2012

James-Stein estimation and the Benjamini-Hochberg procedure

Debashis Ghosh, *Penn State University*

James-Stein estimation and the Benjamini-Hochberg procedure

Debashis Ghosh

Department of Statistics, Penn State University,
514A Wartik Lab, University Park, PA 16802
E-mail: ghoshd@psu.edu
Fax: (814) 863-6699

Abstract

For the problem of multiple testing, the Benjamini-Hochberg (B-H) procedure has become a very popular method in applications. Based on a spacings theory representation of the B-H procedure, we are able to motivate the use of shrinkage estimators for modifying the B-H procedure. Several generalizations in the paper are discussed, and the methodology is applied to real and simulated datasets.

Keywords: clustering; genomics; high-dimensional data; multiple comparisons; minimaxity.

1 Introduction

Multiple testing considerations are becoming much more prevalent, especially in disciplines such as neuroimaging and genomics. A fairly common problem is to identify "interesting" signals for further study. One approach to this signal detection problem is to use multiple comparisons procedures and to treat rejected null hypotheses as evidence for signal in the dataset. While analysts have typically focused on error control based on familywise error rate (FWER), current focus has shifted to methods for controlling the false discovery rate (FDR). The FWER and FDR are defined in Section 2. This change has been due to the fact that control of FWER tends to be quite stringent (i.e. does not reject enough hypotheses) in many situations. A simple example of a method for controlling FWER is the Bonferroni method. It adjusts individual p-values by multiplying them by the number of tests performed. Since current applications are generating thousands or millions of tests, the Bonferroni adjustment will be quite severe in these settings.

Perhaps the most popular method for FDR control is the Benjamini-Hochberg (B-H) procedure (Benjamini and Hochberg, 1995). The B-H procedure proceeds as follows: given a set of p-values $p_1, \ldots, p_n$, using the sorted p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ we compute

$$\hat{k} = \max\{1 \leq i \leq n : p_{(i)} \leq \alpha i/n\}.$$

Define $\hat{k}$ to be zero if the set is empty. If $\hat{k}$ is nonzero, then reject null hypotheses corresponding to $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$; otherwise, reject nothing.

More recently, Efron et al. (2001) showed the equivalence between the B-H procedure with thresholding procedures based on a mixture model for multiple testing. We describe this equivalence in Section 2. As noted by Ghosh (2006, 2009), the mixture model framework can alternatively be used to motivate shrinkage estimators in the classical James-Stein tradition (James and Stein, 1961). Ghosh (2006) proposed shrinking p-values using the mixture model of Efron et al. (2001), while in later work, Ghosh (2009) developed estimation and confidence interval procedures in the high-dimensional setup based on the mixture model. There has also been related work on using shrinkage estimators for the multiple testing problem on the test statistic scale in which shrinkage has been applied to the numerators and/or variances (Cui et al., 2005; Hwang and Liu, 2010; Hwang et al., 2009; Zhao, 2010). In this article, we will develop James-Stein-type estimators within the Benjamini-Hochberg procedure. To do so will require employing a recent result of Ghosh (2011) reexpressing the B-H procedure in terms of spacings (Pyke, 1965). Doing so will lead to a natural characterization for multiple testing in terms of a sample average to which we can then apply the James-Stein theory. The structure of the paper is as follows. In Section 2, we provide background on multiple testing, spacings theory and James-Stein estimation. We describe the proposed shrinkage methodology in Section 3 and apply them to real and simulated datasets in Section 4. Section 5 concludes with some discussion.

2 Background and Preliminaries

2.1 Multiple testing aspects

Suppose that $p_1, \ldots, p_n$ are used to test a set of n hypotheses, where n_0 of them correspond to true null hypotheses. The decision to reject or fail to reject the individual hypotheses, along with whether or they are truly null, can be formulated in terms of a simple 2×2 table: Note that all the analyst will observe is the number of tests/p-values, n, W and V.

