Optimal Shrinkage Estimation of Variances With Applications to Microarray Data Analysis

Tiejun TONG and Yuedong WANG

Microarray technology allows a scientist to study genomewide patterns of gene expression. Thousands of individual genes are measured with a relatively small number of replications, which poses challenges to traditional statistical methods. In particular, the gene-specific estimators of variances are not reliable and gene-by-gene tests have low powers. In this article we propose a family of shrinkage estimators for variances raised to a fixed power. We derive optimal shrinkage parameters under both Stein and squared loss functions. Our results show that the standard sample variance is inadmissible under either loss function. We propose several estimators for the optimal shrinkage parameters and investigate their asymptotic properties under two scenarios: large number of replications and large number of genes. We conduct simulations to evaluate the finite sample performance of the data-driven optimal shrinkage estimators and compare them with some existing methods. We construct F-like statistics using these shrinkage variance estimators and apply them to detect differentially expressed genes in a microarray experiment. We also conduct simulations to evaluate performance of these F-like statistics and compare them with some existing methods.

KEY WORDS: F-like statistic; Gene expression data; Inadmissibility; James–Stein shrinkage estimator; Loss function.

1. INTRODUCTION

The development of microarray technology has revolutionized the study of molecular biology and become a standard tool in genomics research. Instead of working on a gene-by-gene basis, microarray technology allows the scientists to view the expression of thousands of genes from an experimental sample simultaneously (Nguyen, Arpat, Wang, and Carroll 2002; Leung and Cavalieri 2003). Due to the cost, it is common that thousands of genes are measured with a small number of replications (Lönnstedt and Speed 2002; Kendziorski, Newton, Lan, and Could 2003). As a consequence, we are faced with a "large G, small n" paradigm, where G is the total number of genes and n is the number of replications. The standard gene-specific estimators of variances are unreliable due to the relatively small number of replications. Consequently, the commonly used statistical methods, such as t test or F test, for detecting differentially expressed genes on a gene-by-gene basis have low powers (Callow, Dudoit, Gong, Speed, and Rubin 2000). On the other hand, the assumption that variances are equal for all genes is unlikely to be true. Thus, tests based on a pooled common variance estimator for all genes are at the risk of generating misleading results (Cui, Hwang, Qiu, Blades, and Churchill 2005).

A number of approaches to improving variance estimation and hypothesis testing have emerged. Kamb and Ramaswami (2001) suggested a simple regression estimation of local variances. Storey and Tibshirani (2003) added a small constant to the gene-specific variance estimators in their SAM t test to stabilize the small variances. Lin, Nadler, Lan, Attie, and Yandell (2003) proposed a data-adapted robust estimator of array error based on a smoothing spline and standardized local median absolute deviation. Jain et al. (2003) proposed a local-pooled-error estimation procedure, which borrows strength from genes in local intensity regions to estimate array error variability. Baldi and Long (2001) proposed a regularized t test by replacing the usual variance estimator with a Bayesian estimator. Lönnstedt and Speed (2002) proposed an empirical Bayes approach that combines information across genes. Kendziorski et al. (2003) extended the empirical Bayes method using hierarchical gamma–gamma and lognormal–normal models.

Cui and Churchill (2003) compared three variance estimators: the gene-specific estimator, the pooled estimator across genes, and the hybrid estimator as the average of the gene-specific and the pooled estimators. Applying the standard James–Stein shrinkage method to log transformed estimates of variances, Cui et al. (2005) proposed a James–Stein type shrinkage estimator for variances (referred to as the CHQBC estimator in the remainder of this article). Compared to some existing tests, they showed that the F test using the James–Stein type variance estimator has the best or nearly the best power to detect differentially expressed genes over a wide range of situations.

The research so far has concentrated on the methodology. Little is known about the theoretical properties of various shrinkage variance estimators. Shrinkage variance estimation has a long history that began with the amazing inadmissibility result discovered by Stein (1964), where the standard sample variance is improved by a shrinkage estimator using information contained in the sample mean. Much research has been done since then (Maatta and Casella 1990; Kubokawa 1999), most of which concerned single variances (Kubokawa 1999), which are not applicable to microarray data analysis because the homogeneity of the variances is unlikely to be true. Some research has been devoted to the shrinkage estimator of a covariance matrix (Kubokawa and Srivastava 2003). However, all these methods require $n > G$ to ensure nonsingularity of the sample covariance matrix. Therefore, these methods break down for microarray data analysis.

