfall (2005). He noted the difficulty in interpreting the false coverage rate criterion of Benjamini and Yekutieli. In addition, he pointed out the issue of bias of the estimates because of the effect of selection. He suggests that shrinkage-based estimators and associated confidence intervals might be more appropriate for this setting. The importance of estimation in the multiple testing situation has also been advocated by Prentice and Qi (2006): "specialized statistical techniques will be required for estimating the magnitude of odds ratios or other parameters that characterize the strength of such associations, in view of the selection process."

In this paper, we consider the use of Empirical Bayes methods for construction of estimators and confidence intervals. The structure of the paper is

Table 1. Outcomes of m tests of hypotheses.

	Accept	Reject	Total
True Null	U	V	n_0
True Alternative	T	S	n_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

Note: All table entries are mean-squared error estimates.

as follows. Multiple testing concepts and false discovery rate procedures are reviewed in Section 2. A mixture model is introduced there; we use it, and decision-theoretic ideas, to motivate shrinkage estimation procedures in Section 3. In Section 4, we propose estimation procedures for double shrinkage estimators and their associated confidence intervals. Some discussion of optimality for the population version of these estimators is given there as well. We illustrate this methodology using real and simulated data in Section 5. We conclude with some discussion in Section 6.

2. Background and Preliminaries

Suppose we have test statistics $T_1, \ldots, T_n$ for testing hypotheses H_{0i}, $i = 1, \ldots, n$. 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 n_0. Suppose we 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 is conceptualized in Table 1.

Based on Table 1, the FWER is defined as $P(V \geq 1)$. Further discussion

for FWER-controlling procedures can be found in Ge, Dudoit and Speed (2003), Dudoit, van der Laan and Pollard (2004) and van der Laan, Dudoit and Pollard (2004a,b).

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

$$FDR \equiv E\Big[\frac{V}{Q}\Big|Q > 0\Big]P(Q > 0).$$

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) and Sarkar (2002)).

2.2. Mixture model motivation of FDR

An alternative approach 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

$$f(t) \equiv \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \tag{2.1}$$

The mixture model framework represented in (2.1) 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) and Cox and Wong (2004)).

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. Our aim is to use the mixture model (2.1) to motivate new estimation procedures that take into account the multiplicity of parameters being tested.

3. Decision Theory and Mixture Models

Decision theory is an area with a long history in statistics (Raiffa and Schlaifer (1961) and 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 $R(\theta, d) = E\{L(\theta, d)\}$, where the expectation is taken with respect to the distribution of the data. We use the terms parameter and target interchangeably here, and in the sequel.

Consider the following two-stage model:

$$\begin{aligned} T_i|\mu_i &\overset{ind}{\sim} N(\mu_i, 1) \\ \mu_1, \ldots, \mu_n &\overset{i.i.d.}{\sim} F, \end{aligned} \tag{3.1}$$

where μ_i is the mean of T_i, and F is some distribution function. Model (3.1) specifies a two-stage model for the joint distribution of $(T_1, \ldots, T_n)$. Note that we view T_i as an estimator of μ_i, $i = 1, \ldots, n$.

Now take F in (3.1) 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 (3.1), this implies that

$$T_i \overset{i.i.d.}{\sim} \pi_0 N(\mu_0, 1) + (1 - \pi_0)N(\mu_1, 1). \tag{3.2}$$

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), but those authors were only concerned with selection.

In the multiple testing literature, (3.2) would arise in a situation where we wished to test n hypotheses of the form $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, 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). Note that we take T_i to be the absolute value of the test statistic. Otherwise, we would have to consider a three-component mixture model instead of (3.2) for determining differential expression.

We can generalize (3.2) to allow for unequal variances:

$$T_i \overset{i.i.d.}{\sim} \pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0)N(\mu_1, \sigma_1^2). \tag{3.3}$$

To extend the decision theoretic viewpoint here, what the mixture distribution for F, as manifested in (3.2) and (3.3), does is to provide two targets (parameters)

for shrinkage estimators: μ_0 and μ_1. Such a model has been studied by George (1986) in a fully parametric setting.