Table 1: Outcomes of m tests of hypotheses

	Accept	Reject	Total
True Null	U	Q	n_0
True Alternative	Z	S	n_1
	W	V	n

Everything else in the table is unknown. Note that Q and Z represent the entries in the table where we commit mistakes. We can next define error rate quantities based on Table 1. For example, the familywise error rate (FWER) that is controlled by the Bonferroni procedure can be defined as $P(Q \geq 1)$. Benjamini and Hochberg (1995) advocate for control of the FDR, which is given by $E(Q|V)P(V > 0)$. Note that in general, the FWER and FDR are different error quantities so that a direct comparison between them is not feasible. In the situation where $n_0 = n$ (i.e., all hypotheses are true nulls), FWER = FDR.

Instead of $p_1, \ldots, p_n$, now suppose we work the corresponding test statistics, $T_1, \ldots, T_n$. An alternative approach to dealing with multiple testing 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

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

The mixture model framework represented in (1) has been used by Efron et al. (2001) and Efron (2004) to study the false discovery rate. There, the two components of the mixture model are assumed to have normal distributions. Based on a so-called "zero-mean" assumption for the test statistics. Such an assumption effectively makes model (1) identifiable and allows for estimation of the components. Given estimates of π_0 and f_0, one then has an estimator of the local false discovery rate of Efron et al. (2001), defined as

$$\text{locfdr}(t) = \frac{\pi_0 f_0(t)}{f(t)}.$$

As Efron (2010, p. 53) shows, if we transform the local false discovery rate by replacing densities with cumulative distribution functions in the definition of locfdr(t), then we can reexpress the B-H procedure as an Empirical Bayes type of rule based on thresholding $\pi_0 F_0/F$.

2.2 James-Stein Estimation

Another procedure that has been applied in a high-dimensional setting are James-Stein estimators. Let us first recall some definitions from decision theory (Ferguson, 1967). Define θ to be the population parameter of interest, the estimator of θ as d. Let$L(\theta, d)$ be a loss function. Its expectation is called the risk function:

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

Note that the expectation above is taken with respect to the distribution of the data.

We consider a hierarchical model:

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

where μ_i is the mean of T_i, and F is some distribution function. In (2), we are viewing T_i as an estimator of μ_i, $i = 1, \ldots, n$. 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.

If we have one observation T_i for μ_i, we can estimate μ_i in an unbiased fashion by T_i, $i = 1, \ldots, n$. James and Stein (1961) derived the remarkable result that one can construct estimators for μ_i with better risk properties by pooling information across the different parameters. A class of generalized James-Stein estimators is given by

$$\hat{T}_i^{JS} = T_i - \left[1 \wedge \frac{n-2}{\sum_{i=1}^n (T_i - \mu_0)^2}\right](T_i - \mu_0),$$

where μ_0 is the mean corresponding to F. In the case where $\mu_0 = 0$, this reduces to the ordinary estimator initially proposed by James and Stein (1961). They showed that under a quadratic loss function, their proposed estimator had lower risk than the ordinary unbiased estimator when $n \geq 3$. This result led to a proliferation of literature on the topic, with many variations of the James-Stein estimator being developed in the 1960s and the 1970s.

More recently, work on shrinkage estimation has experience a resurgence with the consideration of high-dimensional genomic data. Cui et al. (2005) demonstrate the increase in power for multiple testing by pooling variance estimates using Empirical Bayes ideas. Ghosh (2006) directly applied James-Stein esimtation theory to p-values. While the theory was not completely satisfactory, Ghosh (2012b) has found that there is an implicit shrinkage phenomenon going on in the direct FDR estimation framework of Storey (2002). Hwang and Liu (2010) derive shrunken test statistics in which both estimates are pooled across using both means and variances. A related problem is confidence interval construction with

high-dimensional data. Ghosh (2009), Hwang et al. (2010) and Zhao (2010) have proposed shrinkage-based methods for computing CIs here as well.

