Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 913560, 14 pages
doi:10.1155/2012/913560

Research Article

Robust Semiparametric Optimal Testing Procedure for Multiple Normal Means

Peng Liu[1] and Chong Wang[1,2]

[1] *Department of Statistics, Iowa State University, Ames, IA 50011, USA*
[2] *Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University, Ames, IA 50011, USA*

Correspondence should be addressed to Peng Liu, pliu@iastate.edu

Received 27 March 2012; Accepted 10 May 2012

Academic Editor: Yongzhao Shao

In high-dimensional gene expression experiments such as microarray and RNA-seq experiments, the number of measured variables is huge while the number of replicates is small. As a consequence, hypothesis testing is challenging because the power of tests can be very low after controlling multiple testing error. Optimal testing procedures with high average power while controlling false discovery rate are preferred. Many methods were constructed to achieve high power through borrowing information across genes. Some of these methods can be shown to achieve the optimal average power across genes, but only under a normal assumption of alternative means. However, the assumption of a normal distribution is likely violated in practice. In this paper, we propose a novel semiparametric optimal testing (SPOT) procedure for high-dimensional data with small sample size. Our procedure is more robust because it does not depend on any parametric assumption for the alternative means. We show that the proposed test achieves the maximum average power asymptotically as the number of tests goes to infinity. Both simulation study and the analysis of a real microarray data with spike-in probes show that the proposed SPOT procedure performs better when compared to other popularly applied procedures.

1. Introduction

The problem of statistically testing mean difference for each of thousands of variables is commonly encountered in genomic studies. For example, the popularly applied microarray technology allows the gene expression study of tens of thousands of genes simultaneously. The recent advance of next-generation sequencing technology allows the measurement of gene expression in an even higher dimension. These high-throughput technologies have revolutionized the way genomic studies progress and provided rich data to explore. However, these experiments are expensive, and as a consequence, such experiments typically involve only a few samples for each treatment group. This results in the "large p, small n"

problem for hypothesis testing, and the power of the statistical tests can be very low after controlling the multiple testing error, such as the false discovery rate (FDR).

The normalized signal intensities from microarray experiments are generally assumed to follow normal distributions [1–4]. The recently emerging next-generation sequencing data may also be modeled approximately using normal distributions, when the number of reads are large or under certain transformation [5]. Thus multiple testing problem for normal means has wide applications in genetic and genomic studies, and it is also a general statistical question of interest.

Several testing procedures have been proposed in the context of microarray study, including the SAM test [6], Efron's t-test [7], the regularized t-test [8], the B-statistic [1] and its multivariate counterpart, the MB-statistic [9], the test of Wright and Simon [10], the moderated t-test [2], the F_S test [3] and the test of [11] which is similar to the F_S test, the F_{SS} test [4], and the LEMMA test [12]. Although numerous procedures have been proposed, very few can be justified to achieve the optimal power. Among these procedures, Hwang and Liu [4] proposed a framework and showed that an optimal testing procedure can be derived within such a framework. They also proposed a test with maximum average power (the MAP test) and an approximated version, the F_{SS} test. Here the optimality was defined in terms of maximizing the power averaged across all tests for which the null hypotheses are false while controlling FDR. This method provides theoretical guide for developing optimal multiple testing procedures. The popularly applied moderated t-statistic developed by Smyth [2] can also be shown to achieve optimal power asymptotically under different distributional assumptions from the F_{SS} test. Both the moderated t-statistic and the F_{SS} test assume that the mean expression levels (or the mean of interesting contrasts) of all genes follow a normal distribution although the parameters for this distribution vary between the two tests. Yet in practice such distribution depends on the population of genes selected in a particular study and often does not follow the prespecified parametric distribution. This raises concerns about the robustness of the moderated t and the F_{SS} tests.

The objective of this paper is to develop an optimal and robust multiple testing procedure without any distributional assumptions on the mean. As in Hwang and Liu [4], the optimality is defined in terms of maximizing the power averaged across all tests for which the null hypotheses are false while controlling FDR. We develop a semiparametric optimal testing procedure which we abbreviate as the SPOT procedure. The distribution of the mean expression across genes is not assumed to follow a parametric model which makes our method robust to violations to normal assumptions. We find that the SPOT procedure works very well in simulation studies and in an analysis of real microarray data with spike-in probes.

The remaining of this paper is organized as follows. We first introduce necessary notations in Section 2. Then, in Section 3, we describe the general concepts of optimal testing procedures. We propose our semiparametric optimal testing (SPOT) procedure in Section 4 and describe its implementation in Section 5. Section 6 presents simulation studies. Section 7 shows the analysis result of a real microarray dataset. Section 8 provides a summary of this paper.

2. Notations

An appropriate linear model is typically fitted for each gene based on the design of a microarray experiment. Section 2 of Smyth [2] provides a nice description of this topic.