We propose new optimal shrinkage estimators in this article. Instead of using information in the sample mean (Stein 1964), we borrow information across variances. We will show that the standard sample variance is inadmissible. Therefore,

Tiejun Tong is Postdoctoral Associate, Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520 (E-mail: *tiejun.tong@yale.edu*). Yuedong Wang is Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106 (E-mail: *yuedong@pstat.ucsb.edu*). This research was supported by National Institutes of Health grant R01 GM58533. The authors thank the joint editor, the associate editor, and two referees for their constructive comments and suggestions that have led to a substantial improvement in the article. The authors also thank Xiangqin Cui and Gary Churchill in the Jackson Laboratory for providing the data and answering many technical questions about the MAANOVA package.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000001266

our results extend Stein's theory for multiple means (James and Stein 1961) to multiple variances. An important insight of this article is that a better variance estimator does not necessarily lead to a more powerful test. Specifically, because the variance appears in the denominator, an F test using an estimator of the reciprocal of the variance is more powerful than that using the reciprocal of an estimator of the variance (Sec. 5). We consider optimal shrinkage estimators for the parameter $(\sigma_g^2)^t$, where σ_g^2 is the true variance associated with gene g, $g = 1, \ldots, G$, and t is a fixed nonzero power.

Our methods and theory are general. Nevertheless, we present our methods in the framework of microarray data analysis. In Section 2 we introduce the CHQBC estimator and propose a modified version. In Section 3 we derive optimal shrinkage estimators for $(\sigma_g^2)^t$ under two common loss functions and show that the optimal shrinkage estimators dominate the standard gene-specific variance estimator. We also propose various estimators for the optimal shrinkage parameters and investigate their asymptotic properties under two scenarios: $n \to \infty$ with fixed G and $G \to \infty$ with fixed n. In Section 4 we conduct simulations to evaluate the performance of the optimal shrinkage estimators and compare them with the CHQBC estimator and the modified CHQBC estimator. In Section 5 we construct F-like statistics using our optimal shrinkage estimators to detect differentially expressed genes for microarray data, and conduct simulations to evaluate and compare performances of these F-like statistics. We conclude this article in Section 6 with a brief discussion.

2. CHQBC ESTIMATOR AND ITS MODIFICATION

Let G ($G \geq 3$) be the number of genes, $X_g = \sigma_g^2 \chi_{g,\nu}^2$, $g = 1, \ldots, G$, where $\chi_{g,\nu}^2$ are iid random variables that follow the chi-squared distribution with ν degrees of freedom. Consider the transformation

$$X_g' = \ln \sigma_g^2 + \epsilon_g',$$

where $X_g' = \ln(X_g/\nu) - m$, $\epsilon_g' = \ln(\chi_{g,\nu}^2/\nu) - m$, and $m = \mathrm{E}(\ln(\chi_{g,\nu}^2/\nu))$. Applying the James–Stein shrinkage method to X_g' and then transforming back to the original scale, Cui et al. (2005) proposed the CHQBC estimator

$$\tilde{\sigma}_g^2 = \left(\prod_{g=1}^{G}\left(\frac{X_g}{\nu}\right)^{1/G}\right)B \times \exp\left[\left(1 - \frac{(G-3)V}{\sum(\ln X_g - \overline{\ln X_g})^2}\right)_+ \times (\ln X_g - \overline{\ln X_g})\right], \quad (1)$$

where $V = \mathrm{var}(\epsilon_g')$, $\overline{\ln X_g} = \sum_{g=1}^{G} \ln(X_g)/G$, and $B = \exp(-m)$.

Let $Z_g = X_g/\nu$, $Z_{\mathrm{pool}} = \prod_{g=1}^{G} Z_g^{1/G}$, and $\hat{\alpha}_0 = 1 - (1 - (G-3)V/\sum(\ln X_g - \overline{\ln X_g})^2)_+$. It is easy to check that the CHQBC estimator (1) can be rewritten as

$$\tilde{\sigma}_g^2 = B(Z_{\mathrm{pool}})^{\alpha}(Z_g)^{1-\alpha} \quad (2)$$

with $\alpha = \hat{\alpha}_0$. Note that when $\sigma_g^2 = \sigma^2$ for all g, $\mathrm{E}(Z_{\mathrm{pool}}) \to \sigma^2/B$ as $G \to \infty$. That is, BZ_{pool} is an unbiased estimator of σ^2 when $\sigma_g^2 = \sigma^2$ for all g. On the other hand, Z_g is an unbiased estimator of σ_g^2. Therefore, it is reasonable to consider the combination of two unbiased estimators:

$$\hat{\sigma}_g^2 = (BZ_{\mathrm{pool}})^{\alpha}(Z_g)^{1-\alpha}, \qquad 0 \leq \alpha \leq 1. \quad (3)$$