4. Proposed Theory and Methods

4.1. Double shrinkage estimation: theory

An alternative to the FDR procedures that would address the multiple testing issue in (3.3) is to construct shrinkage estimators that shrink toward the two distributions. We now demonstrate how to do this using (3.3); note that we are considering T_i $(i = 1, \ldots, n)$ to be estimators of the location parameter μ. Assume initially that μ_0 and μ_1 are known and that the variances are known. No gains in borrowing strength across estimators are possible in (3.2) because of the equal variances for the two component distribution of the mixture model.

Take $a \wedge b$ to be the minimum of a and b. To construct a shrinkage "estimator" of μ in model (3.3), we calculate a James-Stein estimator (James and Stein (1961)) relative to each of the component of the mixture distribution and then mix the estimators with appropriate weights. With respect to the first component, the James-Stein estimator is given by

$$T_{0i}^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^{n}(T_i - \mu_0)^2}\right](T_i - \mu_0), \tag{4.1}$$

while for the second component, it is given by

$$T_{1i}^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^{n}(T_i - \mu_1)^2}\right](T_i - \mu_1), \tag{4.2}$$

$i = 1, \ldots, n$. A shrinkage "estimator" combining (4.1) and (4.2) 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

$$\pi_k(T_i) = \frac{\pi_k f_k(T_i)}{\pi_0 f_0(T_i) + \pi_1 f_1(T_i)}. \tag{4.3}$$

and f_0 and f_1 refer to the marginal densities of the distribution of the test statistics under the null and alternative hypotheses. We refer to this as a double shrinkage estimator, since T_i is shrunk toward both μ_0 and μ_1 by construction. We have used quotes for estimator since it is not calculable without further estimation; this is deferred to Section 4.2. Consider (4.3) further with $k = 0$ so (4.3) equals the local false discovery rate (Efron et al. (2001)). There is thus an intimate connection between the double shrinkage estimators and 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 distribution of the test statistics under the null and alternative hypotheses.

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

$$\tilde{T}_i = \frac{\hat{T}_i^{num}}{\hat{T}_i^{den} + a_0}, \quad i = 1, \ldots, n, \tag{4.4}$$

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 $T_1^{den}, \ldots, T_n^{den}$. Typically a_0 is chosen to minimize the coefficient of variation (Tusher, Tibshirani and Chu (2001)). 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 $\tilde{T}_1, \ldots, \tilde{T}_n$ achieves shrinkage for the multiple testing situation; however, the "fudge factor" in (4.4) 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: there is adaptive shrinkage based on data-dependent weights in what we propose.

4.2. Double shrinkage estimation: Empirical Bayes estimators and confidence intervals

We now construct the double shrinkage estimators under (3.1), where $F = \pi_0 F_0 + (1 - \pi_0)F_1$. To do this, we utilize a density estimation method proposed by Efron (2004). Note from (2.1) 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 $T_1, \ldots, T_n$. For estimation of $\pi_0 f_0(t)$, the theoretical null 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. We 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.

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

$$\hat{T}_i^{JS} = \hat{\pi}_0(T_i)\hat{T}_{0i}^{JS} + \{1 - \hat{\pi}_0(T_i)\}\hat{T}_{0i}^{JS}, \tag{4.5}$$

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 a method for practical implementation of double shrinkage estimators for test statistics. The parameter

estimators are obtained by multiplying by the standard error or other appropriate backtransformation. Note that the adjustment for multiple comparisons occurs here in the construction of the multiplier for the statistic minus the estimate mean. In the p-value setting, Ghosh (2006) shows that such shrinkage appropriately controls the false discovery rate.

Along with estimators of the parameters that adjust for multiple testing, it is useful to have confidence intervals that account for the multiple testing phenomenon. While intervals have been proposed for the simple normal/normal model (Morris (1983), Laird and Louis (1987) and Carlin and Gelfand (1990)), the mixing distribution in (3.1) is more general. In addition, we only have one statistic per model in the first stage of (3.1). We focus on the situation where the test statistic for each hypothesis is estimating a parameter of interest, such as a difference in means or an odds ratio. Our confidence intervals will be for the same type of situations as those of Benjamini and Yekutieli (2005). The main advantages of our procedure are that the confidence intervals have the usual interpretation, they adjust for multiple testing, and they account for selection.

