Optimal Tests Shrinking Both Means and Variances Applicable to Microarray Data Analysis

J.T. Gene Hwang  
Cornell University

Peng Liu  
Iowa State University, pliu@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/stat_las_preprints

Part of the Statistics and Probability Commons

Recommended Citation
http://lib.dr.iastate.edu/stat_las_preprints/59
Optimal Tests Shrinking Both Means and Variances Applicable to Microarray Data Analysis

Abstract
As a consequence of “large p small n” characteristic for microarray data, hypothesis tests based on individual genes often result in low average power. There are several proposed tests that attempt to improve power. Among these, FS test developed using the concept of James-Stein shrinkage to estimate the variances, showed a striking average power improvement. In this paper, we derive the FS test as an empirical Bayes likelihood ratio test, providing a theoretical justification. To shrink the means also, we modify the prior distributions leading to the optimal Bayes test which is called MAP test (where MAP stands for Maximum Average Power). Also an FSS statistic is derived as an approximation to MAP and can be computed instantaneously. The FSS shrinks both the means and the variances and has a numerically identical average power as MAP. Simulation studies show that the proposed test performs uniformly better in average power than the other tests in the literature including the classical F test, FS test, the test of Wright and Simon, moderated t-test, SAM, Efron's t test and B statistics. A theory is established which indicates that the proposed test is optimal in power when controlling the false discovery rate (FDR).

Keywords
empirical Bayes test, false discovery rate (FDR), FS test, Neyman-Pearson lemma

Disciplines
Statistics and Probability

Comments

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/stat_las_preprints/59
Optimal Tests Shrinking Both Means and Variances Applicable to Microarray Data Analysis

J. T. Gene Hwang * and Peng Liu†

April 9, 2007

Abstract

As a consequence of “large p small n” characteristic for microarray data, hypothesis tests based on individual genes often result in low average power. There are several proposed tests that attempt to improve power. Among these, $F_S$ test developed using the concept of James-Stein shrinkage to estimate the variances, showed a striking average power improvement. In this paper, we derive the $F_S$ test as an empirical Bayes likelihood ratio test, providing a theoretical justification. To shrink the means also, we modify the prior distributions leading to the optimal Bayes test which is called $MAP$ test (where $MAP$ stands for Maximum Average Power). Also an $F_{SS}$ statistic is derived as an approximation to $MAP$ and can be computed instantaneously. The $F_{SS}$ shrinks both the means and the variances and has a numerically identical average power as $MAP$. Simulation studies show that the proposed test performs uniformly better in average power than the other tests in the literature including the classical $F$ test, $F_S$ test, the test of Wright and Simon, moderated $t$-test, SAM, Efron’s $t$ test and $B$ statistics. A theory is established which indicates

---

*Gene Hwang is Professor, Department of Mathematics and Department of Statistical Science, Cornell University, Ithaca, NY 14853 (email: hwang@math.cornell.edu) and Adjunct Professor, Department of Statistics, National Cheng Kung University, Taiwan.

†Peng Liu is Assistant Professor, Department of Statistics, Iowa State University, Ames, IA 50010 (email: pliu@iastate.edu).
that the proposed test is optimal in power when controlling the false
discovery rate (FDR).

Keywords: False discovery rate (FDR), Neyman-Pearson funda-
mental lemma, $F_S$ test, $F_{SS}$ test, empirical Bayes likelihood ratio
test.

1 INTRODUCTION

Microarray technology has been applied widely in biomedical research to
measure expression levels of thousands of genes simultaneously. A typical
goal of microarray experiments is to identify genes that are differentially
expressed across different treatments. One can apply $F$ test based on data
of each individual gene, a test called $F_1$ in Cui et al. (2005). However,
we are in a “large $p$ small $n$” scenario, i.e., there are a large number ($p$)
of genes and a small number ($n$) of replicates in each gene. The power
of $F_1$ test can be substantially improved by “borrowing strength” across
all genes. Several procedures have been proposed including SAM (Tusher
et al. 2001), Efron’s $t$-test (Efron et al. 2001), regularized $t$-test (Baldi and
Long 2001), $B$-statistic (Lönnstedt and Speed 2002) and its multivariate
counterpart, the $MB$-statistic (Tai and Speed 2006), the tests of Wright and
Simon (2003), moderated $t$-test (Smyth 2004), the $F_S$ test (Cui et al. 2005)
and the test of Tong and Wang (2007) which is similar to $F_S$ test. All these
tests, except the $B$-statistics, modify the $t$-test (or equivalently, $F$ test) by
shrinking the variances or the standard errors only. Take SAM test as an
example, the standard error $\hat{\sigma}$ in $t$-test is replaced by $\hat{\sigma} + s_0$, where $s_0$ is
chosen depending on the data of all genes. If we replace $\hat{\sigma} + s_0$ by $(\hat{\sigma} + s_0)/2$,
the test is unchanged. However, this shows that SAM shrinks the standard
errors toward $s_0$ with the shrinkage factor $1/2$. The $F_S$ test shrinks the variances. Unlike SAM, it uses a shrinkage factor depending on data, which seems more desirable. Specifically in $F_S$, the variance estimator in log scale is based on applying James-Stein-Lindley estimator to the log of unbiased variance estimator. The $F_S$ test, now routinely used in Jackson Lab, has a larger average power than $F_1$ in all the fairly extensive numerical studies. This calls for a theory. In this paper, we derive the $F_S$ test as an empirical Bayes likelihood ratio test, which justifies $F_S$, to some extent, as an optimal test.

The work of Cui et al. (2005) leads to a natural question why only shrinking the variances but not the means? To do so, we modify the prior distribution and derive the most powerful test, MAP test. Here MAP stands for Maximum Average Powerful, a term first coined in Chen et al. (2007). This test is computationally extensive. A first order approximation leads to the $F_{SS}$ test where SS stands for double shrinkage, shrinking both the means and the variances. The $F_{SS}$ test has almost identical power as MAP test and is more powerful than $F_S$ test and all the other tests cited above. Furthermore, the $F_{SS}$ statistic is explicit and can be computed instantaneously. A fast computation is a big advantage considering the dimensionality of tests for microarray data analysis, not to mention that often a large number of permutations are needed for each test.

Two other procedures that are published (or in press) very recently and not included in our numerical studies are commented below. First, Lo and Gottardo (2007) extended the empirical Bayes test developed by Newton et al. (2001) and Kendziorski et al. (2003) to the important case when the
variances corresponding to different genes are assumed different. However, the simulation results of Lo and Gottardo (2007) indicate that their procedures at best behave similarly in power to the moderated $t$-test (Smyth 2004) which is not as powerful as $F_{SS}$ test. Second, Storey’s optimal procedure (Storey 2007) may also be as powerful as $F_{SS}$ test. However, it is computationally intensive and we find it time-consuming to compute for thousands of tests, a typical number of tests for microarray data. Since $F_{SS}$ is instantaneously in computation, it is more applicable for microarray data.

We have focused on maximizing the average power by controlling the average type I error rate when comparing tests, a criterion also used in Cui et al. (2005). Storey (2007) argues that it is the right criterion to use for deriving the optimal multiple test. Inspired by Storey’s paper, we prove a theorem that shows the criterion is equivalent to controlling the false discovery rate (FDR) and maximizing the average power. This shows that our proposed test is optimal either controlling FDR or type I error rate.