Given the linear model, suppose that we have an interesting contrast to test for each gene. This contrast may be the difference between the means of two treatment groups or linear combination of means from several treatment groups. For the simplicity of description, we call the genes whose contrast means are not zero as the differentially expressed (DE) genes and the genes whose contrast means equal to zero as equivalently expressed (EE) genes. After fitting the linear model for each gene, we obtain an estimate for the contrast for each gene, X_g. In addition, we get the estimate of the sample residual variance, s_g^2, for each gene. For each $g = 1, \ldots, G$, X_g and s_g^2 are related to true parameters, μ_g and σ_g^2, by $X_g \mid \mu_g, \sigma_g^2 \sim N(\mu_g, v_g\sigma_g^2)$ and $s_g^2 \mid \sigma_g^2 \sim (\sigma_g^2/d_g)\chi_{d_g}^2$, where μ_g is the contrast mean for gene g, σ_g^2 is the true residual variance for gene g, and the coefficients v_g and d_g are determined by the design of the experiment. Two examples are given as follows.

Example 2.1. Two-channel microarray experiment to compare two treatments. Assume that each sample from treatment A is paired randomly with a sample from treatment B and each pair of samples is cohybridized onto one slide. After normalization and appropriate transformation, the difference of normalized expression measurements between the two samples on each slide is analyzed for each gene. Hence, this is a paired sample case and the number of data points for each gene is n, the number of slides. We are interested in identifying DE genes. In this case, X_g is the mean difference of the paired samples for gene g. s_g^2 is the sample variance for gene g. So $v_g = 1/n$ and $d_g = n - 1$.

Example 2.2. Affymetrix microarray experiment with two independent samples. Assume sample sizes are n_1 and n_2 for treatment A and treatment B, respectively. The statistic X_g is the difference in sample means of normalized expression measurements between two groups for gene g. s_g^2 is the pooled sample variance. Then $v_g = 1/n_1 + 1/n_2$ and $d_g = n_1 + n_2 - 2$.

Given the data X_g and s_g^2, an ordinary t-test with statistic $t_g = X_g/\sqrt{v_g}s_g$ may be used to test the null hypothesis H_g^0: $\mu_g = 0$. However, the power of such tests is low after controlling multiple testing error. So statistical methods with higher power are in demand for such high-dimensional testing problem as encountered in gene expression studies.

3. Optimal Testing Procedures

In the analysis of high-dimensional gene expression data such as microarray data, we are more interested in the average behavior of the tests across all genes rather than the performance of an individual test. Because the dimension of tests is huge, multiple testing errors should be controlled to avoid too many type I errors. Controlling FDR is an important method for controlling multiple testing errors and is widely used for genomic studies. Although many testing procedures have been developed as reviewed in Section 1, the paper by Hwang and Liu [4] provides some theoretical guide on how to derive optimal testing procedures within an empirical Bayes framework. The optimal tests are defined to be the ones that maximize the power averaged across all genes for which the null hypotheses are false while controlling FDR. Such optimal tests have been called MAP tests, where MAP stands for maximum average power [4].

In a Bayesian framework, we assume model parameters like μ_g and σ_g^2 follow some distributions. The residual variances of genes, σ_g^2, have been modeled by prior distribution like inverse gamma [2, 10] or log-normal [4] distribution independent of whether the null hypothesis is true or false. For EE genes, the mean of contrast X_g, μ_g, is equal to 0. For DE

genes, the mean μ_g is not 0 almost surely. Denote the alternative distribution of μ_g by $\pi_1(\cdot)$. Based on the Neyman-Pearson fundamental lemma, for a randomly selected gene g, the most powerful test statistic for testing $H_g^0 : \mu_g = 0$ versus $H_g^1 : \mu_g \sim \pi_1(\mu_g)$ is given by

$$T_g^{\mathrm{NP}} = \frac{\int\int f\left(X_g, s_g^2 \mid \mu_g, \sigma_g^2\right)\pi_1(\mu_g)\pi\left(\sigma_g^2\right)d\mu_g d\sigma_g^2}{\int f(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2)\pi(\sigma_g^2)d\sigma_g^2}, \tag{3.1}$$

where $\pi(\cdot)$ denote the prior distributions of σ_g^2. And the test rejects the null hypothesis H_g^0 when T_g^{NP} is large. The simultaneous testing procedure where all genes are tested using the most powerful statistics T_g^{NP}, $g = 1, 2, \ldots, G$, achieves the highest average power while controlling FDR, as proved in Hwang and Liu [4].

One popular multiple-testing method for microarray data is the moderated t-test proposed by Smyth [2]. Smyth proposed to model the residual variance σ_g^2 with the prior distribution:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2, \tag{3.2}$$

where $\chi_{d_0}^2$ denotes a chi-square distribution with degrees of freedom d_0 and s_0^2 is another hyperparameter. This prior distribution is equivalent to an inverse-gamma distribution and has been shown to fit real data well. Compared to a standard t-test statistic $t_g = x_g/\sqrt{v_g}s_g$, Smyth's moderated t-statistic takes the form of

$$\tilde{t}_g = \frac{x_g}{\sqrt{v_g}\tilde{s}_g}, \tag{3.3}$$

where

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0} \tag{3.4}$$

is a shrinkage estimator of σ_g^2 by shrinking s_g^2 toward s_0^2.