To calculate the confidence intervals, we use the fact that marginally, $T_1, \ldots, T_n$ are i.i.d. with a normal distribution with mean $\pi_0 \int u f_0(u)dt + (1-\pi_0) \int u f_1(u)$ and variance $1 + \sigma_\mu^2$, where $\sigma_{mu}^2 = \pi_0^2 Var_0(\mu) + (1-\pi_0)^2 Var_1(\mu)$. This leads to the following algorithm.

1. Estimate π_0, f_0 and f_1 using the algorithm above.
2. Construct the double shrinkage estimators of $\mu_1, \ldots, \mu_n$, $\hat{T}_1^{JS}, \ldots, \hat{T}_n^{JS}$.
3. Sample with replacement n observations $\mu_1^*, \ldots, \mu_n^*$ from the estimated density $\hat{\pi}_0 \hat{f}_0(t) + (1-\hat{\pi}_0)\hat{f}_1(t)$ and generate data $T_{1,1}^*, \ldots, T_{n,1}^*$, where $T_{i,1}^* \sim N(\hat{T}_i^{JS}, 1 + \hat{\sigma}_*^2)$, and $\hat{\sigma}_*^2$ is the empirical variance of the μ^*'s.
4. Repeat Step 3 B times. Use the empirical distribution of $T_{i,1}^*, \ldots, T_{i,B}^*$ to calculate confidence intervals for μ_i, $i = 1, \ldots, n$.

Based on the empirical distributions derived in Step 3, we can construct confidence intervals for each of the components of $\mu \equiv (\mu_1, \ldots, \mu_n)$. For the last step of the algorithm, we can either get an estimate of the standard error using the empirical distribution, or take the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the distribution to get confidence limits. This can be viewed as an approximate posterior predictive distribution for μ, so arguments such as those in Morris (1983) would suggest that the proposed intervals would have reasonable coverage properties. We explore the finite-sample properties of the methodology in Section 5.1.

The confidence intervals that are proposed here adjust for the multiple testing by using shrinkage principles advocated by Westfall (2005) rather than those advocated by Benjamini and Yekutieli (2005). However, the estimators and intervals found here should be conservative because the shrinkage does not take into account the selection process.

We have assumed that the distribution of the test statistic is known exactly under the null hypothesis when we constructed shrinkage estimators and associated confidence intervals in the previous section. As Efron (2004) has recently written, this is making the assumption of a theoretical null distribution, which might be incorrect. He instead argues for the use of an empirical null distribution.

The idea behind the empirical null distribution is to allow for the fact that, under the null hypothesis, the test statistic might not have a known distribution. However, if we utilize the zero-matching assumption of Efron (2004), then it is possible to estimate the mean and variance for the empirical null distribution. We can then still calculate shrinkage estimators and associated confidence intervals using the method of the previous section. What is different is that the mean and variance of the test statistic under the null hypothesis might not be zero and one, respectively. The mean and variance are estimated using the method of Efron (2004), and this can be done using either a symmetric empirical null distribution or an asymmetric empirical null distribution.

4.3. Double shrinkage estimation: Optimality

One optimality property that can be used to evaluate estimators is minimaxity. As described in (Lehmann and Casella (2002, Sec. 5.1, p.309)), an estimator δ^M of a function of a parameter μ, $g(\mu)$, that 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 background. Let m be a function from R^n to R, and take 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 $\nabla^2 m(t) = \sum_{i=1}^n \nabla_i^2 m(t) \leq 0$, where $\nabla^2 m$ is defined as $\nabla(\nabla m)$.

We consider estimators of the form:

$$\hat{T}_{ki} = T_i + \nabla \log m_k(T_i), \quad i = 1, \ldots, n; \; k = 0, 1, \tag{4.6}$$

where m_0 and m_1 are functions that are twice differentiable. As shown in Brown (1971), estimators of the form (4.6) generate a wide class of rules. A generalized double shrinkage estimator is given by

$$\hat{T}_i = \sum_{k=0}^{1} c_k(T_i) \hat{T}_{ki} \tag{4.7}$$

where

$$c_k(T_i) = \frac{\pi_k m_k(T_i)}{\pi_0 m_0(T_i) + \pi_1 m_1(T_i)}, \tag{4.8}$$

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* (4.6). *If* m_k *and* ∇m_k *are differentiable,* m_k *is superharmonic and satisfies the conditions*