2.3 Spacings and the B-H procedure

In this section, we assume that the n $p-$values are a random sample from the Uniform$(0,1)$. This corresponds to the situation where all hypotheses are true nulls, i.e. $n = n_0$. Define the order statistics of the p-values to be $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \cdots \leq p_{(n)}$. The spacings are defined as

$$p_i^S = p_{(i)} - p_{(i-1)},$$

for $i = 1, \ldots, n+1$, where $p_{(0)} = 0$ and $p_{(n+1)} = 1$. We now recall the following facts from Pyke (1965). First, the joint density of the spacings is

$$f(v_1, \ldots, v_n) = \begin{cases} n! \text{ if } v_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{n+1} v_i = 1 \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

We note that $f(v_1, \ldots, v_n)$ is defined on the region $\{\sum_{i=1}^{n+1} v_i = 1, v_i \geq 0, i = 1, \ldots, n+1\}$. Due to this constraint, it follows that the spacings are not independent random variables. Equivalent, because the order statistics are not independent, neither are the spacings. Note that the marginal distribution of the spacings is Beta$(1, n)$; the pdf is given by

$$f_v(v) = n(1-v)^{n-1}, \quad v > 0.$$

This implies that the mean and variance of the spacing is given by $(n+1)^{-1}$ and $n(n+1)^{-2}(n+2)^{-1}$, respectively. Furthermore, Pyke (1965) shows that $E(p_j^S p_k^S) = (n+1)^{-1}(n+2)^{-1}$ for $1 \leq j, k \leq n+1$. This implies that the correlation between two spacings is given by

$$-\frac{1}{(n+1)^2(n+2)}.$$

Another implication implicit in the derivation of the joint distribution of the spacings is that it is symmetric in all variables and hence invariant under permutation.

A result of Ghosh (2012a) shows that B-H procedure can be represented as the following: reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max\{1 \leq i \leq n : i^{-1} \sum_{j=1}^{i} p_j^S \leq \alpha(n+1)n^{-1}E(p_1^S)\}, \tag{4}$$

with $\hat{k} = 0$ if the set in (4) is empty. Thus, the B-H procedure can be interpreted as comparing a scan statistic or cumulative average of spacings relative to its expected value, scaled by a constant times α, the target FDR that we wish to control. For practical purposes, when n is large, we can remove the $(n+1)/n$ term from (4).

3 Shrinkage methodology

3.1 Proposed estimators

Recognizing that the function involving the observed spacings in (4) is in fact a sample average, our proposals involve replacing the sample average in (4) by a James-Stein estimator for the spacings. To do this will require the facts about spacings that we described in Section 2.3. We consider two strategies for constructing such shrinkage estimators.

The first approach is to use the following working model for the spacings:

$$p_i^S|\mu_i \overset{ind}{\sim} (\mu_i^S, \sigma^2) \tag{5}$$

$$\mu_1^S, \ldots, \mu_n^S \overset{iid}{\sim} F^S. \tag{6}$$

In the hierarchical model defined by (5-6), μ_i^S is the mean of the spacings, and F^S is the distribution of the mean. We also assume at the first stage that the variance of the spacings is known. If we assume that both distributions in (5) and (6) are normal, then this corresponds to the hierarchical model used by Efron and Morris (1973) to motivate the James-Stein estimator. From the properties of spacings described in Section 2.3, we have that $\mu_i^S = (n+1)^{-1}$ and that $\sigma^2 = \mathrm{Var}(p_i^S) \equiv n(n+1)^{-2}(n+2)^{-1}$. This leads to what we term the full-case James-Stein estimator of μ_i^S: for $i = 1, \ldots, n$,

$$\tilde{p}_i^S = \left[1 - \frac{(n-2)\sigma^2}{\sum_{i=1}^n \{p_i^S\}^2}\right] p_i^S, \tag{7}$$

where $\sigma^2 \equiv n(n+1)^{-2}(n+2)^{-1}$. Using the full-case J-S estimator, we can modify our rule (4) to the following: reject $p_{(1)}, \ldots, p_{(\tilde{k})}$, where

$$\tilde{k} = \max\{i : i^{-1} \sum_{j=1}^{i} \tilde{p}_j^S \leq \alpha/n\}, \tag{8}$$

with $\tilde{k} = 0$ if the set in (8) is empty.