2 $F$-LIKE TESTS

Suppose ANOVA models are fitted for each gene. In this section, we focus on testing a one-dimensional parameter $\theta_g$, $1 \leq g \leq G$, which is a linear function of $\beta_g$, the coefficient of the $g$-th ANOVA model corresponding to the $g$-th gene. Let $\hat{\theta}_g$ be the ANOVA estimator of $\theta_g$. A typical $F$ tests for $H_0^g : \theta_g = 0$ vs $H_1^g : \theta_g \neq 0$ is to reject if

$$\frac{(\hat{\theta}_g)^2}{\hat{\sigma}_g^2} > \text{crit}$$

(2.1)
where $\hat{\sigma}^2_g (MSE_g)$ is the unbiased estimator of the variance of $\hat{\theta}_g$ and crit denotes some generic critical value that is also used later in this paper. Traditionally, it is assumed that

$$\hat{\theta}_g \sim N(\theta_g, \sigma^2_g) \quad (2.2)$$

and

$$\hat{\sigma}^2_g \sim \sigma^2_g \chi^2_d \quad (2.3)$$

where $\chi^2_d$ is a chi-squared random variable with $d$ degrees of freedom and $d$ depends on the ANOVA model. Under these assumptions, crit can be determined according to an $F$ distribution with one and $d$ degrees of freedom. However, in real application, crit is better determined by permutation, so that the procedure is applicable even without distributional assumptions (2.2) and (2.3). The comment about permutation applies to all the tests discussed in the paper and is applied in some of the figures.

The test in (2.1) is called the $F_1$ test in Cui et al. (2005). If one assumes that all $\sigma^2_g, g = 1, ..., G$, are identical, then it is desirable to use $F_3$ test which, as defined in Cui and Churchill (2003) and Cui et al. (2005), rejects $H^g_0$ if and only if

$$\frac{(\hat{\theta}_g)^2}{(\sum \hat{\sigma}^2_g) / G} > \text{crit},$$

(2.4)

The test $F_3$ is expected to have a larger power when $\sigma^2_g, g = 1, ..., G$, are identical. But it fails miserably when $\sigma^2_g, g = 1, ..., G$ are very different.

This prompts the authors in Cui et al. (2005) to propose the $F_S$ test that is similar to $F_1$ except that the variance estimator shrinks $\hat{\sigma}^2_g$ by a logarithmic transformation and an application of James–Stein–Lindley estimator (Lindley 1962). Taking the log of (2.3) gives $ln(\hat{\sigma}^2_g) = ln(\sigma^2_g) + ln(\chi^2_d / d)$. Let
\( X_g' = \ln(\hat{\sigma}_g^2) - E(\ln(\chi^2_d/d)). \) Then \( X_g' \sim \ln \sigma_g^2 + \epsilon_g' \), where \( \epsilon_g' = \ln(\chi^2_d/d) - E(\ln(\chi^2_d/d)) \) with mean zero and variance \( V = \text{Var}(\ln(\chi^2_d/d)) \). Both the mean and the variance of \( \ln(\chi^2_d/d) \) can be determined easily by numerical method. Then empirical Bayes or James–Stein–Lindley shrinkage estimator of \( \ln(\sigma_g^2) \) is:

\[
\hat{X}' + \left( 1 - \frac{(G-3)V}{\sum (X'_g - \bar{X}')^2} \right) \times (X'_g - \bar{X}').
\]

Taking exponential of the estimator produces a shrinkage estimator of \( \sigma_g^2 \) and is denoted as \( \hat{\sigma}_{EB}^2 \). Now, \( F_S \) test rejects the null hypothesis if

\[
\frac{(\hat{\theta}_g)^2}{\hat{\sigma}_{EB}^2} \text{ is large.} \quad (2.5)
\]

The hope is that \( F_S \) would have good power no matter whether \( \sigma_g^2 \)'s are similar or are very different across genes. Indeed, Cui et al. (2005) showed that \( F_S \) has average power never less than \( F_1 \) and \( F_3 \) and is strikingly more powerful than \( F_1 \) and \( F_3 \) in various situations.

### 3 Optimality of the \( F_S \) Test

The results of Cui et al. (2005) show that their rejection region has good average power and also satisfies the condition that the average type I error is controlled to be less than or equal to \( \alpha \). Note that the average power is

\[
\frac{1}{G_1} \sum P_{\theta, \sigma_g^2}(H_0^g \text{ is rejected}) \quad (3.1)
\]

and the average type I error rate is similar to (3.1) with \( \theta=0 \):

\[
\frac{1}{G_0} \sum P_{\sigma_g^2}(H_0^g \text{ is rejected}). \quad (3.2)
\]

In the above notation, \( G_0 \) and \( G_1 \) denote the numbers of \( \theta_g \)'s (genes) which satisfy null hypotheses and alternative hypotheses, respectively. The total
number of genes is \( G = G_0 + G_1 \). Here we focus on the case that \( \theta \) does not depend on \( g \). A more complicated theory will be derived that applies to the more realistic setup in Section 4.

Similar to the works in Cui et al. (2005) and Storey (2007), we focus on the rejection regions, a collection of \((\hat{\theta}_g, \hat{\sigma}^2_g)\), that do not depend on \( g \). Storey (2007) gave a theory that, under an exchangeable setting, there is no loss of power to focus on such rejection regions. When \( \sigma^2_g \)'s are assumed to be random variables having the same distribution with the probability density function (p.d.f.) \( \pi(\cdot) \), (3.1) converges to

\[
\int P_{\theta, \sigma^2}(H_0 \text{ is rejected})\pi(\sigma^2)d\sigma^2. \tag{3.3}
\]

Here the subscript \( g \) in \( \sigma^2_g \) (and later in \( \hat{\theta}_g \)) is suppressed since (3.3) does not depend on \( g \) anymore. Also (3.2) converges to (3.3) with \( \theta = 0 \).

Since \( G, G_0 \) and \( G_1 \) are big for microarray data, we should look at the approximate problem of maximizing (3.3) given that (3.3) with \( \theta = 0 \) is controlled to be \( \alpha \).

The most powerful test can then be constructed for testing \( H_0 : \theta_g = 0 \) vs. \( H_1 : \theta_g = \theta, \theta \neq 0 \), using Neyman–Pearson fundamental lemma which rejects \( H_0 \) if

\[
\frac{\int f(\hat{\theta} | \theta_g = \theta, \sigma^2)f(\hat{\sigma}^2 | \sigma^2)\pi(\sigma^2)d\sigma^2}{\int f(\hat{\theta} | \theta_g = 0, \sigma^2)f(\hat{\sigma}^2 | \sigma^2)\pi(\sigma^2)d\sigma^2} \text{ is large.} \tag{3.4}
\]

Here and later \( f \) is a generic notation representing the p.d.f. For example, \( f(\hat{\sigma}^2 | \sigma^2) \) denotes the conditional distribution of \( \hat{\sigma}^2 \) given \( \sigma^2 \). The left hand side of (3.4) is also called the Bayes factor by Bayesian statisticians. See, for example, Robert (2001), page 227.
However, θ is unknown. More generally, we test $H_0 : \theta_g = 0$ vs $H_1 : \theta_g \neq 0$. Then, a likelihood ratio test statistic should maximize the left hand side of (3.4) with respect to θ, i.e.,

$$\sup_\theta \int f(\hat{\theta} | \theta, \sigma^2) f(\hat{\sigma}^2 | \sigma^2) \pi(\sigma^2) d\sigma^2 \over \int f(\hat{\theta} | \theta = 0, \sigma^2) f(\hat{\sigma}^2 | \sigma^2) \pi(\sigma^2) d\sigma^2.$$  

(3.5)

This leads to replacing θ by $\hat{\theta}$ where $\hat{\theta}$ is the maximum likelihood estimate (MLE). Hence (3.5) can be interpreted as the estimated most powerful test.