In practice, the unknown hyperparameters d_0 and s_0^2 for the distribution of the variance σ_g^2 can be estimated consistently by the method of moments, that is, equating the empirical and expected first two moments of $\log s_g^2$ [2]. Smyth [2] showed that the moderated t-test is equivalent to the B statistic proposed in Lönnstedt and Speed [1] which was derived as the posterior odds under the assumption that the distribution of μ_g under the alternative hypothesis follows $N(0, v_0\sigma_g^2)$, where v_0 is a constant. In fact, we can prove the claim that the moderated t-test achieves the optimal average power asymptotically under their assumptions for μ_g and σ_g^2. The proof is in the appendix.

However, note that the assumption that μ_g of DE genes follows a normal distribution with mean zero is restrictive. It is likely that, for example, there are more upregulated genes than downregulated genes for some studies which suggests that the mean of μ_g should be

positive. Hwang and Liu [4] have proposed a more general normal prior distribution of μ_g for DE genes:

$$\pi_1(\mu_g) \sim N\left(\theta, \tau_g^2\right), \tag{3.5}$$

where the mean of this distribution is not necessarily zero but to be estimated based on data. In addition, the variance for this distribution does not depend on the residual variance. Under this model, they have derived an optimal test and an approximated version of the test statistic (F_{SS} test) that is computationally faster. The F_{SS} statistic shrinks both the estimate of mean μ_g and the estimate of variance σ_g^2.

Both the moderated t-test and the F_{SS} test have been shown to achieve optimal power asymptotically under the assumption of normal distribution for the alternative means. Simulation studies also confirm that the power of the tests is superior under the model assumptions. However, a single normal distribution assumption on μ_g for DE genes may not be appropriate for all cases and the distribution of $\pi_1(\mu_g)$ may consist of a mixture of different subgroup distributions, for example, a mixture of two normal distributions with one having a negative mean and the other having a positive mean. If the parametric distributional assumptions of $\pi_1(\mu_g)$ are violated, the power of an optimal test built under those assumptions will suffer.

4. Semiparametric Optimal Testing (SPOT) Procedure

To obtain a more robust procedure, we propose to model the distribution of the mean μ_g nonparametrically while still deriving the optimal procedure. For the variance σ_g^2, the inverse gamma distributional assumption is reasonable and works well in practice, so we still keep this assumption. Hence, we will derive a semiparametric optimal testing procedure that we call the SPOT procedure.

Note that the numerator and denominator of the most powerful test statistic (3.1) are the joint marginal distributions of (X_g, s_g^2), under the alternative and null hypothesis, respectively. By denoting the marginal distributions by

$$\begin{aligned} m_1\left(X_g, s_g^2\right) &= \iint f\left(X_g, s_g^2 \mid \mu_g, \sigma_g^2\right)\pi_1(\mu_g)\pi\left(\sigma_g^2\right)d\mu_g d\sigma_g^2, \\ m_0\left(X_g, s_g^2\right) &= \int f\left(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2\right)\pi\left(\sigma_g^2\right)d\sigma_g^2, \end{aligned} \tag{4.1}$$

statistic (3.1) becomes

$$T_g^{\mathrm{NP}} = \frac{m_1\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)}. \tag{4.2}$$

The null marginal distribution $m_0(X_g, s_g^2)$ only involves integration with respect to variance σ_g^2. With consistent estimators of hyperparameters as proposed in Smyth [2], we can estimate $m_0(X_g, s_g^2)$ consistently. For the alternative marginal distribution $m_1(X_g, s_g^2)$, it is hard to find

a consistent estimator without any distributional assumption on μ_g. If we were to know which genes are DE, then we could estimate $m_1(X_g, s_g^2)$ nonparametrically with observed values of (X_g, s_g^2) from the DE gene population. Many nonparametric density estimators are consistent, for example, the histogram estimators and the kernel density estimators with proper choices of bandwidths [13]. However, the knowledge of differential expression is the research question of the study and of course is not available for all genes. Considering all genes without separating those that are differentially expressed from those that are not, we have a mixture distribution of differentially expressed and nondifferentially expressed genes. The mixture density of the marginal distributions, denoted by $m_m(X_g, s_g^2)$, can be estimated consistently by nonparametric density estimators with observed (X_g, s_g^2) for all genes $g = 1, \ldots, G$. Can this consistent estimator of $m_m(X_g, s_g^2)$ help us construct a most powerful test statistic, together with a consistent estimator of $m_0(X_g, s_g^2)$?

Suppose that p_0 and p_1 are proportions of EE and DE genes, respectively, with $0 \leq p_0, p_1 \leq 1$ and $p_0 + p_1 = 1$, then the mixture marginal density is

$$m_m\left(X_g, s_g^2\right) = p_0 m_0\left(X_g, s_g^2\right) + p_1 m_1\left(X_g, s_g^2\right). \tag{4.3}$$

The ratio of mixture marginal density $m_m(X_g, s_g^2)$ and the null marginal density $m_0(X_g, s_g^2)$ is a monotonic function of the statistic T_g^{NP} expressed in formula (4.2) because

$$\begin{aligned} \frac{m_m\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)} &= \frac{p_0 m_0\left(X_g, s_g^2\right) + p_1 m_1\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)}, \\ &= p_0 + p_1 \frac{m_1\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)}. \end{aligned} \tag{4.4}$$

Thus the test that rejects the null hypothesis when $m_m(X_g, s_g^2)/m_0(X_g, s_g^2)$ is large is also a most powerful test. Note that to calculate this statistic, we only need to estimate $m_m(X_g, s_g^2)$ and $m_0(X_g, s_g^2)$ but do not have to estimate the proportions p_0 and p_1.