A few remarks are in order. First, we term this the full James-Stein estimator because the estimator defined in (7) uses information from all n p-values. It is directly analogous to the classical James-Stein estimator in the sample mean setting. Second, strictly speaking, the model (5)-(6) cannot be true for the spacings. In particular, the model would imply *positive* correlation for the spacings, whereas the results from Pyke (1965) show that the opposite is true. One alternative approach is to construct full-case shrinkage estimators that exploit correlation information (e.g., Bock, 1975). However, because n is so large and the covariance matrix is not full-rank, we found the estimator to be too computationally extensive in our numerical experiments. We have chosen to shrink the spacings to zero as evidenced in (7). However, we could also shrink to other targets by modification of the formula, replacing p_i^S by $p_i^S - c$ and adding c, where c represents the new value to which we wish to shrink.

We also explored the use of a sequential James-Stein estimator. The idea here is to explicitly use the ordering of the p-values in the construction of the shrinkage estimators for the spacings. The modification of the estimator $\tilde{p}_i^S$, $i = 1, \ldots, n$, is as follows: define

$$\dot{p}_i^S = \left[1 - \frac{(i-1)\sigma^2}{\sum_{k \le i}\{p_k^S\}^2}\right] p_i^S, \tag{9}$$

and replace $\tilde{p}_i^S$ with $\dot{p}_i^S$ in the rule (8). Now note that this is a variable shrinkage estimator in which different spacings will receive differing amounts of shrinkage depending on which order statistic the spacing corresponds to. In particular, spacings associated with smaller p-values will receive less shrinkage those those corresponding to larger p-values. However, our simulation studies showed that this estimator performed poorly relative to the full James-Stein estimator, so we did not consider it further in our study.

3.2 A paradox: single hypothesis versus multiple hypothesis testing

Before describing the proposed methodology, we digress and describe a basic paradox in multiple testing using the p-values which is brought out in the spacings viewpoint adopted here. . In the case of a single hypothesis H_0, we perform a hypothesis test, obtain a p-value, and compare the p-value to the significance level. If the former is smaller, then we reject H_0; otherwise, we fail to reject H_0.

Now, we move to the multiple testing setup in which we have p-values $p_1, \ldots, p_n$. Now what is more germane is the *ensemble behavior* of the p-values. The rule (4) implies that we reject hypotheses that are grouped closer to each other than expected. However, this could be in any part of the distribution of p-values. As an example, we simulate $n = 10000$ p-values from the following mixture distribution:

$$p_1, \ldots, p_n \overset{iid}{\sim} 0.9\text{Uniform}(0,1) + 0.05\text{Beta}(30,1) + 0.05\text{Beta}(1,30).$$

A plot of the data is given in Figure 1. Note that there are two peaks in the data, one around 0.2, the other around 0.8. If we take the classical view regarding hypothesis testing, then we should only care about p-values on the left-hand side of the picture, i.e. the p-values that are close to zero. However, what is suggested to us by the spacings interpretation of the B-H rule is that **both** sets of p-values near the peak and their clumping are suggestive of evidence against the null hypothesis. Thus, while any individual p-value near the second peak (the one on the right in Figure 1) would typically imply lack of evidence against rejection of the null hypothesis, the fact that there are many p-values near 0.8 more than is expected under an assumption that all the p-values are uniformly distributed on the interval $[0, 1]$. This is the source of the paradox between single-hypothesis and multiple hypothesis testing. It is for this reason that Efron (2004) has argued that the theoretical null distribution is an inappropriate one for hypothesis and that one should use an empirical null hypothesis instead. The idea of the empirical null hypothesis is to estimate the null distribution of test statistics

using the observed test statistics so that decisions about rejecting hypotheses are done using the estimated null distribution. To estimate this distribution, Efron (2004) utilizes a mixture model formulation in conjunction with a zero-matching assumption that allows for estimation. In more recent work, Ghosh (2012a) directly incorporates the empirical null hypothesis into the B-H procedure without requiring use of the zero-matching assumption.

The simulation also suggests that multiple hypothesis testing should be viewed more as detection of local phenomenon. Coming back to Figure 1, we see that the evidence comes from p-values near 0.2 and 0.8 and suggests that the evidence in the multiple testing framework is local in nature. This is one of the immediate implications of the spacings results for the B-H procedure.