We shall work with a model similar to (2.2) and (2.3) with the exception that $\hat{\sigma}^2 = \sigma^2 K$, where $\sigma^2$ and $K$ both have log-normal distributions. More specifically, we assume that

$$\hat{\theta} \sim N(\theta, \sigma^2) \text{ and } \hat{\rho}_0 = \rho + \ln K,$$  

(3.6)

where $\hat{\rho}_0 = \ln \hat{\sigma}^2$, $\rho = \ln \sigma^2$, $\rho \sim N(\mu_V, \tau_V^2)$, and $\ln K \sim N(\mu_K, \sigma_K^2)$. We use the subscript $V$ in $\mu_V$ and $\tau_V^2$ since they are related to the variance $\sigma^2$. Note that if we set $K$ to be $\chi^2_d/d$, then $\hat{\sigma}^2$ would reduce to (2.3). Instead, we approximate $\ln(\chi^2_d/d)$ by $N(\mu_K, \sigma_K^2)$ where $\mu_K$ and $\sigma_K^2$ are taken to be the mean and variance of $\ln(\chi^2_d/d)$. This would simplify the test and its computation. Simulation indicates that the approximation works well. See a comment at the end of Section 5. We could also subtract $\mu_K$ from both sides of the equation in (3.6) and write it as $\hat{\rho} = \rho + \ln K - \mu_K = \rho + \delta$ where $\delta = \ln K - \mu_K \sim N(0, \sigma_K^2)$ and $\hat{\rho} = \hat{\rho}_0 - \mu_K$. Hence $\hat{\rho}$ is identical to $X'$ in Section 2.

**Theorem 1.** Under (3.6) with a fixed $\mu_V$ and $\tau_V^2$, the likelihood ratio test for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ rejects $H_0$ if and only if the statistic
$\hat{\theta}^2/\hat{\sigma}_{\rho}^2$ is larger than some critical value, where

$$\hat{\rho}_p = M_V \hat{\rho} + (1 - M_V) \mu_V \quad \text{and} \quad M_V = \frac{\tau_V^2}{(\tau_V^2 + \sigma_K^2)}.$$  \hspace{1cm} (3.7)

The $\hat{\rho}_p$ and $\hat{\sigma}^2_p$ with subscript $p$ are the estimators of $\rho$ and $\sigma^2$ based on the posterior distribution. In particular, $\hat{\rho}_p$ is the posterior mean of $\rho$ given $\hat{\rho}$. Note that $\hat{\sigma}^2_p$ is equivalent to $(\hat{\sigma}^2)_M$ after omitting constants such as $exp((1 - M_V) \mu_V)$. Hence the statistic $\hat{\theta}^2/\hat{\sigma}_{\rho}^2$ is equivalent to

$$\hat{\theta}^2/(\hat{\sigma}^2)_M. \hspace{1cm} (3.8)$$

When $M_V = 1$, the statistic (3.8) reduces to $F_1$, which is the right statistic to use since $M_V = 1$ implies that $\sigma^2_g$ are very different from each other. Similarly if $M_V = 0$, the statistic (3.8) is equivalent to $F_3$, since the denominator of $F_3$ is equivalent to a constant by the law of large numbers. The statistic $F_3$ is the right statistic to use since $M_V = 0$ implies that $\sigma^2_g$ are identical.

However, the most practical case is that $M_V$ is unknown and should be estimated by data, leading to the empirical Bayes test below. Following the Lindley–James–Stein approach (Lindley 1962), we replace $\hat{\rho}_p$ and $\hat{\sigma}^2_p$ by

$$\hat{\rho}_{EB} = \hat{\rho} + \left(1 - \frac{(G - 3)\sigma_K^2}{\sum(\hat{\rho}_g - \hat{\rho})^2}\right) (\hat{\rho}_g - \hat{\rho}), \quad \text{and} \quad \hat{\sigma}^2_{EB} = e^{\hat{\rho}_{EB}}. \hspace{1cm} (3.9)$$

This results in the test statistic in (2.5), i.e., the $F_S$ statistic in Cui et al. (2005).

The above argument derives $F_S$ as an empirical Bayes likelihood ratio test. The likelihood ratio test can be viewed as an approximation of the most powerful test. Hence the derivation explains why $F_S$ can have high power.
The test proposed by Wright and Simon (2003) and Smyth (2004) assumes \( \tilde{\sigma}_g^2 | \sigma_g^2 \sim \sigma_g^2 \chi^2_d / d \), i.e., \( K \) is distributed as \( \chi^2_d / d \) and \( \sigma_g^2 \) has a prior distribution of inverse gamma with parameters \( a \) and \( b \). We found that these two tests have power similar to \( F_S \) under the four possible combinations of distributional assumptions that (i) \( K \sim \chi^2_d / d \) or \( K \sim \text{log-normal} \), and (ii) \( \sigma_g^2 \) is either inverse gamma or log-normally distributed. Unlike \( F_S \), these two tests need to estimate \( a \) and \( b \) and are slightly more computationally intensive.

4 DERIVING A TEST MORE POWERFUL THAN \( F_S \)

The \( F_S \) test shrinks only the variances. Wouldn’t it be better if we shrink the mean too? We have tried to construct tests with shrinkage estimators for both the means and the variances, but the power of the resulting test is not necessarily better than \( F_S \) by our numerical results. The better way is to use the empirical Bayes approach to guide us in the search. In order to shrink the means, we assume, in addition to (3.6), that

\[
\theta \sim N(\mu, \tau^2). \tag{4.1}
\]

Similar normal assumption for the mean \( \theta \) has been used for deriving \( B \) statistic in Lönnstedt and Speed (2002) and the regularized \( t \) statistic in Baldi and Long (2001). The difference is that Lönnstedt and Speed (2002) assumed \( \mu = 0 \) and \( \tau^2 = c\sigma_g^2 \) for some constant \( c \) and that Baldi and Long (2001) assumed \( \mu \) is equal to the sample mean and \( \tau^2 = \sigma_g^2 / \lambda_0 \) for some
constant \( \lambda_0 \). Now we are testing

\[ H_0 : \theta = 0 \text{ vs } H_a : \theta \sim N(\mu, \tau^2). \]

At this point, we assume that \( \mu \) and \( \tau^2 \) are known. In real applications, \( \mu \) and \( \tau^2 \) should be estimated and we will describe the estimation in Section 7. Although we are making parametric assumptions in deriving our tests, the cutoff points of these tests could be determined using permutations, leading to tests valid non-parametrically.

By Neyman–Pearson fundamental lemma, the test that maximizes the average power

\[
\int \int \int \int_C f(\hat{\theta} \mid \theta, \sigma^2) f(\hat{\sigma}^2 \mid \sigma^2) \pi(\theta) \pi(\sigma^2) d\hat{\theta} d\hat{\sigma}^2 d\sigma^2 d\theta
\]

among all critical regions \( C \) such that

\[
\int \int \int_C f(\hat{\theta} \mid \theta = 0, \sigma^2) f(\hat{\sigma}^2 \mid \sigma^2) \pi(\theta) \pi(\sigma^2) d\hat{\theta} d\hat{\sigma}^2 d\sigma^2 \leq \alpha,
\]

is the test with \( C \) defined by

\[
\frac{\int \int f(\hat{\theta} \mid \theta, \sigma^2) f(\hat{\sigma}^2 \mid \sigma^2) \pi(\theta) \pi(\sigma^2) d\hat{\theta} d\hat{\sigma}^2 d\sigma^2 d\theta}{\int f(\hat{\theta} \mid \theta = 0, \sigma^2) f(\hat{\sigma}^2 \mid \sigma^2) \pi(\theta) \pi(\sigma^2) d\hat{\theta} d\hat{\sigma}^2 d\sigma^2} > \text{crit}
\]