When necessary, the dependence of $\hat{\sigma}_g^2$ on α will be expressed explicitly as $\hat{\sigma}_g^2(\alpha)$. We refer to $\hat{\sigma}_g^2(\hat{\alpha}_0)$ as the modified CHQBC estimator. When variances σ_g^2 are similar, it is likely that $\hat{\alpha}_0 \approx 1$ and thus $\hat{\sigma}_g^2(\hat{\alpha}_0) \approx \tilde{\sigma}_g^2(\hat{\alpha}_0)$. When $\hat{\alpha}_0 \approx 0$, $\tilde{\sigma}_g^2(\hat{\alpha}_0) \approx BZ_g$, which is biased when $B \neq 1$. Simulations in Section 4.1 indicate that the modified CHQBC estimator $\hat{\sigma}_g^2(\hat{\alpha}_0)$ always performs better than the original CHQBC estimator $\tilde{\sigma}_g^2(\hat{\alpha}_0)$ for estimating σ_g^2.

The estimator $\hat{\sigma}_g^2$ has a very simple structure: it borrows information across genes by shrinking each gene-specific variance toward the bias corrected geometric mean of variances for all genes. The shrinkage parameter $\hat{\alpha}_0$ was obtained by applying the James–Stein method to the logarithm of sample variances that do not follow the normal distribution (Cui et al. 2005). Therefore, $\hat{\alpha}_0$ may not be optimal and theoretical properties of $\tilde{\sigma}_g^2$ and $\hat{\sigma}_g^2$ are unknown.

3. OPTIMAL SHRINKAGE

We now consider the family of shrinkage estimators $\hat{\sigma}_g^2$ in (3) with the shrinkage parameter α unfixed. There is no shrinkage when $\alpha = 0$, and all variance estimates are shrunken to the bias corrected geometric mean when $\alpha = 1$. Our goal is to find the optimal shrinkage parameter α under the Stein loss function (James and Stein 1961)

$$L_1(\sigma^2, \hat{\sigma}^2) = \hat{\sigma}^2/\sigma^2 - \ln(\hat{\sigma}^2/\sigma^2) - 1 \quad (4)$$

and the squared loss function

$$L_2(\sigma^2, \hat{\sigma}^2) = (\hat{\sigma}^2/\sigma^2 - 1)^2. \quad (5)$$

Equation (4) is also called the entropy loss or Kullback–Leibler loss function (Kubokawa 1999). The Stein loss function penalizes gross underestimation as heavily as gross overestimation, whereas the squared loss function penalizes the gross underestimation less than the gross overestimation.

Throughout the remainder of this article, we assume that $Z_g = \sigma_g^2 \chi_{g,\nu}^2/\nu$, $g = 1, \ldots, G$, are independent random variables and $G \geq 2$. As discussed in Section 1, we will derive an optimal shrinkage estimator for $(\sigma_g^2)^t$ for any power $t \neq 0$. The estimators for σ_g^2 and $1/\sigma_g^2$ are special cases with $t = 1$ and $t = -1$. We will first generalize the estimator (3) for estimating $(\sigma_g^2)^t$. Define

$$h_n(t) = \left(\frac{\nu}{2}\right)^t \left(\frac{\Gamma(\nu/2)}{\Gamma(\nu/2 + t/n)}\right)^n, \quad (6)$$

where $\Gamma(\cdot)$ is the gamma function.

Lemma 1. For any nonzero $t > -\nu/2$, we make the following statements:

(a) The term $h_1(t)Z_g^t$ is an unbiased estimator of $(\sigma_g^2)^t$.

(b) When $\sigma_g^2 = \sigma^2$ for all g, $h_G(t)Z_{\mathrm{pool}}^t$ is an unbiased estimator of $(\sigma^2)^t$.

The proof is straightforward. Note that $h_1(t)Z_g^t$ is the gene-specific estimator of $(\sigma_g^2)^t$. Combining (3) and Lemma 1, we now propose a family of shrinkage estimators for $(\sigma_g^2)^t$:

$$\hat{\sigma}_g^{2t} = (h_G(t)Z_{\text{pool}}^t)^\alpha (h_1(t)Z_g^t)^{1-\alpha}, \qquad 0 \le \alpha \le 1. \tag{7}$$

Note that $h_1(1) = 1$ and $h_G(1) \to B$ as $G \to \infty$. Therefore, when $t = 1$ and G is large, (7) reduces to (3). Let $\sigma_{\text{pool}}^2 = (\prod_{g=1}^G \sigma_g^2)^{1/G}$, $\boldsymbol{\sigma}^{2t} = (\sigma_1^{2t}, \ldots, \sigma_G^{2t})$ and $\hat{\boldsymbol{\sigma}}^{2t} = (\hat{\sigma}_1^{2t}, \ldots, \hat{\sigma}_G^{2t})$.