3.3 Optimality properties

One property for the full James-Stein estimator is minimaxity (Lehmann and Casella 2002, §5.1, p. 309). A minimax estimator T^M for a parameter μ minimizes the maximum expected loss, i.e.,

$$\inf_T \sup_\mu R(\mu, T) = \sup_\mu R(\mu, T^M).$$

We have the following result:

Theorem 1: Under model (5)-(6), the full-case James-Stein estimator is asymptotically minimax.

Proof: The form of $\tilde{p}_i^S$ satisfies the regularity conditions and represents a super harmonic function in the sense of Brown (1971). By the results there, the full-case James-Stein estimator is minimax.

Note that ultimately our interest is not in the estimators per se, but rather the operating characteristics of the stopping rules corresponding to modifications of the B-H procedure with the James-Stein estimators. This will be assessed by simulation studies in Section 4.2.

3.4 Shrunken generalized B-H procedures

An extension of the B-H procedure was developed in Ghosh (2011). He proposed the rule to select the hypotheses corresponding to $p_{(1)}, \ldots, p_{(\tilde{k}^*)}$ as being rejected, where

$$\tilde{k}^* = \max\{i : i^{-1} \sum_{j=1}^{i} g_\lambda(p_j^S) \leq \alpha E\{g_\lambda(p_1^S)\}\}, \tag{10}$$

and $g(z) = z^\lambda$. As usual, if the set in (4) is empty, then we fail to reject any hypotheses. This is referred to in Ghosh (2011) as the generalized Benjamini-Hochberg procedure. Inspection of (10) relative to (4) reveals that we simply applied the $g()$ function to the spacings as well as to the expectation on the right-hand side of the inequality in (4). Note that if $\lambda = 1$, we get back the usual B-H procedure. The idea behind generalized Benjamini-Hochberg procedures is that by transforming the spacings, we can implicitly incorporate information about the

distribution of the spacings beyond first moments. Ghosh (2011) found using simulation studies that the choice of $\lambda = 2$ provided FDR control as well as increased power relative to the usual B-H procedure. Here and in the sequel, we will deal with the $\lambda = 2$ case.

Again noting that the rule in (4) involves a sample average, we can apply James-Stein estimation in a manner analogous to that given in Section 3.1. Similarly, we can develop minimaxity results for the James-Stein versions of the generalized B-H procedures as well. For the sake of exposition, we take $g(z) = z^2$ here. Then using algebraic manipulations, the variance of $g(p_i^S)$ is given by

$$\sigma_*^2 \frac{20n^2 + 44n}{(n+1)^2(n+2)^2(n+3)(n+4)}.$$

For $i = 1, \ldots, n$, if $q_i = g(p_i^S)$, then a shrunken estimator is given by

$$\tilde{q}_i = \left[1 - \frac{(n-2)\sigma^2}{\sum_{k=1}^n \{q_k\}^2}\right] q_i, \tag{11}$$

The rule becomes to reject the hypotheses $p_{(1)}, \ldots, p_{(\tilde{k}^*)}$ as being rejected, where

$$\tilde{k}^* = \max\{1 \leq i \leq n : i^{-1} \sum_{j=1}^{i} \tilde{q}_j \leq \alpha E\{q_1\}\}, \tag{12}$$

with no rejections occurring if the set defined in (12) is empty.

4 Numerical Examples

4.1 Microarray datasets

We applied the proposed methodology to several gene expression datasets. The first was the acute leukemia (AML/ALL) dataset from the Golub et al. (1999) gene expression study. In this study, there were $n = 7128$ p-values considered, calculated from unpooled two-sample t-tests comparing the samples in the AML group (11 samples) to those in the ALL group (27 samples). Using an FDR of 0.05, the BH procedure selected 1270 genes, the shrunken BH procedure 1495, the generalized BH procedure 1986 genes, and the shrunken generalized BH procedure 2189 genes.

Next, we applied the methods to the HIV dataset that has been considered in Efron (2004). In this study, there were $n = 7680$ p-values from two-sample tests comparing individuals with and without HIV. Based on controlling the FDR at level 0.05, the BH procedure selected 1649 genes, the shrunken BH procedure 2012 genes, the generalized BH procedure 3238 genes, and the shrunken generalized BH procedure 3365 genes.