where \( \text{crit} \) is determined so that this rejection region makes (4.3) achieve equality. Note also that \( \pi(\theta) \) and \( \pi(\sigma^2) \) are generic notation for the p.d.f.’s of \( \theta \) and \( \sigma^2 \). This test will be called the maximum average power (MAP) test, a term borrowed from Chen et al. (2007). This is also a Bayes test statistic. Integrate out \( \theta \) in the numerator, the left hand side of (4.4) equals

\[
\frac{\int \frac{1}{\sqrt{\sigma^2 + \tau^2}} e^{-\frac{1}{2}(\hat{\theta} - \mu)^2/\sigma^2 + \tau^2} f(\hat{\sigma}^2 \mid \sigma^2) \pi(\sigma^2) d\sigma^2}{\int \frac{1}{\sigma} e^{-\frac{1}{2}\hat{\sigma}^2/\sigma^2} f(\hat{\sigma}^2 \mid \sigma^2) \pi(\sigma^2) d\sigma^2}.\]
Here we are merely using the fact that, \( \hat{\theta} | \sigma^2 \sim N(\theta, \sigma^2) \) and \( \theta | \sigma^2 \sim N(\mu, \tau^2) \) imply that \( \hat{\theta} \sim N(\mu, \sigma^2 + \tau^2) \). Furthermore the above MAP test statistic can be written as

\[
\frac{E\left[ (\sigma^2 + \tau^2)^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta} - \mu)^2/(\sigma^2 + \tau^2)} | \hat{\sigma}^2 \right]}{E\left[ \sigma^{-1} e^{-\frac{1}{2}(\hat{\theta})^2/\sigma^2} | \hat{\sigma}^2 \right]}
\]

(4.5)

where \( E[\cdot | \hat{\sigma}^2] \) represents the integration of \( \sigma^2 \) with respect to the conditional distribution of \( \sigma^2 \) given \( \hat{\sigma}^2 \).

To apply (4.5), we need to calculate two integrals for each gene, which is computationally intensive. To avoid integration, we may use the first order approximation by estimating \( \sigma^2 \) with \( \hat{\sigma}^2_{EB} \) as defined in (3.9). This gives the statistic

\[
\left( \frac{\hat{\sigma}^2_{EB} + \tau^2}{\hat{\sigma}^2_{EB}} \right)^{-\frac{1}{2}} \cdot \left( e^{-\frac{1}{2}(\hat{\theta} - \mu)^2/(\hat{\sigma}^2_{EB} + \tau^2)}/ e^{-\frac{1}{2}(\hat{\theta})^2/\hat{\sigma}^2_{EB}} \right)
\]

(4.6)

The test that rejects \( H_0 \) if (4.6) is large is called \( F_{SS} \) test, where \( SS \) stands for the double shrinkage, shrinking both the means and the variances which will be explained in Section 6. The test \( F_{SS} \) is explicit and can be computed instantaneously.

5 NUMERICAL STUDIES OF POWER

We perform many numerical calculation partly based on simulation and partly based on real data and plot the power of various tests as reported in Figures 1 and 2. In all these graphs, we observe that \( F_{SS} \), having indistinguishable power from the computationally more intensive optimum \( MAP \) test, is more powerful than \( F_S \), which is more powerful than \( F_1 \) and \( F_3 \).

In numerical studies that generate Figures 1 and 2, we simulate data based on the canonical form of (2.2), (2.3) and \( \theta_g \sim N(\mu_\theta, \tau^2_\theta) \) for \( g=1, 2, ... \),
$G$, where $G$ is taken to be 10,000. The variances $\sigma_g^2$ are drawn randomly from the 15,600 residual variance estimates based on the tumor data set described in Cui et al. (2005). We also vary the coefficient of variation (CV) of the variances while keeping their geometric mean constant as in the same paper. This enables us to draw four different plots in each of Figures 1 and 2. In all these tests, the cutoff points are determined using simulated data when $\theta=0$ so that the average type I error rate is controlled to be 0.05. Different realistic values of degrees of freedom (df) and $\tau_\theta$ are used. These results are similar as what we present in Figures 1 and 2.

The result shows that the parametric assumption, although different from the tumor data, does not diminish the superiority of $F_{SS}$ over $F_S$ (and $F_S$ over $F_1$ and $F_3$).

We also studied the power of these tests under the model assumption that $K$ in (3.6) is simulated from log normal distribution instead of $\chi^2_{d/d}$. The power of these tests are similar to what are shown in Figures 1 and 2 and are not reported here.

Thus far, we derive the $F_{SS}$ test using log normal approximation. Without this, we could assume directly that $K \sim \chi^2_{d/d}$ and $\sigma^2$ is inverse gamma distributed with $a$ and $b$ as parameters and derive a test identical to (4.6) except that $\hat{\sigma}_{EB}^2$ is an empirical Bayes estimator for $\sigma^2$ based on the new setting and estimated $a$ and $b$ which is slightly more complicated than $F_{SS}$. The resultant test is demonstrated to have power indistinguishable from $F_{SS}$ under the four models depicted at the end of Section 3 and is not reported here. This comment applies to Figures 1, 2, 3 and 5. We expect it applies to Figure 4 as well.
Figure 1: The average power of MAP test and $F_{SS}$ test compared with other $F$-like tests. Data were simulated according to the canonical form as described in the text. Variances $\sigma^2_g$'s were randomly drawn from a data set in Cui et al. (2005) and the true mean effect $\theta_g$ were simulated from $N(\mu_\theta, \tau^2_\theta)$. Significance level was controlled at nominal 5% level of Type I error rate. The performances of MAP test and $F_{SS}$ test are compared with $F_1$, $F_3$ and $F_S$ for various coefficient of variation (CV) of variances.
6 SHRINKING BOTH MEANS AND VARIANCES

To understand what (4.6) does, it is worthwhile to look at the likelihood ratio:

\[ LR(\theta, \sigma^2) = \frac{e^{-\frac{1}{2}(\hat{\theta} - \theta)^2/\sigma^2}}{e^{-\frac{1}{2}\hat{\theta}^2/\sigma^2}}, \quad (6.1) \]

which is the Neyman–Pearson statistic for testing \( H_0 : \theta_g = 0 \) vs. \( H_1 : \theta_g = \theta \) based on \( \hat{\theta}_g \sim N(\theta, \sigma^2) \). Consider the case where both \( \theta \) and \( \sigma^2 \) are unknown and if we replace \( \theta \) and \( \sigma^2 \) by the intuitive estimators \( \hat{\theta} \) and \( \hat{\sigma}^2 \), then \( LR(\hat{\theta}, \hat{\sigma}^2) \) is increasing in \( \frac{\hat{\theta}}{\hat{\sigma}^2} \), leading to the \( t \)-statistic. Hence in this sense, a \( t \)-statistic estimates \( \theta \) and \( \sigma^2 \) by its unbiased estimators.

If we replace \( \theta \) by \( \hat{\theta} \) and \( \sigma^2 \) by the shrinkage estimator \( \hat{\sigma}^2_{EB} \), then it leads to the statistic in Theorem 1. Hence in this sense, \( F_S \) test is based on a statistic that shrinks the variances but not the means.

In order to derive (4.6), or actually its exponential part, we may replace \( \sigma^2 \) by \( \hat{\sigma}^2_{EB} \) and choose \( \theta \) so that

\[ \frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2_{EB}}} = \frac{\hat{\theta} - \mu}{\sqrt{\hat{\sigma}^2_{EB} + \tau^2}}, \quad (6.2) \]

which would imply that LR in (6.1) becomes the second ratio involving the two exponential terms in (4.6). (Numerical studies show that the first ratio can be dropped without affecting too much of the power.) Simple algebraic calculation shows that one should take \( \theta \) to be

\[ \hat{\theta} \left( 1 - \sqrt{\frac{\hat{\sigma}^2_{EB}}{\hat{\sigma}^2_{EB} + \tau^2}} \right) + \sqrt{\frac{\hat{\sigma}^2_{EB}}{\hat{\sigma}^2_{EB} + \tau^2}} \mu. \quad (6.3) \]

Hence, other than the first ratio in (4.6), the second ratio behaves as if \( \theta \) is estimated by the above estimator which both shrinks the variance as in \( F_S \)
and shrinks $\hat{\theta}$ toward $\mu$.