3.1 Optimal Estimator Under the Stein Loss Function

Under the Stein loss function (4), it is easy to check that the average risk is

$$\begin{aligned} R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) &\triangleq \frac{1}{G}\sum_{g=1}^G \mathrm{E}(L_1(\sigma_g^{2t}, \hat{\sigma}_g^{2t})) \\ &= \frac{h_G^\alpha(t)h_1^{1-\alpha}(t)}{h_1^{G-1}(\alpha t/G)h_1((1-\alpha+\alpha/G)t)}(\sigma_{\text{pool}}^2)^{\alpha t}\frac{1}{G}\sum_{g=1}^G(\sigma_g^2)^{-\alpha t} \\ &\quad - \ln(h_G^\alpha(t)h_1^{1-\alpha}(t)) - t\Psi\left(\frac{\nu}{2}\right) + t\ln\left(\frac{\nu}{2}\right) - 1, \end{aligned} \tag{8}$$

where $t > -\nu/2$, $\Psi(t) = \Gamma'(t)/\Gamma(t)$ is the digamma function (Abramowitz and Stegun 1972), and the second equality is derived after some tedious but straightforward algebra using Lemma 1 and the fact that $\mathrm{E}\ln(\chi_{g,\nu}^2) = \Psi(\nu/2) + \ln(2)$. Then the optimal α under the Stein loss function is $\alpha_1^* = \arg\min_{\alpha\in[0,1]} R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$. Denote the optimal estimator under the Stein loss function as $\hat{\sigma}_g^{2t}(\alpha_1^*)$. In the following discussion, the derivatives $R_k'(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ and $R_k''(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ are with respect to α, $k = 1, 2$.

Theorem 1. For any fixed G, ν, and nonzero $t > -\nu/2$, $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ is a strictly convex function of α on $[0, 1]$ that satisfies

(a) $R_1'(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=0} < 0$

(b) $R_1'(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=1} \ge 0$, where the equality holds if and only if $\sigma_g^2 = \sigma^2$ for all g

(c) $R_1''(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) > 0$ for all $\alpha \in [0, 1]$.

Corollary 1. For any fixed G, ν, and nonzero $t > -\nu/2$, under the Stein loss function,

(a) There exists a unique α_1^* in $(0, 1]$ that is the solution to $R_1'(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) = 0$.

(b) The estimator $h_1(t)Z_g^t$ is inadmissible for $(\sigma_g^2)^t$ because $\alpha_1^* > 0$.

(c) The parameter $\alpha_1^* = 1$ iff $\sigma_g^2 = \sigma^2$ for all g. Thus, $h_G(t)Z_{\text{pool}}^t$ is also inadmissible for $(\sigma_g^2)^t$ when σ_g^2 are not all the same. When $\sigma_g^2 = \sigma^2$ for all g, $h_G(t)Z_{\text{pool}}^t$ dominates all other estimators in the family (7).

Theorem 2. For any fixed G and nonzero $t > -\nu/2$, as $\nu \to \infty$, we have

(a) $\alpha_1^* \to 0$ when σ_g^2 are not all the same

(b) $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ approaches a constant function of α when $\sigma_g^2 = \sigma^2$ for all g.

Proofs of Theorems 1 and 2 can be found in Tong and Wang (2005). Theorem 2 indicates that it is unnecessary to borrow information from other genes when the number of replications for each gene is large.

3.2 Optimal Estimator Under the Squared Loss Function

Under the squared loss function (5), it is easy to check that the average risk is

$$\begin{aligned} R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) &\triangleq \frac{1}{G}\sum_{g=1}^G \mathrm{E}(L_2(\sigma_g^{2t}, \hat{\sigma}_g^{2t})) \\ &= \frac{h_G^{2\alpha}(t)h_1^{2(1-\alpha)}(t)}{h_1^{G-1}(2\alpha t/G)h_1(2(1-\alpha+\alpha/G)t)} \\ &\quad \times (\sigma_{\text{pool}}^2)^{2\alpha t}\frac{1}{G}\sum_{g=1}^G(\sigma_g^2)^{-2\alpha t} \\ &\quad - \frac{2h_G^\alpha(t)h_1^{1-\alpha}(t)}{h_1^{G-1}(\alpha t/G)h_1((1-\alpha+\alpha/G)t)} \\ &\quad \times (\sigma_{\text{pool}}^2)^{\alpha t}\frac{1}{G}\sum_{g=1}^G(\sigma_g^2)^{-\alpha t} + 1, \end{aligned} \tag{9}$$

where $t > -\nu/4$.

Theorem 3. For any fixed G, ν, and nonzero power $t > -\nu/4$, we have

(a) $R_2'(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=0} < 0$

(b) $R_2'(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=1} > 0$.