Let $\widehat{m}_m(X_g, s_g^2)$ denote any consistent density estimator of $m_m(X_g, s_g^2)$, and let $\widehat{m}_0(X_g, s_g^2)$ denote any consistent estimator of $m_0(X_g, s_g^2)$, such that

$$\begin{aligned} \widehat{m}_m\left(X_g, s_g^2\right) &\xrightarrow{P} m_m\left(X_g, s_g^2\right) \quad \text{as } G \nearrow \infty, \\ \widehat{m}_0\left(X_g, s_g^2\right) &\xrightarrow{P} m_0\left(X_g, s_g^2\right) \quad \text{as } G \nearrow \infty, \end{aligned} \tag{4.5}$$

where $\xrightarrow{P}$ denotes convergence in probability. Then the statistic $\widehat{m}_m(X_g, s_g^2)/\widehat{m}_0(X_g, s_g^2)$ has the optimal testing power asymptotically. Notice the convergence with respect to G, which is usually huge in the microarray and RNA-seq studies.

We have already discussed the availability of a parametric consistent estimator of $m_0(X_g, s_g^2)$ through estimating the hyperparameters d_0 and s_0^2 of σ_g^2 in Section 3. For $m_m(X_g, s_g^2)$, any theoretically consistent density estimator $\widehat{m}_m(X_g, s_g^2)$ of joint data (X_g, s_g^2)

can be used to construct the test statistic (4.4) with asymptotically optimal average power. For example, nonparametric estimators such as histograms, kernel density estimates, and local polynomial estimators can all be utilized. As our test statistic $\widehat{m}_m(X_g, s_g^2)/\widehat{m}_0(X_g, s_g^2)$ involves both parametric and nonparametric parts, we name it the semiparametric optimal test (SPOT).

5. Implementation of SPOT

In this section, we discuss details in implementation of the proposed SPOT procedure.

5.1. Estimation of $m_0(X_g, s_g^2)$

The null marginal density

$$\begin{aligned} m_0\left(X_g, s_g^2\right) &= \int f\left(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2\right)\pi\left(\sigma_g^2\right)d\sigma_g^2 \\ &= \int \frac{e^{-x_g^2/(2v_g\sigma_g^2)}}{\left(2\pi v_g\sigma_g^2\right)^{1/2}}\left(\frac{d_g}{2\sigma_g^2}\right)^{d_g/2}\frac{s^{2(d_g/2-1)}e^{-d_g s_g^2/(2\sigma_g^2)}}{\Gamma\left(d_g/2\right)}\left(\frac{d_0 s_0^2}{2}\right)^{d_0/2}\frac{\sigma_g^{-2(d_0/2+1)}e^{-d_0 s_0^2/2\sigma_g^2}}{\Gamma(d_0/2)}d\sigma_g^2 \\ &= C_2 \cdot s_g^{2(d/2-1)}\left(\frac{x_g^2/v_g + d_0 s_0^2 + d_g s_g^2}{2}\right)^{-(1+d_0+d_g)/2}, \end{aligned} \tag{5.1}$$

where C_2 is a constant. As in Smyth [2] and Hwang and Liu [4], we assume that the distribution of σ_g^2 does not depend on whether a gene is DE or EE. Then, all genes are used to estimate the parameters d_0 and s_0^2. We apply the method of moments proposed in Smyth [2] to get estimates of d_0 and s_0^2. Replacing unknown parameters d_0 and s_0^2 in $m_0(X_g, s_g^2)$ by their consistent method of moments estimates leads to a consistent estimator $\widehat{m}_0(X_g, s_g^2)$ of $m_0(X_g, s_g^2)$.

5.2. A Hybrid Method for Estimation of $m_m(X_g, s_g^2)$

Although any consistent estimator $\widehat{m}_m(X_g, s_g^2)$ can be used to construct a SPOT statistic of the form $\widehat{m}_m(X_g, s_g^2)/\widehat{m}_0(X_g, s_g^2)$, in practice, a density estimator that converges fast would be always preferred. It is known that the accuracy of the density estimators goes down quickly as the dimension increases [13]. We have tried a few two-dimensional density estimators for $\widehat{m}_m(X_g, s_g^2)$, including the kernel estimators. Due to the curse of dimensionality, the direct two-dimensional density estimators do not perform as satisfactory as a hybrid estimator that we develop and would suggest to use. This hybrid estimator has a component that is similar to kernel estimators, whereas it also utilizes the prior information on variances σ_g^2 to help improving the accuracy.