Interestingly, in the typical shrinkage estimator, there is no square root. To check the effect of the square root on $F_{SS}$ test, we drop the square root in (6.3) and plug into (6.2) which results in the modified test $F_{nsr}$:

$$F_{nsr} = \left( \frac{\hat{\sigma}^2_{EB} + \tau^2}{\hat{\sigma}^2_{EB}} \right) ^{-\frac{1}{2}} \cdot \left( \frac{e^{-\frac{1}{2}M_{MV}(\hat{\theta} - \mu)^2/\hat{\sigma}^2_{EB}}}{e^{-\frac{1}{2}(\hat{\theta}^2)/\hat{\sigma}^2_{EB}}} \right)$$  \hspace{1cm} (6.4)

where ‘nsr’ stands for ‘no square root’ and $M_{MV} = \frac{\hat{\sigma}^2_{EB}}{\hat{\sigma}^2_{EB} + \tau^2}$. We also generate another modified test, $F_{2r}$, by using the second ratio in (4.6):

$$F_{2r} = \frac{e^{-\frac{1}{2}(\hat{\theta} - \mu)^2/(\hat{\sigma}^2_{EB} + \tau^2)}}{e^{-\frac{1}{2}(\hat{\theta}^2)/\hat{\sigma}^2_{EB}}},$$  \hspace{1cm} (6.5)

where ‘2r’ stands for ‘second ratio only’. Figure 2 shows that $F_{nsr}$ is slightly less powerful as $F_{SS}$ test. It also shows that the test $F_{2r}$ does not behave in power too differently from $F_{SS}$, justifying the derivation in (6.2). Hence the statistic is more subtle than just shrinking the means and the variances. Nevertheless, it does have the ingredient of shrinking both the means and the variances as suggested by (6.3).

7 COMPARISON WITH OTHER TESTS IN A MORE REALISTIC SETTING

In this section, we show in a realistic setting that the proposed $F_{SS}$ test has higher power than the tests proposed in the literature. Although we tried Storey’s statistic (Storey 2007) which may be quite powerful, we are unable to report since its intensive computation prevents us to simulate the power in a reasonable amount of time for $G = 15000$ that we consider.
Figure 2: The average power of MAP test and $F_{SS}$ test compared with modified forms of $F_{SS}$ test. Data were simulated according to the canonical form as described in Section 5. Variances $\sigma_g^2$ were randomly drawn from a data set in Cui et al. (2005) and the true mean effect $\theta_g$ were simulated from $N(\mu_\theta, \tau^2_\theta)$. Significance level was controlled at nominal 5% level of Type I error rate. The performances of MAP test and $F_{SS}$ test are compared with other tests for various coefficient of variation (CV) of variances. $F_{nssr}$ indicates the test statistic corresponding to (6.4) and $F_{2r}$ indicates the test statistic corresponding to (6.5).
Below we address two practical issues in applying $F_{SS}$ test.

First, the $F_{SS}$ test in (4.6) assumes the knowledge of $\mu$ and $\tau^2$ (mean and variance of $\theta$), which are unknown in real application. An obvious approach is to estimate $\mu$ and $\tau^2$ based on data. To do so, we assume a mixture model which has become popular recently in analyzing microarray data:

$$\theta_g \sim \begin{cases} 
N(\mu, \tau^2) & \text{with probability } \pi_1 = 1 - \pi_0 \\
0 & \text{with probability } \pi_0 
\end{cases} .$$

Assuming $\hat{\theta}_g | \theta_g, \sigma^2_g \sim N(\theta_g, \sigma^2_g)$, we have $\hat{\theta}_g | \sigma^2_g \sim \pi_1 N(\mu, \sigma^2_g + \tau^2) + \pi_0 N(0, \sigma^2_g)$. The maximum likelihood approach using the distributional assumption is not very good. Instead, in the likelihood function, one replaces $\mu$ and $\tau^2$ by functions of $\pi_1$ and the first two moments that result from solving the two equations for $\mu$ and $\tau^2$:

$$E\left[\hat{\theta}_g\right] = \pi_1 \mu,$$

$$E\left[\hat{\theta}_g^2\right] = E[\sigma^2_g] + \pi_1 \tau^2 + \pi_1 \mu^2 .$$

Also, we replace the first two moments by its sample moments. This reduces the likelihood function to a function of $\pi_1$. Maximizing it leads to an estimate of $\pi_1$. In the calculation above, all $\sigma^2_g$ and $E[\sigma^2_g]$ are replaced respectively by $\hat{\sigma}^2_{EB}$ for the corresponding gene and its average across all genes. Although, we are not so interested in estimating $\pi_1$, its estimate can be substituted into the expressions above to arrive at estimates of $\mu$ and $\tau^2$, which can then be substituted into the $F_{SS}$ statistic.

The other practical problem is that the test statistics $F_3, F_S$ and $F_{SS}$ are not standard $F$ statistics. Consequently, their distributions can not be obtained by analytic calculation. The same as in Cui et al. (2005), we approximate the null distributions for all $F$-like statistics by permutation analysis.
We also use permutation to get the null distribution for $F_1$ statistic because distributional assumptions are sometime questionable for microarray data and it is fair to establish all critical values by permutation. The two modifications depicted above are applied to the $F_{SS}$ test in the numerical studies reported in Figures 3 and 5.

Permutation analysis is briefly reviewed in Cui and Churchill (2003). It is a nonparametric approach to establish the null distribution of a test statistic. We apply the permutation test with two treatment groups (2-sample tests) as described in Cui et al. (2005) and $p$-values are calculated according to the approximated null distribution. Then the average power is estimated by taking the proportion of differentially expressed genes that are found significant at the nominal type I error rate of 5%. The results are shown in Figure 3. For moderated $t$-test, we directly use the p-values generated by the Limma package which is developed by Dr. Smyth and downloaded from www.r-project.org.

In Figures 3, it is demonstrated that $F_{SS}$ test is more powerful than $B$-statistic of Lönnstedt and Speed (2002), $F_S$ test, the test of Wright and Simon (2003), moderated $t$-test (Smyth 2004) and SAM (Tusher et al. 2001). The last four tests shrink only the variances or standard errors. In particular, the tests of Wright and Simon, and moderated $t$-test are derived based on a prior on $\sigma^2_\theta$ only, which amounts to shrinking the variance. The numerical studies show that the power of these tests are similar to $F_S$, which seems reasonable since they all shrink the variances or standard errors only.

The $B$ statistic of Lönnstedt and Speed (2002) shrinks both the mean and the variance. However it shrinks the mean toward zero. Hence when $\mu_\theta$ is
Figure 3: Average power comparison of $F_{SS}$ test with other proposed tests. For simulation for all three plots, the total number of genes, $G$, is 15K with 10% being differentially expressed for each of 10 data sets. Power were averaged over 10 simulated data sets. Gene expression data were simulated as described in Section 5. Hyper-parameters for $F_{SS}$ were estimated as described in the text. Permutation was used to get the null distribution of test statistics. Significance level was controlled at nominal 5% level of Type I error rate. (a) The $F_{SS}$ test is compared with $F_1$, $F_3$ and $F_S$. (b) The $F_{SS}$ test is compared with SAM, moderated $t$-test (Limma) and RVM test of Wright and Simon (2003). (c) Paired data were simulated in order to compare with $B$-statistic. The performance of $F_{SS}$ test is compared with $B$ and other $F$-like tests.
not zero, $F_{SS}$ is more powerful. In the simulation setup of their paper when $\mu_\theta = 0$, our test is only slightly more powerful. In calculating the $B$ statistic, we use the Limma package of Smyth to estimate the hyperparameters. This was suggested to us by Professor Terry Speed who considers it to be better than what was originally proposed in their paper. The modification however does not make a difference in our tested cases.