Theorem 3 implies that the gene-specific estimator, $h_1(t)Z_g^t$, is inadmissible. Contrary to the result under the Stein loss function, the pooled estimator, $h_G(t)Z_{\text{pool}}^t$, is also inadmissible even when $\sigma_g^2 = \sigma^2$ for all g. By Theorem 3 and the fact that $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) \ge 0$, there exists an α_2^* that minimizes $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$. However, $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ is not guaranteed to be a convex function of α on $[0, 1]$. Therefore, α_2^* may not be unique. A counterexample with very large ν was provided by Tong and Wang (2005). Nevertheless, for small ν, $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ is always strictly convex in millions of simulations under various situations. Denote the optimal estimator under the squared loss function as $\hat{\sigma}_g^{2t}(\alpha_2^*)$.

Theorem 4. For any fixed G and nonzero $t > -\nu/4$, as $\nu \to \infty$,

(a) $\alpha_2^* \to 0$ when σ_g^2 are not all the same

(b) $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ approaches a constant function of α when $\sigma_g^2 = \sigma^2$ for all g.

Proofs of Theorems 3 and 4 can be found in Tong and Wang (2005).

3.3 Estimation of the Optimal Shrinkage Parameters

Both α_1^* and α_2^* depend on the unknown quantity

$$b(\boldsymbol{\sigma}^2, \eta) = (\sigma^2_{\text{pool}})^{\eta} \frac{1}{G} \sum_{g=1}^{G} (\sigma_g^2)^{-\eta}, \tag{10}$$

where $\eta = \alpha t$ or $\eta = 2\alpha t$. A simple estimator of $b(\boldsymbol{\sigma}^2, \eta)$ is $b(\mathbf{Z}, \eta)$, where $\mathbf{Z} = (Z_1, \ldots, Z_G)$. Denote $\hat{\alpha}_1^*$ and $\hat{\alpha}_2^*$ as the estimates of α_1^* and α_2^* with $b(\boldsymbol{\sigma}^2, \eta)$ in (8) and (9) replaced by $b(\mathbf{Z}, \eta)$. The following theorem shows that $\sigma_g^{2t}(\hat{\alpha}_1^*)$ and $\sigma_g^{2t}(\hat{\alpha}_2^*)$ are asymptotically optimal, and $\hat{\alpha}_1^*$ and $\hat{\alpha}_2^*$ are consistent under certain conditions as $\nu \to \infty$.

Theorem 5. For any fixed G and nonzero t, when $\nu \to \infty$, we have

(a) $b(\mathbf{Z}, \alpha t) \xrightarrow{\text{a.s.}} b(\boldsymbol{\sigma}^2, \alpha t)$ uniformly for $\alpha \in [0, 1]$
(b) $R_k(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\hat{\alpha}_k^*(\nu))) - R_k(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha_k^*(\nu))) \xrightarrow{\text{a.s.}} 0, k = 1, 2$
(c) $\hat{\alpha}_1^*(\nu) \xrightarrow{\text{a.s.}} 0$ and $\hat{\alpha}_2^*(\nu) \xrightarrow{\text{a.s.}} 0$ when σ_g^2 are not all the same.

Proof of Theorem 5 can be found in Tong and Wang (2005). For microarray data, ν is small and G is large. In the following text, we investigate asymptotic properties as $G \to \infty$. We now consider σ_g^2 as random variables and assume that $\sigma_g^2 \overset{\text{iid}}{\sim} F$, $g = 1, \ldots, G$.

Lemma 2. For any fixed nonzero t with $\nu > 2t$, $\mathrm{E}(\sigma_1^2)^{-t} < \infty$, and $\mathrm{E}(\ln(\sigma_1^2)) < \infty$, we have $w(\alpha t) b(\mathbf{Z}, \alpha t) - b(\boldsymbol{\sigma}^2, \alpha t) \xrightarrow{\text{a.s.}} 0$ uniformly for $\alpha \in [0, 1]$ as $G \to \infty$, where $w(\alpha t) = (\nu/2)^{\alpha t} h_1(-\alpha t) \exp[-\alpha t \Psi(\nu/2)]$.

For a fixed t, let $H_k(\boldsymbol{\sigma}^2, \alpha, G) = R_k(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha))$ and let $H_k(\mathbf{Z}, \alpha, G)$ be the functions with $b(\boldsymbol{\sigma}^2, k\alpha t)$ in $H_k(\boldsymbol{\sigma}^2, \alpha, G)$ replaced by $w(k\alpha t) b(\mathbf{Z}, k\alpha t)$, $k = 1, 2$. Denote $\alpha_k^*(G) = \arg\min_{\alpha \in [0,1]} H_k(\boldsymbol{\sigma}^2, \alpha, G)$ and $\breve{\alpha}_k^*(G) = \arg\min_{\alpha \in [0,1]} H_k(\mathbf{Z}, \alpha, G)$.