In constructing this estimator, we first estimate the marginal density of X_g by the typical kernel density estimate:

$$\widehat{f}(x_g) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{h}K\left(\frac{x_g - x_i}{h}\right), \tag{5.2}$$

where h is a positive value known as bandwidth. We estimate the conditional density of $f(s_g^2 \mid x_g)$ by using

$$f\left(s_g^2 \mid x_g\right) = \int f\left(s_g^2 \mid \sigma_g^2, x_g\right) f\left(\sigma_g^2 \mid x_g\right) d\sigma_g^2 = \int f\left(s_g^2 \mid \sigma_g^2\right) f\left(\sigma_g^2 \mid x_g\right) d\sigma_g^2, \tag{5.3}$$

where the second equality is a result of the independence between s_g^2 and x_g given the parameter σ_g^2. The distribution of $s_g^2 \mid \sigma_g^2$ is $(\sigma_g^2/d_g)\chi_{d_g}^2$ for normal-distributed observations. Now we need to estimate $f(\sigma_g^2 \mid x_g)$. Denote the set of genes that lie within bandwidth distance to gene g as $\{A_g : i \in A_g$ if and only if $|x_i - x_g| < h\}$. We estimate $f(\sigma_g^2 \mid x_g)$ by the following approximation that is based on the neighborhood of x_g, A_g:

$$\frac{1}{\#\{A_g\}}\sum_{i\in A_g}\frac{f(s_i^2 \mid \sigma^2)\pi(\sigma^2)}{\int f(s_i^2 \mid \sigma^2)\pi(\sigma^2)d\sigma^2}. \tag{5.4}$$

The $\#\{A_g\}$ in formula denotes the number of genes in set A_g. Substituting exact parametric form of $f(s_g^2 \mid \sigma^2)$ and $\pi(\sigma^2)$ into above formulas leads to the explicit form

$$\widehat{f}\left(s_g^2 \mid x_g\right) = C_3 \cdot \frac{1}{\#\{A_g\}}\sum_{i\in A_g} s_g^{2(d/2-1)}\left(\frac{d_g s_i^2 + d_0 s_0^2 + d_g s_g^2}{2}\right)^{-(d_0+d_g)/2}, \tag{5.5}$$

where C_3 is a constant. The product between the kernel estimate $\widehat{f}(x_g)$ and the conditional estimate $\widehat{f}(s_g^2 \mid x_g)$ provides us a joint density estimator of the mixture

$$\widehat{m}_m\left(X_g, s_g^2\right) = \widehat{f}(x_g) \cdot \widehat{f}\left(s_g^2 \mid x_g\right). \tag{5.6}$$

With this approximation, we cannot theoretically show that the resulting estimator, $\widehat{m}_m(X_g, s_g^2)$, is consistent but it works better in practice than the consistent kernel density estimator of the joint density $m_m(X_g, s_g^2)$.

6. Simulation Study

In order to evaluate the performance of our proposed SPOT procedure, we performed three simulation studies. The gene expression data were simulated from Normal (μ_{gi}, σ_g^2) for observations of gene g in treatment group i. The way to sample μ_{gi} and σ_g^2 differs across

simulation studies. We assume that there are two treatment groups and 3 replicates per treatment group. For each simulation setting, one hundred sets of gene expression data were independently simulated, and each dataset included 10,000 genes. The performances of the SPOT, moderated t, F_{SS}, and ordinary t-test statistics were evaluated for by comparing their average behavior averaged across the 100 datasets.

6.1. Simulation Study I

In the first simulation study, we have two settings that differ in the number of DE genes. For the first setting, $G_1 = 2,500$ are DE genes whereas the other $G_0 = 7,500$ are EE. In the second setting, only $G_1 = 1,800$ are DE while the other $G_0 = 8,200$ are EE. Gene expression means μ_{gi} and variances σ_g^2 were simulated as follows:

$$\begin{aligned}
&\mu_{g1} = 0 \quad \forall g;\\
&\mu_{g2} \sim \text{Normal}\left(0.5, 0.3^2\right) \quad \text{for } g = 1 \text{ to } 0.3G_1;\\
&\mu_{g2} \sim \text{Normal}\left(1, 0.3^2\right) \quad \text{for } g = (0.3G_1 + 1) \text{ to } 0.9G_1;\\
&\mu_{g2} \sim t_1\ (0.5) \quad \text{for } g = (0.9G_1 + 1) \text{ to } G_1;\\
&\mu_{g2} = 0 \quad \text{for } g = (G_1 + 1) \text{ to } 10000;\\
&\sigma_g^2 \sim \text{Gamma}\ (2, 4) \quad \forall g.
\end{aligned} \tag{6.1}$$