8 EXTENSIONS TO MULTIPLE REGRESSION

In the previous sections about $F_{SS}$ test, the case of testing a single parameter or a single contrast was considered. In this section, we extend the MAP test and $F_{SS}$ test to the case of testing multiple linear contrasts of parameters.

We look at the model (8.1) with the parameter $\beta$ being a $p \times 1$ vector,

$$Y = X\beta + \epsilon. \quad (8.1)$$

In microarray context, $Y$ is the $n$-dimensional vector of observed gene expression levels, which are usually log-ratios for two-color microarray data or log-intensities for single channel data, properly normalized. The matrix $X$ is the design matrix for the fixed effects $\beta$.

The estimated parameter $\hat{\beta}$ is assumed to follow $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. To derive a procedure easy to compute, we assume that $\beta \sim N(\mu_\theta, \tau^2(X'X)^{-1})$. A common interesting case is to test $H_0 : A\beta = \eta$, where $A$ is a full-rank $k \times p$ matrix with $k \leq p$. If we define $\theta = A\beta - \eta$, the null hypothesis is
equivalent to $H_0: \theta = 0$. Let $\hat{\theta} = A\beta - \eta$. Then

$$\hat{\theta} | \theta \sim N(\theta, \sigma^2 A(X'X)^{-1}A') \quad \text{and} \quad \theta \sim N(\mu, \tau^2 A(X'X)^{-1}A'),$$

where $\mu = A\mu_0 - \eta$. Integrating out $\theta$, the marginal distribution of $\hat{\theta}$ is $N(\mu, (\sigma^2 + \tau^2) A(X'X)^{-1}A')$. Hence as in Section 4, the statistic of the MAP test is

$$E \left[ \frac{(\sigma^2 + \tau^2)^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (\hat{\theta} - \mu)'(A(X'X)^{-1}A')^{-1}(\hat{\theta} - \mu)/(\sigma^2 + \tau^2) \right)}{\exp \left( -\frac{1}{2} \hat{\theta}'(A(X'X)^{-1}A')^{-1}\hat{\theta}/\sigma^2 \right)} \right],$$

(8.2)

where the expectation is with respect to the conditional distribution of $\sigma^2$ given $\hat{\sigma}^2 \equiv |Y - X\hat{\beta}|^2/(n - p)$. To avoid the integration, we use the first order approximation of $\sigma^2$ by replacing it with $\hat{\sigma}^2_p$ in (3.7) and in turn by $\hat{\sigma}^2_{EB}$ defined in (3.9) by the empirical Bayes approach. This results in the statistic

$$\left( \frac{\hat{\sigma}^2_{EB} + \tau^2}{\hat{\sigma}^2_{EB}} \right)^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (\hat{\theta} - \mu)'(A(X'X)^{-1}A')^{-1}(\hat{\theta} - \mu)/(\hat{\sigma}^2_{EB} + \tau^2) \right) \exp \left( -\frac{1}{2} \hat{\theta}'(A(X'X)^{-1}A')^{-1}\hat{\theta}/\hat{\sigma}^2_{EB} \right).$$

(8.3)

One can then use the approach in Section 7 to estimate $\pi_1$ and $\mu$. However in doing so, we may focus on one element of $\mu$ at a time to simplify computation.

Naturally, the moment calculation involves $X$. After substituting $\mu$ by its estimate, the corresponding test is denoted as $F_{MSS}$, where the subscript MSS stands for double shrinkage in multiple regression.

To compare the tests based on the four $F$ statistics, $F_1, F_3,$ and $F_{MSS}$ involving multiple parameters, we performed a simulation study similar to Figure 1 in Section 5. We simulate sufficient statistics based on the model $y_{g,t} = \theta_{g,t} + \epsilon_{g,t}$ for treatment $t, t = 1, 2, ..., 5$ and gene $g$ where $g = 1, 2,$
..., $G$. Here $\theta_g$ is a five-dimensional random vector in the simulation. As in Section 5, the residual variance $\sigma_g^2$ are drawn from the tumor data set of Cui et al. (2005) and $CV$ of the variances are similarly modified.

We simulate $\hat{\theta}_{g,t}$ by $N(\theta_{g,t}, \sigma_g^2)$ where the relative expression level $\theta_g$ is equal to zero for non-differentially expressed genes and follows $N(a \mu_\theta, \tau^2 I)$ for differentially expressed genes. For Figure 4, $\mu_\theta = (-0.5, -0.25, 0.25, 0.5, 0)'$ and $a$ is a scalar to tune the magnitude of the mean effects and are shown as the $X$-axis in all sub-plots. The null distribution for all $F$ tests are constructed by setting $\theta = 0$ and the critical values are determined by using the 95\% quantiles of the corresponding null distributions. Then the average power are calculated by taking the proportion of differentially expressed genes that are found significant.

In Figure 4, $F_{MSS}$, the analog of $F_{SS}$ test, is shown to have power substantially larger than any other test including $F_S$ test and $F_1$ test, with a larger improvement than that of $F_{SS}$ over $F_S$.

9 EQUIVALENCE OF CRITERIA

It is important to relate the work to the false discovery rate (FDR) control. We would focus on the setting that leads to the $F_{SS}$ test. Following the notations in Storey (2007), the expected number of true positives (ETP) is (4.2) times $G_1$. Similarly, the expected number of false positives (EFP) is the left hand side of (4.3) times $G_0$. One major difference between our approach and his approach is that he considered the unweighted version whereas our weights are the p.d.f. of $(\theta_g, \sigma_g^2)$. Hence in this paper as well as in his paper,
Figure 4: The average power of $F_{MSS}$ test compared with other $F$-like tests. Data were simulated according to the description in Section 8. Variances $\sigma_g^2$ were randomly drawn from a data set in Cui et al. (2005) and the true mean effect $\theta$ were simulated from $N(a\mu, \tau^2 I)$. The x-axis in the plots indicates the magnitude of $a$, the scalar of means. Significance level was controlled at nominal 5% level of Type I error rate. The performances of $F_{MSS}$ test are compared with $F_1$, $F_3$ and $F_S$ for various coefficient of variation (CV) of variances.
the aim is to find test that maximizes ETP given that EFP is controlled to be no more than \( \alpha \). This is referred to as Criterion I below.

As in Storey (2007), it can be shown

\[
FDR \simeq \frac{EFP}{EFP + ETP}
\]  

(9.1)

where “\( \simeq \)” denotes equal either asymptotically as \( G \to \infty \) or exactly in some exchangeable settings. The FDR in (9.1) could be interpreted as the false discovery rate defined in Benjamini and Hochberg (1995) or the pFDR defined in Storey (2002). Because of this, Storey (2007) argued forcefully and convincingly that EFP and ETP are more fundamental than FDR.

In the following discussion, we shall, as in Storey (2007), ignore the difference between the right hand side and the left hand side of (9.1). Also, define the missed discovery rate as in Storey (2007):

\[
MDR = \frac{EFN}{EFN + ETN}
\]  

(9.2)

where EFN and ETN are expected values of FN and TN respectively. Here FN (or TN) denotes the number of false negatives (or true negatives) as in Table 1.