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

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

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

Theorem 1 provides sufficiency conditions 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 similar to the regularity conditions that densities must satisfy for the usual asymptotic results for maximum likelihood estimation procedures (i.e., consistency, asymptotic normality and efficiency). Recall, however, that the mixture model for the 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* (4.7) *is minimax.*

Proof. Note that if m_0 and m_1 are superharmonic, then $\sum_{k=0}^{1} \pi_k m_k$ is also superharmonic. By Theorem 1, (4.7) is minimax.

Based on the results of Theorem 1 and Lemma 1, the population version of the proposed double shrinkage estimators from Sections 4.1 and 4.2 are minimax. This provides some robustness to the estimation procedure, as well as some theoretical justification for their construction.

5. Numerical Examples

5.1. Simulation studies

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 was normal with mean η_i and variance $2i + 1$, $i = 0, 1$. The number of hypotheses tested was

$n = 10,000$. 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:** μ was 0.25 with probability 0.75 and 0.5 with probability 0.25.
- **Medium:** μ was 0.25 with probability 0.5 and 0.5 with probability 0.5.
- **Large:** μ was 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 weights, i.e., $\pi_0(t) = \pi_0$ instead of (4.3). Thus, we are not incorporating the data-adaptive nature of the weights at this stage; this is considered further in Section 5. With regard to the target in the double shrinkage estimators, we considered two situations: where the true target is used, and where the target is misspecified. The misspecified target is taken to be one. The mean-squared error results are shown in Table 2. Based on the true target results, there is a major increase in risk in 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 data-dependent shrinkage toward the two targets has better risk properties relative to shrinkage toward one in this multiple testing framework.

Next, we explored the properties of the proposed double shrinkage estimators and confidence intervals through a small simulation study. Other goals of the study were to assess the sensitivity of the procedure to normality and dependence. Our comparison with estimators was with unadjusted estimators, and we compared our proposed confidence interval procedure with that of Benjamini and Yekutieli (2005). Their procedure works as follows.

1. Sort the univariate p-values, $p_{(1)} \leq \cdots \leq p_{(n)}$ in increasing order.
2. Calculate $R = \max\{i : p_{(i)} \leq iq/n\}$.
3. Select the R parameters for which $p_i \leq Rq/n$, corresponding to the rejected hypotheses.
4. Construct a $(1 - Rq/n)$ CI for each parameter selected.

Note that step (2) is the Benjamini and Hochberg (1995) procedure based on p-values. By contrast, our procedure provides simultaneous confidence intervals for all parameters. To compare the proposed confidence interval methodology to that of Benjamini and Yekutieli, we consider confidence intervals of selected hypotheses. In particular, we use α as an error control parameter, find the hypotheses that are rejected given a certain level of α, and then report the confidence interval coverages for the selected parameters using both methodologies. Notice that

Table 3. Simulation results for proposed methods.

Independent	Gaussian	α	MSE(U)	MSE(S)	Cov(EB)	COV(BY)
Yes	Yes	0.1	-0.03	0.003	0.97	0.96
		0.05	-0.031	0.002	0.98	0.97
		0.01	-0.029	0.001	0.98	0.96
		0.005	-0.028	0.003	0.97	0.95
Yes	No	0.1	-0.031	0.002	0.97	0.96
		0.05	-0.032	0.001	0.98	0.97
		0.01	-0.033	0.002	0.98	0.95
		0.005	-0.034	0.001	0.97	0.95
No	Yes	0.1	-0.036	0.002	0.97	0.96
		0.05	-0.032	0.002	0.98	0.97
		0.01	-0.033	0.001	0.98	0.96
		0.005	-0.031	0.001	0.97	0.96
No	No	0.1	-0.031	0.002	0.97	0.97
		0.05	-0.032	0.003	0.98	0.96
		0.01	-0.033	0.002	0.98	0.97
		0.005	-0.031	0.003	0.97	0.96

by definition, the coverage probability for a confidence interval in the Benjamini-Yekutieli framework does not exist, while our procedure calculates confidence intervals for all parameters.