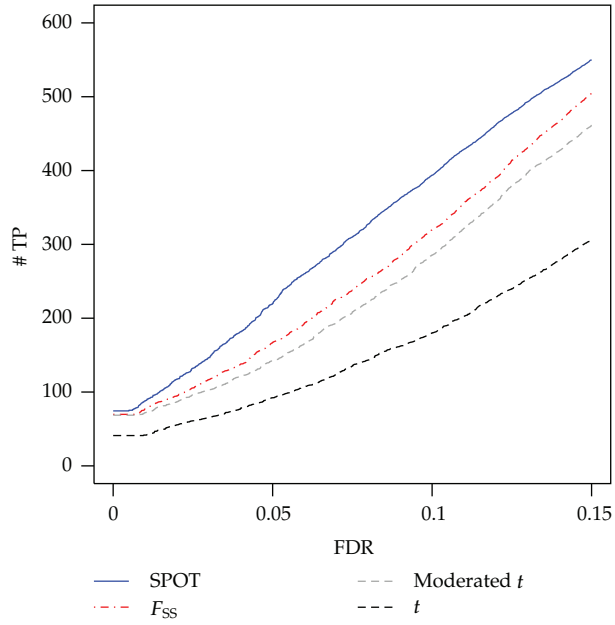
For each simulated data, SPOT, moderated t, F_{SS}, and ordinary t-test statistics were calculated and evaluated using the number of selected true positives at various FDR levels. The plots of number of true positives versus FDR for SPOT, moderated t, F_{SS}, and ordinary t-test statistics are shown in Figure 1. Simulation settings 1 and 2 generated similar results. The ordinary t-test is the poorest method under comparison. The moderated t-test is considerably better than the ordinary t-test although it is worse than F_{SS} test. Our proposed SPOT test is superior to all other three methods, with the largest number of true positive findings than the other three statistics at the same FDR levels.

6.2. Simulation Study II

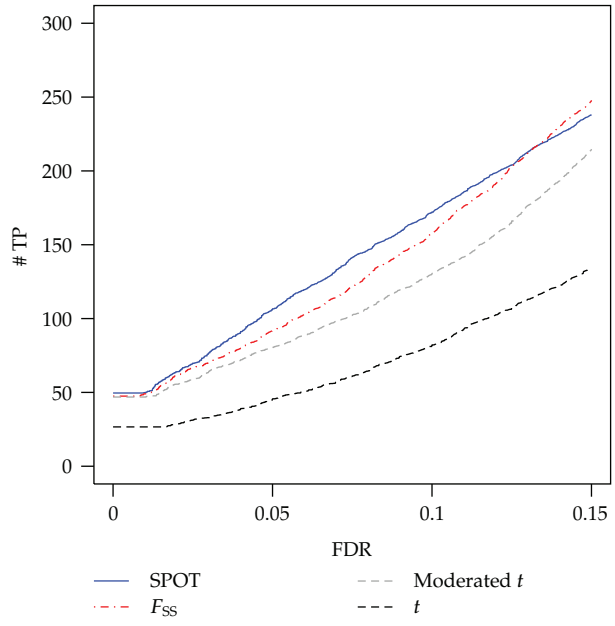
To check how the variance distribution affects the relative ranking of the SPOT procedure, we did another simulation study the same as the setting 1 of simulation study I except that the variances were simulated from a log-normal distribution, which is the assumption under which the F_{SS} test was derived. As Figure 2 shows, the results are similar to those from simulation I. The SPOT procedure still performs much better than all the other three methods.

6.3. Simulation Study III

Typically, the parametric test achieves higher power than the nonparametric test if the parametric assumption is appropriate. To check the robustness of the SPOT procedure, we



(a) Simulation I, setting 1



(b) Simulation I, setting 2

Figure 1: Simulation study I: plots of number of true positives (# TP) versus false discovery rate (FDR) from analyses using SPOT, moderated t, and F_{SS} methods.

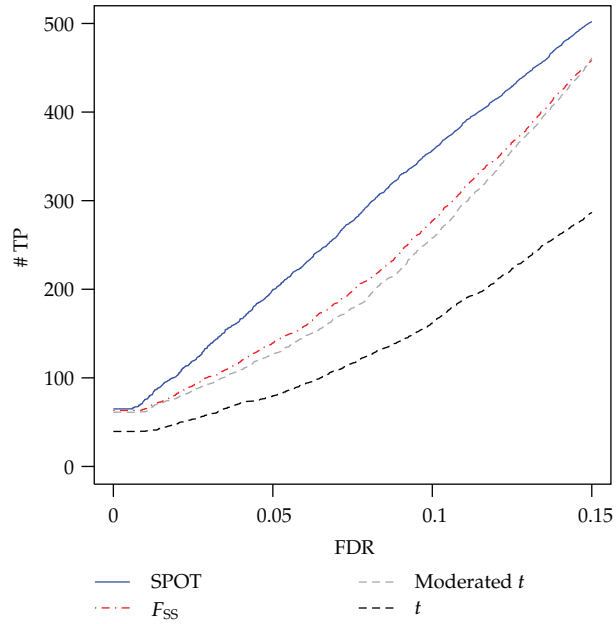


Figure 2: Simulation study II: plots of number of true positives (# TP) versus false discovery rate (FDR) from analyses using SPOT, moderated t, and F_{SS} methods.

simulated data under the parametric assumption for both μ_{gi} and σ_g^2 under which the F_{SS} test was derived. Specifically, for the 2,500 differentially expressed genes, μ_{gi} were drawn from a normal distribution with mean 1.2 and standard deviation 0.3, σ_g^2 were sampled from a log-normal distribution with parameters −0.96 and 0.8. Figure 3 shows that the SPOT procedure and the F_{SS} test are comparable to each other when FDR is small (less than 0.05) and they are both much better than the moderated t-test and the ordinary t-test. When FDR is between 0.05 and 0.15, the F_{SS} test is the best while the SPOT procedure is the next best performing procedure, which is still much better than the moderated t-test and the ordinary t-test.