Table 1: Outcomes when testing G hypotheses. The expected number of outcomes for results of hypothetical Test 1 and Test 2 are listed to the right of the possible outcomes.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Accepted</th>
<th>Test 1 (Test 2)</th>
<th>Rejected</th>
<th>Test 1 (Test 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Hypothesis</td>
<td>( TN )</td>
<td>0 (1K)</td>
<td>( FP )</td>
<td>2K (1K)</td>
</tr>
<tr>
<td>False Hypothesis</td>
<td>( FN )</td>
<td>1K (10K)</td>
<td>( TP )</td>
<td>18K (9K)</td>
</tr>
</tbody>
</table>

25
Storey’s Lemma 2 (2005) claims that Criterion I is equivalent to minimizing MDR for each fixed FDR, which we call MDR criterion. The result is quite interesting. We, however, consider that a criterion better than MDR Criterion is to minimize $\text{EFN}/G_1$ among tests that control FDR (Criterion II). Criterion II is the criterion used in Table 5 and Figure 7B of Cui et al. (2005). Table 1 reports the expected values of TN, FN etc of two hypothetical tests, Test 1 and Test 2, where K represents thousand. For example, for Test 1, ETN=0 and ETP=18K. For Test 2, ETN=1K and ETP=9K.

These two tests both have FDR=10% = 2K/(2K+18K) = 1K/(1K+9K). Intuitively, it seems that Test 1 is more powerful because among 19K alternative hypotheses, it identifies 18K true positives. In contrast, Test 2 only identifies, among 19K alternatives, 9K true positives. However, MDR is 100% for Test 1 and is $10/11 \approx 90\%$ for Test 2. The MDR Criterion would conclude that Test 1 is inferior. According to Criterion II, Test 1 is better since its $\text{EFN}/G_1$ is smaller. This agrees with the intuition.

One major reason that the MDR of Test 2 is smaller is due to its ETN being larger. However, we argue that ETN is a quantity related to the true null and should not be used to measure the power of the test.

The following theorem relating Criterion I and Criterion II is precisely stated and proved in the Appendix B. Storey (2007) in Lemma 5 states assumptions (basically exchangeability of distributions of genes) applicable to microarray experiments under which one may assume without loss of generality that the optimal rejection region is the same for each $g$ (or each gene). This would be assumed in Theorem 2 below.

**Theorem 2.** The optimal solution to Criterion I gives the optimal solution
Figure 5: Comparison of FDR and FN for $F_{SS}$ test with other $F$-like tests. The average of FDR and FN over results of 10 simulated data sets that generate Figures 3 are plotted for $F_{SS}$, $F_1$, $F_3$ and $F_S$ and Efron’s $t$ in panel (a) and (b). Still, significance level was controlled at nominal 5% level of Type I Error.

to Criterion II.

The proof in the Appendix B gives a possible constructive solution by solving (B.2), where $G_0$ can be replaced by the fraction $\pi_0$ of hypotheses which are null and $G_1$ by $\pi_1 = 1 - \pi_0$. For t-tests, Liu and Hwang (2007) show that this has a unique solution when exists. However, the tests considered in this paper are more complicated. Obviously in a real application, $\pi_0$ should be estimated by, for example, the method in Section 7.

Figure 5 shows that, in agreement of Theorem 2, $F_{SS}$ test minimizes both FDR and EFN when compared to other tests.
10 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we derive $F_S$ statistic using sound statistical principles. Similar principles were used to derive more powerful test $F_{SS}$ that shrinks both means and variances while $F_S$ statistic shrinks only the variances. The statistic $F_{SS}$ is more powerful compared to all other statistics considered here and has an explicit form, hence is computationally very fast.

We also found that $F_{SS}$ has smallest FDR and smallest false negatives among the test statistic. A future important research project is to provide a method to control FDR. Preliminary numerical studies in Cui et al. (2005) show that permutation procedure does the job reasonably well for $F_S$. We expect similar result for $F_{SS}$.

ACKNOWLEDGEMENT

We are thankful to Professors Xiangqin Cui and Jing Qiu who provided the residual variance data and simulation code in Cui et al. (2005). We would like to thank Professors Terry Speed and Chong Wang for insightful discussion of this work.

APPENDICES

Appendix A: Proof of Theorem 1

As in Section 3, we work with the model $\hat{\sigma}^2 = \sigma^2 K$ and use the notation of $\hat{\rho} = \ln \hat{\sigma}^2 - \mu_K$ and $\rho = \ln \sigma^2$. We first focus on the numerator of (3.5),
without taking supremum over $\theta$,
\[
\int f(\hat{\theta} \mid \theta, e^\rho) f(\hat{\rho} \mid \rho) \pi(\rho)d\rho. \tag{A.1}
\]

We make the same notational assumption as those stated right above Theorem 1. Hence $\hat{\rho}$ is normally distributed with mean $\rho$ and $\rho$ has a Bayes normal prior. A classical Bayesian calculation leads to
\[
\rho \mid \hat{\rho} \sim N(\hat{\rho}_p, M_V \sigma_K^2) \quad \text{and} \quad \hat{\rho} \sim N(\mu_V, \sigma_K^2 + \tau^2), \tag{A.2}
\]
where $\hat{\rho}_p$ and $M_V$ are the same as in (3.7). Since $f(\hat{\rho} \mid \rho)\pi(\rho) = \pi(\rho \mid \hat{\rho})f(\hat{\rho})$,
(A.1) equals
\[
f(\hat{\rho})E \left[ e^{-\frac{1}{2}(\hat{\theta} - \theta)^2 / e^\rho} \frac{1}{\sqrt{2\pi e^\rho}} \mid \hat{\rho} \right], \tag{A.3}
\]
where $f(\hat{\rho})$ is the density of $\hat{\rho}$ according to the second part of (A.2) and the expectation is taken with respect to the first part of (A.2). Taking the supremum of (A.3) over $\theta$ leads to the substitution of $\theta$ by $\hat{\theta}$. Furthermore setting $\theta = 0$ in (A.3) gives the denominator of (3.5). Canceling out $f(\hat{\sigma})$ and some constants demonstrates the statistic in (3.5) is equal to
\[
E \left[ e^{-\rho^2 / \hat{\rho}^2} \right] / E \left[ e^{-\frac{1}{2}\hat{\theta}^2 / e^\rho} e^{-\frac{1}{2}\rho} \mid \hat{\rho} \right]. \tag{A.4}
\]
From (A.2), we note that the conditional distribution of $\rho$ given $\hat{\rho}$ has the same distribution as $\hat{\rho}_p + \sigma_K \sqrt{M_VZ}$, where $Z$ is the standard normal random variable. Substituting this expression in the numerator and denominator of (A.4) and canceling out the term relating to $\hat{\rho}_p$ shows that (A.4) equals
\[
E \left[ e^{-\frac{1}{2}\sigma_K \sqrt{M_VZ}} \right] / E \left[ e^{-\frac{1}{2}\hat{\theta}^2 / (e^{\sigma_K \sqrt{M_VZ} - e^{\hat{\rho}_p}}) e^{-\frac{1}{2}\sigma_K \sqrt{M_VZ}}} \right]. \tag{A.5}
\]
Note that the numerator has nothing to do with the statistic $\hat{\theta}$ and $\hat{\rho}$. On the other hand, the denominator equals
\[
E \left\{ \left[ \exp \left( -\frac{1}{2} (\hat{\theta}^2 / e^{\hat{\rho}_p}) \right) \right] e^{-\sigma_K \sqrt{M_VZ}} e^{-\frac{1}{2}\sigma_K \sqrt{M_VZ}} \right\}. \tag{A.6}
\]
Since $M_V$ and $\sigma_K$ are all constants, the main focus is on statistics $\hat{\theta}$ and $\hat{\rho}_p$. It is obvious then that (A.6) decreases according to $\hat{\theta}^2/e^{\hat{\rho}_p} = \hat{\theta}^2/\hat{\sigma}_p^2$. Hence (3.5) is equivalent to the assertion that $\hat{\theta}^2/\hat{\sigma}_p^2$ is large.