In terms of the simulation model, we fit a mixture model, much as in the previous simulation study. We again used $n = 10,000$ and two groups with 10 samples each. 1,000 simulation samples were generated for each scenario; 500 resamplings were used for the bootstrap method. In what we report here, we take $\pi_0 = 0.8$. For differentially expressed genes, the diference in means was assumed to be 0.5. The variance for all genes was 1 and 2 for the two groups. We considered four scenarios for the differentially expressed genes based on independence between differentially expressed genes (independent or dependent) and their distribution (Gaussian or t). For dependence, we assumed that the genes were equicorrelated with correlation 0.3. For the t-distribution, we took 3 degrees of freedom. The results are given in Table 3. We find that use of the shrunken estimator outperforms that of the unshrunken estimator. This result seems to be insensitive to dependence and normality for the situations in the simulation study. In addition, the coverage for the confidence intervals from the proposed methodology tends to be fairly conservative across all scenarios considered. The Benjamini-Yekutieli procedure tends to have good coverage properties as well. Again, note that the coverage is for selected parameters. Given the finding that the q-value enjoys a shrinkage property (Ghosh (2006)), the same might hold for the false coverage rate criterion of Benjamini and Yekutieli (2005).

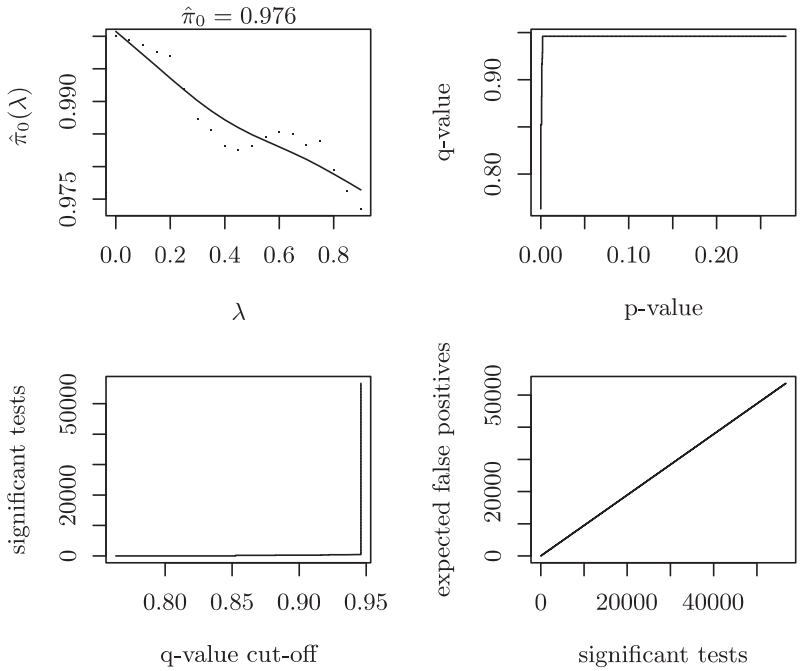

Figure 1. Plot of q-values for Alzheimer's disease genome scan data from Maraganore et al. (2005), using the method of Storey (2002).

5.2. Parkinson's disease genome scan

We apply the proposed methodology to genomic data from a study of Parkinson's Disease conducted by Maraganore, de Andrade, Lesnick, Strain, Farrer, Rocca, Pant, Frazer, Cox and Ballinger (2005). They performed a two-tiered, whole-genome association study of Parkinson disease (PD). For Tier 1, 198,345 uniformly spaced and informative single-nucleotide polymorphisms (SNPs) were genotyped in 443 sibling pairs discordant for PD. The second tier of the study involved collecting information on 1,793 PD-associated SNPs and 300 genomic control SNPs in 332 matched case-unrelated control pairs. For the purposes of illustrating the proposed methodology, we focus on the Tier 1 data. In addition, we excluded noninformative SNPs (SNPs that showed no variation across the samples). This resulted in a total of $m = 197,222$ SNPs for analysis. Note that Maraganore et al. (2005) considered a slightly different subset of SNPs using external biological knowledge.