A third data example is the prostate cancer dataset described in Efron (2010). This example consists of performing hypothesis tests on $n = 6033$ genes on subjects with and without prostate cancer. Suppose we wish to find differentially expressed genes at an FDR of 0.05. For these data, the BH procedure selected 21 genes, the shrunken BH procedure 59

genes, the generalized BH procedure 90 genes, and the shrunken generalized BH procedure 182 genes.

4.2 Simulation studies

Here, we report on the results of some simulation studies to assess the finite-sample properties of the proposed methodologies. We use a model as in Ghosh (2011), generating $n = 300$ statistics based on a multivariate normal distribution with mean μ_i and variance one, $i = 1, \ldots, n$. For the alternative statistics, n_1 of them have $\mu_i = 2$; for the true null hypotheses, for $i = 1, \ldots, n$, $\mu_i = 0$. Data were simulated assuming a correlation of zero (independent case) and 0.3 (dependent case). We compared the performance of the BH procedure, gBH procedure and their shrunken counterparts. The results of the first two are taken from Ghosh (2011). The setup of FDR and power is similar to that article as well.

The results are given in Tables 2 and 3. All procedures are controlled at an FDR of 0.05. We find that in terms of FDR, the operating characteristics of the BH procedure and the proposed methodology with $\lambda = 2$ are quite similar in the independent case. In addition, the proposed procedures enjoy much higher power relative to the B-H procedure; the difference is more pronounced in the dependent case.

5 Discussion

In this article, we have shown how one can develop James-Stein estimators within the classical multiple testing procedures. By resorting to results from spacings theory, several facts and implications obtain. First, the B-H procedure has a natural interpretation as comparing the empirical averages of spacings relative to its theoretical expected value, scaled by the target FDR one wishes to control at. Second, one can think of multiple hypothesis testing as a clustering problem. In particular, the spacings framework shows that *local* deviations from the ensemble null hypothesis are indicative of rejection in the multiple comparisons problem. Exactly how to specify the ensemble null distribution is a problem that remains to be investigated further. The results indicate that shrinkage appears to increase power for rejection in the multiple testing framework.

The proposed methodology developed here is a sensible extension of James-Stein estimation ideas to multiple testing. What is currently lacking is a theory to justify the error control properties of the proposed procedures. As alluded to in Efron (2010), this is a common type of dilemma encountered in the development of Empirical Bayes solutions to problems.

References

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

Bock, M. E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *Annals of Statistics* **3**, 209 – 218.

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

Cui, X., Hwang J. T., Qiu, J., Blades, N. J. and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59-75.

Efron, B. (2004). Selection and estimation for large-scale simultaneous inference. *Journal of the American Statistical Association* **99**, 96 – 104.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge: Cambridge University Press.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – An empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117 – 130.

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

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

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

Ghosh, D. (2009). Empirical Bayes methods for estimation and confidence intervals in high-dimensional problems. *Statistica Sinica* **19**, 125 – 143.

Ghosh, D. (2011). Generalized Benjamini-Hochberg procedures using spacings. *Journal of the Indian Society of Agricultural Statistics* **65**, 213 – 220.

Ghosh, D. (2012a). Incorporating the empirical null hypothesis into the Benjamini-Hochberg procedure, *Statistical Applications in Genetics and Molecular Biology*, in press.

Ghosh, D. (2012b). Shrinkage in adaptive false discovery rate procedures for multiple testing: structure and synthesis, submitted. Available at `http://works.bepress.com/debashis_ghosh/53/`.

Golub, T. R., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531 – 537.

Hwang, J. T. and Liu, P. (2010). Optimal tests shrinking both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology* **9**, Article 36.

Hwang, J. T., Qiu, J., and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society Series B* **71**, 265 – 285.

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.

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

Pyke R. (1965). Spacings (with discussion). *Journal of the Royal Statistical Society Series B* **27**, 395 – 449.

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

Zhao, Z. (2010). Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Statistics and Its Interface* **3**, 533 – 541.