**Appendix B: Proof of Theorem 2**

Theorem 2 is now precisely stated:

*Theorem 2.* Let the FDR of the Neyman-Pearson test (which is optimal according to Criterion I) $C_\lambda = \{x : f_1(x) \geq \lambda f_0(x)\}$ be $f$ and assume that $f < 1$. Then among all tests that have $\text{FDR} \leq f$, $C_\lambda$ minimizes $\text{EFN}$.

*Remark 1.* The following theorem aims at the setting of Section 4. However, it applies to a general problem of testing $H_0 : f(x) = f_0(x)$ v.s. $H_1 : f(x) = f_1(x)$. Here $f_0$ and $f_1$ are assumed to be probability density functions (p.d.f.) with respect to the Lebesgue measure. However the same theory holds with other measures including the counting measure corresponding to the discrete case. When applying to Section 4, we take $f_0$ and $f_1$ to be the marginal p.d.f. of $x = (\hat{\theta}_g, \hat{\sigma}_g)$, namely the p.d.f. of $(\hat{\theta}_g, \hat{\sigma}_g)$ after integrating out the prior distribution of $(\theta_g, \sigma_g)$. Hence $x$ in general is a vector.

*Remark 2.* Note that for a critical region $C$

\[ \text{FDR} \equiv \text{FDR}(C) = G_0 A(C)/(G_0 A(C) + G_1 B(C)), \]

where $A(C) = \int_C f_0(x)dx$ and $B(C) = \int_C f_1(x)dx$ and $G_0$ and $G_1$ are assumed to be positive. Simple algebraic calculation shows that

\[ \text{FDR}(C) \leq f \Leftrightarrow A(C) - B(C)G_1f/[G_0(1-f)] \leq 0. \quad (B.1) \]

In particular, since $\text{FDR}(C_\lambda) = f$, (B.1) implies

\[ A(C_\lambda) = B(C_\lambda)G_1f/[G_0(1-f)]. \quad (B.2) \]
To minimize $E\!F\!N/G_1 = 1 - B(C)$, under the constraint (B.1), it would be convenient to study how to choose $C$ to minimize

$$A(C) - B(C)G_1 f/[G_0(1 - f)] + k[(1 - B(C)].$$

The following Lemma provides the solution. Below we use $A$ and $B$ to denote $A(C)$ and $B(C)$.

**Lemma 1.** One rejection region that minimizes (B.3) is

$$\{x : [G_1 f/(G_0(1 - f)] + k f_1(x) \geq f_0(x)\}.$$  \hspace{1cm} (B.4)

**Proof of Lemma 1** Since in (B.2), $A$ and $B$ are the only quantities depending on $C$, minimizing (B.3) is equivalent to minimizing

$$A - \left[\frac{G_1 f}{G_0(1 - f)} + k\right] B = \int_C f_0(x) - \left[\frac{G_1 f}{G_0(1 - f)} + k\right] f_1(x)dx.$$  

Hence (B.4) obviously minimizes the above expression and hence (B.3) establishes the lemma.

**Proof of Theorem 2.** If $\lambda \leq 0$, then $C_\lambda$ is the whole Euclidian space. The corresponding $E\!F\!N/G_1$ is zero and is minimized. Hence we assume $\lambda > 0$ below. Choose $k$ so that

$$G_1 f/[G_0(1 - f)] + k = \lambda^{-1}.$$  \hspace{1cm} (B.5)

Below we argue that $k \geq 0$. Note

$$A(C_\lambda) = \int_{C_\lambda} f_0(x)dx \leq \frac{1}{\lambda} \int_{C_\lambda} f_1(x)dx = \frac{1}{\lambda} B(C_\lambda).$$

Hence $A(C_\lambda) - \frac{1}{\lambda} B(C_\lambda) \leq 0$, implying that

$$A(C_\lambda) - \frac{G_1 f}{G_0(1 - f)} B(C_\lambda) - kB(C_\lambda) \leq 0.$$
The first two terms cancel by (B.2). Further, the assumption \( f < 1 \) implies that \( B(C_\lambda) > 0 \), and consequently \( k \geq 0 \).

Now we show that if \( k > 0 \), then \( C_\lambda \) minimizes EFN among all \( C \) that satisfy (B.1). By Lemma 1,

\[
A(C_\lambda) - \frac{G_1 f}{G_0 (1 - f)} B(C_\lambda) - kB(C_\lambda) \leq A(C) - \frac{G_1 f}{G_0 (1 - f)} B(C) - kB(C).
\]

Applying (B.4) to the left-hand side of the above equation and canceling out the first two terms establish that

\[
-kB(C_\lambda) \leq A(C) - \frac{G_1 f}{G_0 (1 - f)} B(C) - kB(C) \leq -kB(C),
\]

where the last inequality follows from (B.1). Hence if \( k > 0 \), \( B(C_\lambda) \geq B(C) \), or equivalently \( C_\lambda \) minimizes EFN.

The proof would be complete if we could show that \( C_\lambda \) minimizes EFN even when \( k = 0 \). This step is proved in the next two lemmas.

**Lemma 2.** If \( k = 0 \), then \( f_1(x) = \lambda f_0(x) \) for almost all \( x \in C_\lambda \).

**Proof of Lemma 2.** Note that

\[
A(C_\lambda) = \int_{C_\lambda} f_0(x) dx \leq \int_{C_\lambda} \frac{1}{\lambda} f_1(x) dx = \frac{1}{\lambda} B(C_\lambda).
\]

However (B.2) and (B.5) with \( k = 0 \) assert that the above inequality is an equality. Hence

\[
\int_{C_\lambda} \left[ f_0(x) - \frac{1}{\lambda} f_1(x) \right] dx = 0.
\]

Since on \( C_\lambda \), \( f_0(x) - \frac{1}{\lambda} f_1(x) \leq 0 \), we conclude that \( f_0(x) - \frac{1}{\lambda} f_1(x) = 0 \) on \( C_\lambda \), establishing the lemma.

**Lemma 3.** If \( k = 0 \), then \( C_\lambda \) minimizes EFN among all \( C \) satisfying (B.1).
Proof of Lemma 3 From (B.1),

\[ 0 \geq A(C) - \frac{G_1 f}{G_0 (1 - f)} B(C) = A(C) - \lambda^{-1} B(C). \]  

(B.6)

The above equation follows from \( k = 0 \) and (B.5). Now the right hand side equals

\[ \int_C \left[ f_0(x) - \frac{1}{\lambda} f_1(x) \right] dx = \int_{C \cap C'_{\lambda}} \left[ f_0(x) - \frac{1}{\lambda} f_1(x) \right] dx, \]  

(B.7)

where \( C'_{\lambda} \) is the complement of \( C_{\lambda} \) and the last equation holds because of Lemma 2. On \( C'_{\lambda} \), \( f_0(x) - \lambda^{-1} f_1(x) > 0 \), we now conclude that the Lebesgue measure of \( C \cap C'_{\lambda} \) is zero. Otherwise the right hand side of (B.7) would be positive, contradicting (B.6). Now almost surely, \( C \) is included in \( C_{\lambda} \). The one maximizes \( B(C) \) under the constraint (B.1) is the largest set \( C_{\lambda} \) which satisfies (B.1) by (B.2).

References