Theorem 6. For any fixed nonzero t, we make the following statements:

(a) When $\nu > 2|t|$, $\mathrm{E}(\sigma_1^2)^{-t} < \infty$, and $\mathrm{E}(\ln(\sigma_1^2)) < \infty$, we have $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\breve{\alpha}_1^*(G))) - R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha_1^*(G))) \xrightarrow{\text{a.s.}} 0$ and $\breve{\alpha}_1^*(G) - \alpha_1^*(G) \xrightarrow{\text{a.s.}} 0$ as $G \to \infty$.
(b) When $\nu > 4|t|$, $\mathrm{E}(\sigma_1^2)^{-2t} < \infty$ and $\mathrm{E}(\ln(\sigma_1^2)) < \infty$, we have $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\breve{\alpha}_2^*(G))) - R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha_2^*(G))) \xrightarrow{\text{a.s.}} 0$ as $G \to \infty$.

Proofs of Lemma 2 and Theorem 6 are in Appendix. Note that there is no corresponding consistent result for α_2^* because it may not be unique. For the special case that F is a gamma distribution with shape parameter γ and scale parameter β, it is easy to check that $\mathrm{E}(\ln(\sigma_1^2)) = \Psi(\gamma) + \ln\beta < \infty$, $\mathrm{E}(\sigma_1^2)^{-t} = \beta^{-t}\Gamma(\gamma - t)/\Gamma(\gamma) < \infty$ for $\gamma > t$, and $\mathrm{E}(\sigma_1^2)^{-2t} = \beta^{-2t}\Gamma(\gamma - 2t)/\Gamma(\gamma) < \infty$ for $\gamma > 2t$.

Note that Theorem 6 does not apply for small ν. We propose an alternative two-step procedure: (1) substitute $b(\boldsymbol{\sigma}^2, \eta)$ in (8) and (9) by $b(\mathbf{Z}, \eta)$, and compute temporary optimal shrinkage parameters and the corresponding shrinkage estimators, say $\hat{\boldsymbol{\sigma}}^2_-$; (2) substitute $b(\boldsymbol{\sigma}^2, \eta)$ in (8) and (9) by $b(\hat{\boldsymbol{\sigma}}^2_-, \eta)$ to get the final optimal shrinkage parameters $\breve{\alpha}_1^*$ and $\breve{\alpha}_2^*$. When $t > 0$, because σ_g^2 appears in the denominator in (8) and (9), extreme small values of Z_g make estimates of α_1^* and α_2^* unstable. We truncate the smallest 1% of Z_g's in our procedures. We find that the truncation is unnecessary when $t < 0$. Simulations indicate that $\breve{\alpha}_1^*$ and $\breve{\alpha}_2^*$ perform better than $\hat{\alpha}_1^*$, $\hat{\alpha}_2^*$, $\breve{\alpha}_1^*$, and $\breve{\alpha}_2^*$ when ν is small. Therefore, $\breve{\alpha}_1^*$ and $\breve{\alpha}_2^*$ will be used in our simulations.

Computation of optimal shrinkage parameters amounts to finding minimizers of some loss functions in the interval $[0, 1]$. Optimization methods such as quasi-Newton or conjugate-gradient algorithms may be used. Because the computations involved are cheap and fast, for simplicity, we use grid search in the following data analysis and simulations. The R code is available from the authors.

4. SIMULATIONS AND COMPARISONS

In this section we conduct simulations to compare the performance of our estimators with the CHQBC estimator and the modified CHQBC estimator for the purpose of estimation. All estimators considered in this section perform substantially better than the standard gene-specific estimator. For simplicity, we will not present the results for the gene-specific estimator. We evaluate the performance for estimating σ_g^2 in Section 4.1 and the performance for estimating $(\sigma_g^2)^{-1}$ in Section 4.2. We set $G = 5{,}000$ in this section.