We first applied the q-value method of Storey (2002) to identify candidate alleles that might be associated with disease. The results from the q-value analysis are given in Figure 1. Key features to note are the following. First, the estimated proportion of alleles not associated with disease is 97.6%, although

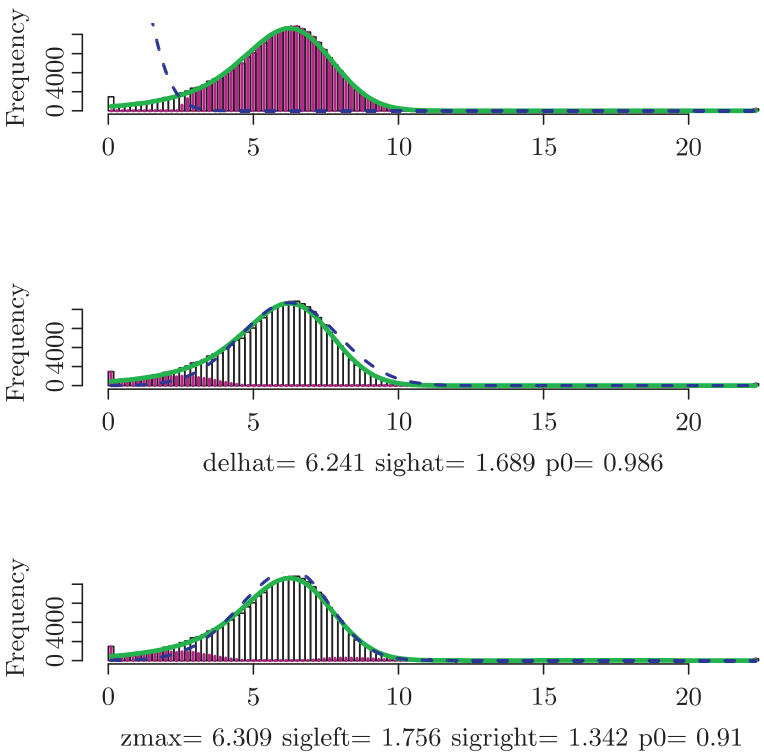

Figure 2. Plot of estimated false discovery rates using the method of Efron (2004). The upper plot assumes a theoretical null distribution of $N(0,1)$, the middle plot assumes an empirical null distribution symmetric about the point 6.241; the bottom plot assumes an asymmetric empirical null distribution. Further details on the symmetric versus asymmetric null distribution can be found in Efron (2004).

one also observes nonmonotonic behavior in the q-values themselves. Second, all the genes are nonsignificant for most levels of desired false discovery rate control. One would need to use a level close to 0.85 to find significant genes. Next, the local false discovery rate method of Efron (2004) was used to estimate the null and alternative distributions. There are three choices of a null distribution to use here: the theoretical null distribution, the symmetric empirical null distribution, and the asymmetric empirical null distribution. Plots from the three distributions are given in Figure 2. The algorithm failed to converge if we wished to fit the theoretical null distribution here, while it did converge for the empirical null distributions. We find that the estimate of π_0 decreases or increases, relative to the q-value analysis, based on the choice of empirical null distribution used. The implication for the double shrinkage estimators is that there will be *a priori* higher shrinkage toward the null distribution component for these data.

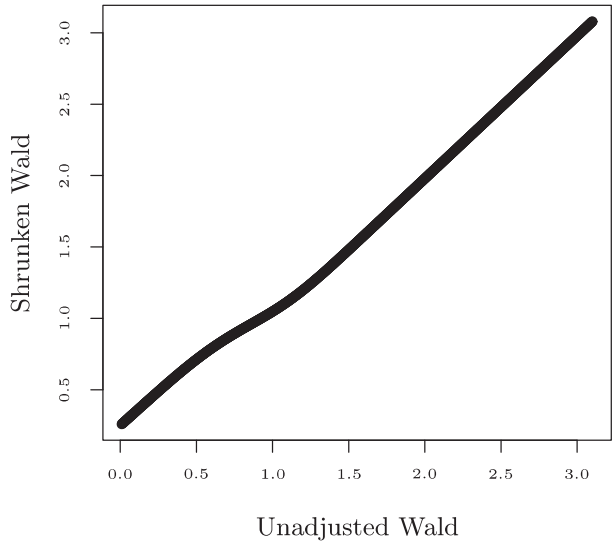

Figure 3. Plot of Wald statistic based on odds ratio (horizontal axis) versus double shrinkage estimator (vertical axis).