7. Evaluation Using the Golden Spike Microarray Data

In this section, we compare the performances of different methods using a real microarray dataset from experiments conducted using Affymetrix GeneChip in the Golden Spike Project. The Golden Spike Project generated microarray datasets comparing two replicated groups in which the relative concentrations of a large number of genes are known. The two groups are the spike-in group and the control group, each with three chips. Data and information related to this project are available through the website http://www2.ccr.buffalo.edu/halfon/spike/. More specifically, the Golden Spike dataset included 1309 individual cRNAs "spiked in" at known relative concentrations between the two groups. The fold-changes between the spike-in and control group were assigned at different levels for different cRNAs, and the levels ranged from 1.2 to 4. Hence, these cRNAs were truly "differentially expressed" between groups and we consider them as DE genes. In addition, a background sample of 2551 RNA species was present at identical concentrations in both samples. So these 2551 RNA species were not differentially expressed between the two

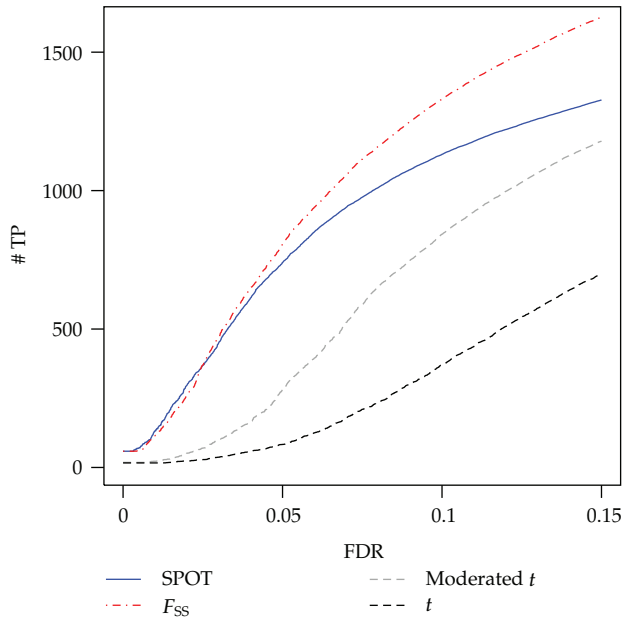


Figure 3: Simulation study III: plots of number of true positives (# TP) versus false discovery rate (FDR) from analyses using SPOT, moderated t, and F_{SS} methods.

Table 1: Golden Spike data: number of true positives selected by three testing procedures at critical FDR levels.

Method	FDR					
	0.01	0.02	0.05	0.1	0.15	0.2
SPOT	754	847	947	986	1015	1051
Moderated t	466	588	821	911	969	1018
F_{SS}	442	563	824	908	975	1016

groups. With the knowledge of the true differential expression status, this real microarray dataset provides an ideal case to evaluate the performances of different methods without imposing any distributional assumption for variances and means as usually is done in simulation studies.

With the summary dataset downloaded from the Golden Spike Project website, we calculated the SPOT, the moderated t, the ordinary t, and the F_{SS} statistics and evaluated their performances using the true statuses of RNA based on the design. Figure 4 shows the plots of number of true positives versus FDR for the ordinary t, moderated t, F_{SS}, and SPOT procedures over a range of FDR $\in [0, 0.15]$ which is of most practical interest. It can be observed that the performance of the SPOT procedure improves over the performances of the other three methods throughout the whole range of FDR in these plots. In addition, the improvement is substantial at lower FDR levels. For example, the SPOT procedure detects 754 true positives at the FDR level of 1% while the moderated t-test only detects 466 and the F_{SS} test only detects 442 true positives (Table 1).

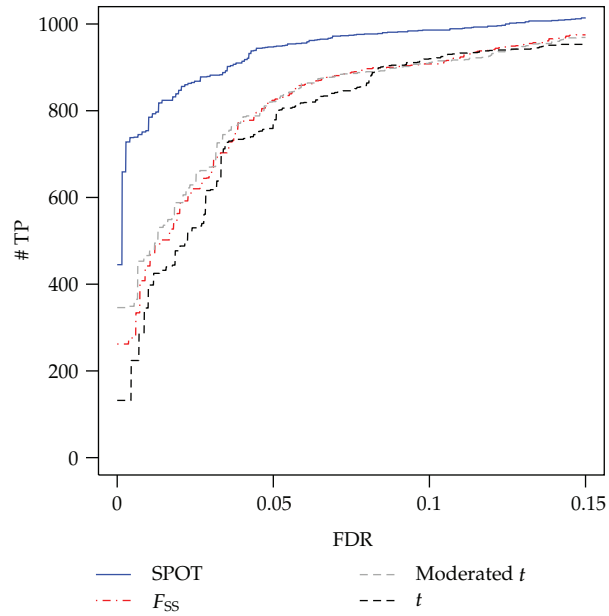


Figure 4: Golden spike data: plots of number of true positive (TP) genes versus false discovery rate (FDR) from analysis using SPOT, moderated t, and F_{SS} tests.