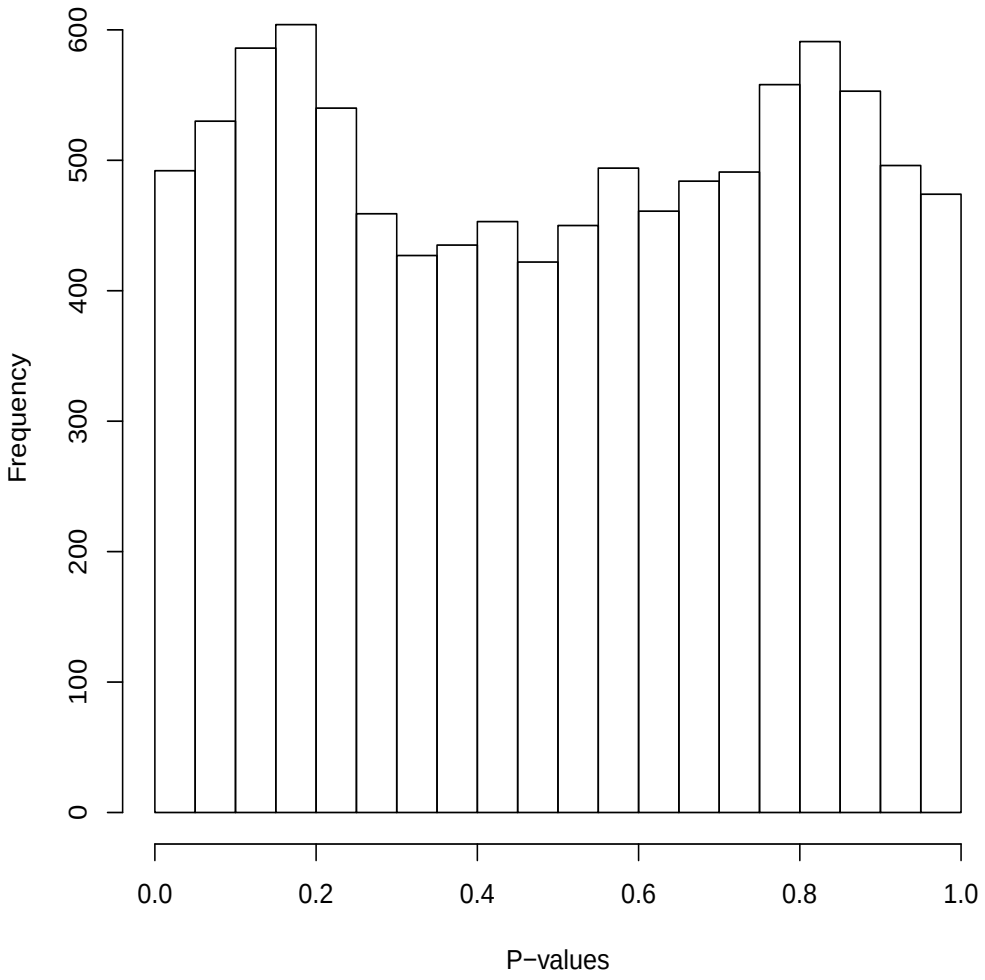

Figure 1: Simulated p-values from mixture model in Section 3.2.

Table 2: FDR results of simulation studies from Section 4.2.

	Independent			Dependent		
Method/n_1	20	60	100	20	60	100
BH	0.04	0.04	0.03	0.001	0.005	0.004
sBH	0.03	0.04	0.4	0.02	0.01	0.03
gBH	0.002	0.03	0.002	0.04	0.03	0.02
sgBH	0.04	0.04	0.04	0.04	0.04	0.04

Table 3: Power results of simulation studies from Section 4.2.

	Independent			Dependent		
Method/n_1	20	60	100	20	60	100
BH	0.60	0.96	0.99	0.34	0.63	0.71
sBH	0.98	1.00	1.00	0.78	0.89	0.99
gBH	0.94	0.98	0.99	0.73	0.89	0.93
sgBH	0.99	1.00	1.00	0.8	0.90	0.98