We next calculated double shrinkage estimators for the odds ratio using the proposed methodology in Section 5, using the symmetric empirical null hypothesis. A plot comparing the Wald statistic based on the odds ratio to the double shrinkage test statistic is given in Figure 3. Based on the plot, we find evidence of some shrinkage; the slope of a linear regression to the line is about 0.9. Based on the multiple testing procedures discussed previously, we perform the following multiple testing/selection/estimation procedure.

1. Reject all hypotheses with q-value less than 0.8.
2. Report adjusted odds ratios and confidence intervals for those SNPs selected in Step 1.

The results are listed in Table 4. Note that because the selection process is not modelled in the above algorithm, the estimators and odds ratios have some bias associated with them. In addition, the confidence intervals will be on the conservative side. Note that there appears to be a reversal in the direction of associations of several SNPs after applying the shrinkage estimation procedure. This could be due either to the overshrinkage phenomenon of the James-Stein estimator or to the fact that the empirical null distribution is being used here. In theory, one could check if the latter is the cause simply by redoing the analysis based on the theoretical null distribution. As mentioned above, the analysis using the theoretical null distribution failed to converge. This issue is currently being explored further.

Table 4. Results for SNPs - shrunken odds ratios and associated 95% CIs.

SNP ID	OR	sOR	95% CI
24235124	0.45	0.19	(0.07, 0.55)
46538934	0.46	0.19	(0.07, 0.52)
23204984	0.49	0.19	(0.07, 0.51)
23869311	0.48	0.2	(0.08, 0.52)
23166462	2.1	0.2	(0.08, 0.53)
23864204	1.8	0.19	(0.09, 0.42)
24620648	0.56	0.29	(0.13, 0.63)
24395156	0.55	0.17	(0.08, 0.4)
46542193	0.17	0.32	(0.02, 5.32)
23361109	0.17	0.32	(0.02, 5.33)
23265602	0.32	0.79	(0.15, 4)
23657057	1.96	0.37	(0.15, 0.9)
23457423	0.42	0.2	(0.06, 0.62)
23772641	0.47	0.45	(0.16, 1.27)
24422280	0.53	0.21	(0.09, 0.48)
24105865	1.96	0.39	(0.16, 0.97)
24650291	2.04	0.31	(0.13, 0.76)
46551827	0.56	0.23	(0.11, 0.51)
24443744	0.54	0.25	(0.11, 0.54)
24548027	0.56	0.19	(0.09, 0.42)
24690480	1.8	0.19	(0.09, 0.42)
24690643	1.85	0.22	(0.11, 0.48)
23824090	15.75	0.5	(0.11, 3.97)
24561474	0.55	0.21	(0.09, 0.47)
24098016	0.5	0.41	(0.16, 1.06)
23958252	1.71	0.19	(0.09, 0.41)
24650302	1.83	0.29	(0.12, 0.66)

Note: SNP ID refers to SS ID from Table 2 of the supplementary information by Maraganore et al. (2005); OR is the unadjusted odds ratio; sOR is the adjusted odds ratio using shrunken estimator from the methods proposed in Section 5; 95% CI is the confidence interval using standard error from resampling distribution based on 1000 resamplings discussed in Section 5.

While the q-value analysis suggests that we have a high level of false discoveries, the analysis is properly calibrated so that investigators have a ranked set of SNPs to follow up on for further validation or confirmatory studies. In Maraganore et al. (2005), they had another collection of SNPs that were considered from those found to have significant unadjusted association in Tier 1, as well as from known biological pathways in the literature; this is their Tier 2 study. In analysis of the combined Tier 1 and Tier 2b data, they found that the two SNPs with the lowest unadjusted P values ($P = 9.07 \times 10^{-6}$; $P = 2.96 \times 10^{-5}$) tagged the PARK10 late-onset PD susceptibility locus.

While the two-stage design is important, attempting to model that, or to account for it in the analysis, is beyond the scope of the paper but an important topic for future research.