4.1 Estimation of σ_g^2

We consider four different estimators in this subsection: $\hat{\sigma}_g^2(\breve{\alpha}_1^*)$, $\hat{\sigma}_g^2(\breve{\alpha}_2^*)$, $\hat{\sigma}_g^2(\hat{\alpha}_0)$, and $\tilde{\sigma}_g^2(\hat{\alpha}_0)$. We simulate σ_g^2, $g = 1, \ldots, G$, from a gamma distribution with shape parameter γ and scale parameter β. We set $\beta = 1$ because it has little impact on the comparative performance. To evaluate performance under different levels of variance heterogeneity, we consider three different shape parameters, $\gamma = .25$, 1, and 4, which correspond to three different coefficients of variation [$\mathrm{CV} = \sqrt{\gamma\beta^2}/(\gamma\beta) = \sqrt{\gamma}/\gamma$] at levels 2, 1, and 0.5, respectively. For each σ_g^2, we simulate $\nu + 1$ observations from $\mathrm{N}(\mu_g, \sigma_g^2)$, where μ_g is a random sample from $\mathrm{N}(0, 1)$. We calculate Z_g as the sample variance for each g. We use a factorial design that consists of three levels for γ and seven levels for ν, $\nu = 1, \ldots, 7$. Therefore, we have 21 combinations of parameter settings. For each setting, we repeat simulation 100 times. We compute the average risk

$$\mathrm{AR}_k = \frac{1}{100G} \sum_{r=1}^{100} \sum_{g=1}^{G} L_k(\sigma_{gr}^2, \hat{\sigma}_{gr}^2), \qquad k = 1, 2,$$

where r represents simulation replications, and $k = 1$ and $k = 2$ correspond to the Stein and the squared loss functions, respectively. We plot $\ln(\mathrm{AR}_k)$ in Figure 1 as a function of ν for all four methods.

Standard errors of these log average risks range from .00007 to .01871 under the Stein loss function and from .00040 to .07398 under the squared loss function. Therefore, most of the differences in log average risks are statistically significant. The modified CHQBC estimator $\hat{\sigma}_g^2(\hat{\alpha}_0)$ has smaller average risk than the original CHQBC estimator $\tilde{\sigma}_g^2(\hat{\alpha}_0)$ in all settings. When the variance heterogeneity is not small ($\gamma = .25$ and $\gamma = 1$), two optimal estimators $\hat{\sigma}_g^2(\hat{\alpha}_1^*)$ and $\hat{\sigma}_g^2(\breve{\alpha}_2^*)$ have similar risks, which are smaller than those of $\hat{\sigma}_g^2(\hat{\alpha}_0)$. When ν and