8. Summary

In this paper, we have derived a semiparametric optimal testing (SPOT) procedure for high-dimensional gene expression data analysis. Although the method is illustrated for analyzing microarray data, it can be applied to any high-dimensional testing problem with normal model. Our test statistic is justified to be asymptotically most powerful, without any assumption on the mean parameter of differential expression. The asymptotic property is derived when the number of genes is large, which is reasonable for high-dimensional gene expression studies. We also provided an approximate version to implement the SPOT procedure in practice and evaluated the performance of our proposed test statistic using both simulation studies and real microarray data analysis. The proposed SPOT method is shown to outperform the popularly applied moderated t and the F_{SS} statistics, which are optimal only under certain normality conditions of the mean. There is still potential in improving the performance of SPOT procedure if better density estimates can be found for the marginal distributions $m_m(X_g, s_g^2)$ and $m_0(X_g, s_g^2)$.

Appendix

Proof of the Claim That the Moderated t-Test Achieves the Optimal Average Power Asymptotically under the Assumptions That $\mu_g \sim N(0, v_0\sigma_g^2)$ and $1/\sigma_g^2 \sim 1/d_0 s_0^2 \chi_{d_0}^2$

Under Smyth's [2] model assumptions, the most power test statistic formula (3.1) derived under the Neyman-Pearson lemma becomes

$$T_g^{\mathrm{NP}} = \frac{\iint f\left(X_g, s_g^2 \mid \mu_g, \sigma_g^2\right)\pi_1(\mu_g)\pi_1\left(\sigma_g^2\right)d\mu_g d\sigma_g^2}{\int f(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2)\pi_0(\sigma_g^2)d\sigma_g^2}$$

$$= \frac{\int\int\left(e^{-(x_g-\mu_g)^2/2v_g\sigma_g^2}/\left(2\pi v_g\sigma_g^2\right)^{1/2}\right)\left(d_g/2\sigma_g^2\right)^{d_g/2}\left(s^{d_g-2}e^{-d_gs_g^2/2\sigma_g^2}/\Gamma\left(d_g/2\right)\right)\mathcal{A}}{\int\left(e^{-x_g^2/2v_g\sigma_g^2}/\left(2\pi v_g\sigma_g^2\right)^{1/2}\right)\left(d_g/2\sigma_g^2\right)^{d_g/2}\left(s^{2(d_g/2-1)}/\Gamma\left(d_g/2\right)\right)e^{-d_gs_g^2/2\sigma_g^2}\cdot\mathcal{B}}$$

$$= C_1\cdot\left(\frac{x_g^2/\left(v_0+v_g\right)+d_0s_0^2+d_gs_g^2}{x_g^2/v_g+d_0s_0^2+d_gs_g^2}\right)^{-(1+d_0+d_g)/2}, \tag{A.1}$$

where $\mathcal{A}$ denotes $(e^{-\mu_g^2/(2v_0\sigma_g^2)}/(2\pi v_0\sigma_g^2)^{1/2})(d_0s_0^2/2)^{d_0/2}(\sigma_g^{-d_0+2}/\Gamma(d_0/2))e^{-d_0s_0^2/(2\sigma_g^2)}d\mu_g d\sigma_g^2$, and $\mathcal{B}$ denotes $(d_0s_0^2/2)^{d_0/2}(\sigma_g^{-2(d_0/2-1)}/\Gamma(d_0/2))e^{-d_0s_0^2/(2\sigma_g^2)}d\sigma_g^2$, which is a monotonic function of Smyth's [2] moderated t-statistic, with C_1 being some constant. Thus the claim follows with existence of consistent estimates of d_0 and s_0^2, which has been shown in Smyth [2].

References

[1] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, no. 1, pp. 31–46, 2002.

[2] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, article 3, 2004.

[3] X. Cui, J. Hwang, J. Qiu, N. J. Blades, and A. Churchill, "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, vol. 6, no. 1, pp. 59–75, 2005.

[4] J. T. G. Hwang and P. Liu, "Optimal tests shrinking both means and variances applicable to microarray data analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 9, article 36, 2010.

[5] T. Cai, J. Jeng, and H. Li, "Robust detection and identification of sparse segments in ultra-high dimensional data analysis," *Journal of the Royal Statistical Society: Series B*, vol. 14, part 4, 2012.

[6] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[7] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayes analysis of a microarray experiment," *The Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1151–1160, 2001.

[8] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[9] Y. C. Tai and T. P. Speed, "A multivariate empirical Bayes statistic for replicated microarray time course data," *The Annals of Statistics*, vol. 34, no. 5, pp. 2387–2412, 2006.

[10] G. W. Wright and R. M. Simon, "A random variance model for detection of differential gene expression in small microarray experiments," *Bioinformatics*, vol. 19, no. 18, pp. 2448–2455, 2003.

[11] T. Tong and Y. Wang, "Optimal shrinkage estimation of variances with applications to microarray data analysis," *The Journal of the American Statistical Association*, vol. 102, no. 477, pp. 113–122, 2007.

[12] H. Bar, J. Booth, E. Schifano, and M. T. Wells, "Laplace approximated EM microarray analysis: an empirical Bayes approach for comparative microarray experiments," *Statistical Science*, vol. 25, no. 3, pp. 388–407, 2010.

[13] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer, New York, NY, USA, 2006.