6. Discussion

In this article, we have provided an Empirical Bayes-oriented approach to testing and estimation in the multiple testing issue. While this type of methodology has been extensively explored for the problem of testing in this area, the issue of estimation and confidence intervals using such methods is relatively in its infancy. The paper by Benjamini and Yekutieli (2005) focused on estimation of the confidence intervals, while Efron (2004) discusses estimation of the density for the alternative hypothesis.

Central to the proposed development in the paper is 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 proposed double shrinkage methods proposed. It is shown that shrinkage toward the two targets that comprise the mixture distribution potentially leads to better risk behavior than existing procedures. While shrinkage toward multiple targets was studied from a risk point of view by George (1986), we extend that view 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 estimation provides a natural method for adjusting for the multiple testing problem. In particular, we find that there is a reduction in strength of evidence after one accounts for the multiplicity of hypotheses being tested. 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 and weight functions for the shrinkage estimators. This gives a natural intuition as to why shrinkage of estimators work for the multiple testing problem considered here.

There are several possible extensions of the proposed approach that we shall be pursuing in later work. First, we want to develop procedures in which we account for the selection process, as referred to by Prentice and Qi (2006).

Second, we have assumed the nuisance parameters to have known values. However, one could give distributions to these parameters and perform shrinkage estimation with these quantities as well. Third, we have assumed in our development that the test statistics are a random sample. In reality, one could assume dependence due to linkage disequilibrium or coregulation and coexpression of genes in genetic pathways. While the theoretical development of properties of Empirical Bayes methods in this setting might be difficult, it may be possible to find double shrinkage estimators. Finally, it might be useful to combine the selection and estimation procedures in the following two-step approach.

1. Select genes whose shrunken p-value (Ghosh (2006)) is below some threshold.
2. For those genes selected in Step 1, report the associated parameter estimates and confidence intervals using the procedures in this paper.

These are areas of current investigation.

Acknowledgements

The author would like to thank Tom Nichols and Trivellore Raghunathan for helpful discussions, and Radu Craiu for bringing the data of Maraganore et al. (2005) to his attention. This research is supported by the National Institutes of Health.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. 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. Plann. Inference* **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.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate controlling confidence intervals for selected parameters (with discussion). *J. Amer. Statist. Assoc.* **100**, 71-80.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Annals of Mathematical Statistics* **42**, 855-903.

Carlin, B. P. and Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. *J. Amer. Statist. Assoc.* **85**, 105-114.

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

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. Amer. Statist. Assoc.* **96**, 96-104.

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

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

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. Roy. Statist. Soc. Ser. 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.

Ghosh, D. (2006). Shrunken p-values for assessing differential expression, with applications to genomic data analysis. *Biometrics* **62** , 1099-1106.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the 4th Berkeley Symposium in Mathematical Statistics and Probability* 1, 361-380, Univ. California Press, Berkeley.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82**, 739-750.

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

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

Maraganore, D. M., de Andrade, M., Lesnick, T. G., Strain, K. J., Farrer, M. J., Rocca, W. A., Pant, P. V., Frazer, K. A., Cox, D. R. and Ballinger, D. G. (2005). High-resolution whole-genome association study of Parkinson disease. *Americal Journal of Human Genetics* **77**, 685-693.

Morris, C. N. (1983). Parametric Empirical Bayes Confidence Intervals. In *Scientific Inference, Data Analysis and Robustness* (Edited by G. E. P. Box, T. Leonard and C.-F. Wu), 25-50. Academic Press, New York.

Prentice, R. L. and Qi, L. (2006). Aspects of the Design and Analysis of High-Dimensional SNP Studies for Disease Risk Estimation. *Biostatistics* **7**, 339-354.

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

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

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

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. Roy. Statist. Soc. Ser. B* **66**, 187-205.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**, 5116-5121.

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.

Westfall, P. H. (2005). Comment on "False discovery rate controlling confidence intervals for selected parameters" by Benjamini and Yekutieli. *J. Amer. Statist. Assoc.* **100**, 85-89.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley, New York.

Department of Statistics and Public Health Sciences, Penn State University, 514A Wartik Lab, University Park, PA, 16802, U.S.A.

E-mail: ghoshd@psu.edu

(Received July 2006; accepted August 2007)