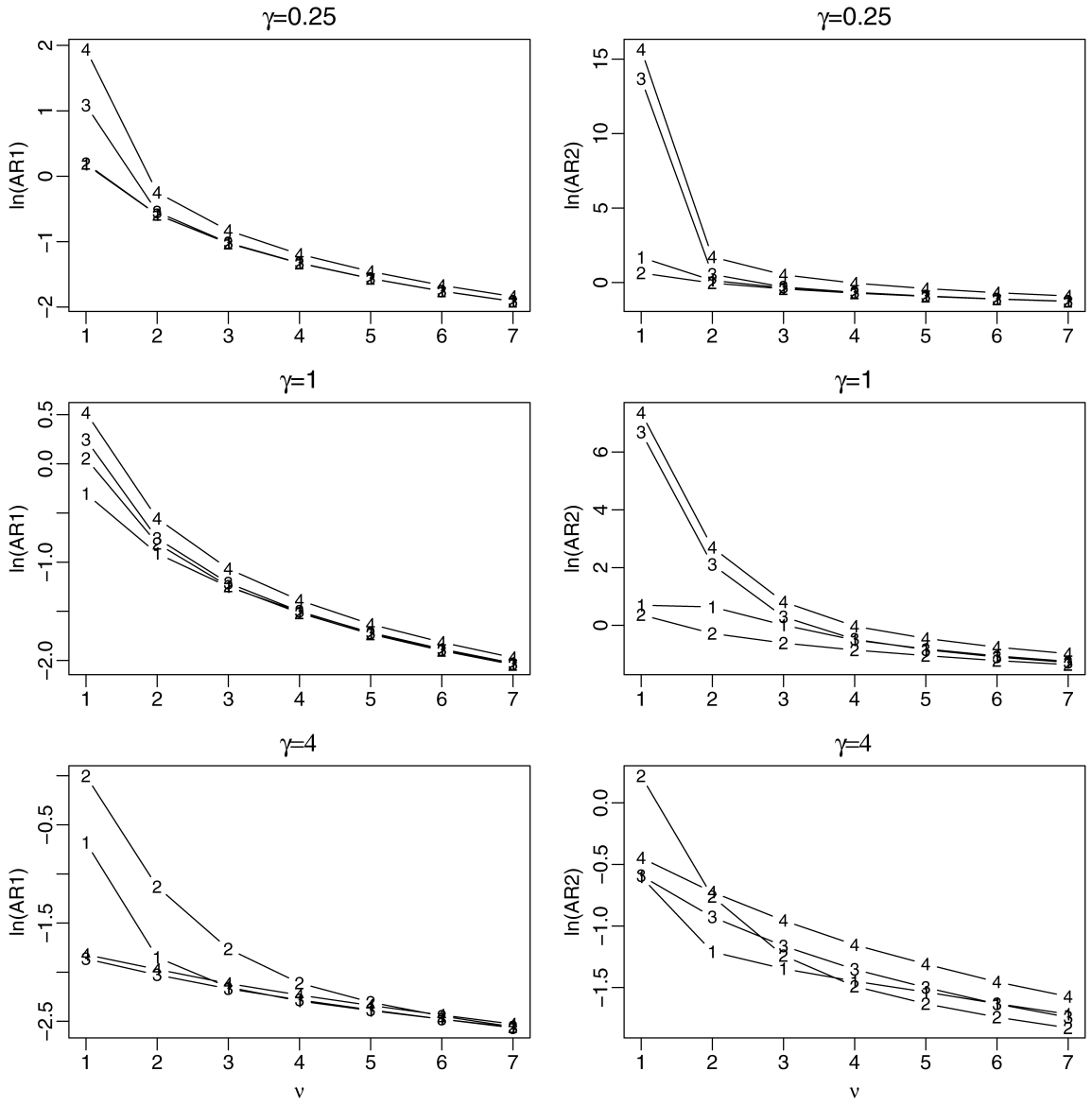

Figure 1. Plots of Log Average Risks for Estimating σ_g^2 Under the Stein Loss Function (left) and the Squared Loss Function (right). The three rows correspond to the three shape parameters. The lines in each plot marked 1, 2, 3, and 4 correspond to the optimal estimator under the Stein loss function $\hat{\sigma}_g^2(\check{\alpha}_1^*)$, the optimal estimator under the squared loss function $\hat{\sigma}_g^2(\check{\alpha}_2^*)$, the modified CHQBC estimator $\hat{\sigma}_g^2(\hat{\alpha}_0)$, and the CHQBC estimator $\tilde{\sigma}_g^2(\hat{\alpha}_0)$, respectively.

the variance heterogeneity is small ($\gamma = 4$), $\hat{\sigma}_g^2(\hat{\alpha}_0)$ has smaller risks under the Stein loss function than the optimal estimators. Overall, the optimal estimator under the Stein loss function performs well. Note that the estimates of optimal shrinkage parameters $\check{\alpha}_1^*$ and $\check{\alpha}_2^*$ do not guarantee the optimal performance, especially when ν is small. Simulations (not shown) indicate that α_1^* and α_2^* do guarantee the optimal performance for the Stein and the squared loss functions, respectively.

4.2 Estimation of σ_g^{-2}

As discussed in Section 1, it is better to use an estimator of σ_g^{-2} directly than to use the reciprocal of an estimator of σ_g^2 in the F test (Sec. 5). In this subsection we evaluate performance for estimating σ_g^{-2}.

Simulations (not shown) indicate that $\hat{\sigma}_g^{-2}(\check{\alpha}_1^*)$ always performs better than $\hat{\sigma}_g^{-2}(\check{\alpha}_2^*)$. For simplicity, we present results for $\hat{\sigma}_g^{-2}(\check{\alpha}_1^*)$ only. We consider four estimators, $\hat{\sigma}_g^{-2}(\check{\alpha}_1^*)$, $\hat{\sigma}_g^2(\check{\alpha}_1^*)$, $\hat{\sigma}_g^2(\hat{\alpha}_0)$, and $\tilde{\sigma}_g^2(\hat{\alpha}_0)$, where $\hat{\sigma}_g^{-2}(\check{\alpha}_1^*)$ is the estimator of $(\sigma_g^2)^t$ with $t = -1$. We take the reciprocal of the last three as estimators of σ_g^{-2}. Theorem 1 requires that $\nu \geq 3$ when $t = -1$. Therefore, we take ν from 3 to 9. All other settings are the same as those in Section 4.1.

Standard errors of the log average risks range from .00017 to .00678 under the Stein loss function and from .00038 to .04529 under the squared loss function. Again, most of the differences in log average risk are statistically significant. Figure 2 shows that under both the Stein and the squared loss functions, risk of $\hat{\sigma}_g^{-2}(\check{\alpha}_1^*)$ < risk of $\tilde{\sigma}_g^2(\hat{\alpha}_0)$ < risk of $\hat{\sigma}_g^2(\hat{\alpha}_0)$ < risk of $\hat{\sigma}_g^2(\check{\alpha}_1^*)$. It is interesting to note that $\tilde{\sigma}_g^2(\hat{\alpha}_0)$ outperforms $\hat{\sigma}_g^2(\hat{\alpha}_0)$, which, again, confirms that a better estimator for σ_g^2 may not lead to a better estimator for σ_g^{-2}. We have performed many more simulations with different parameters for both Sec-

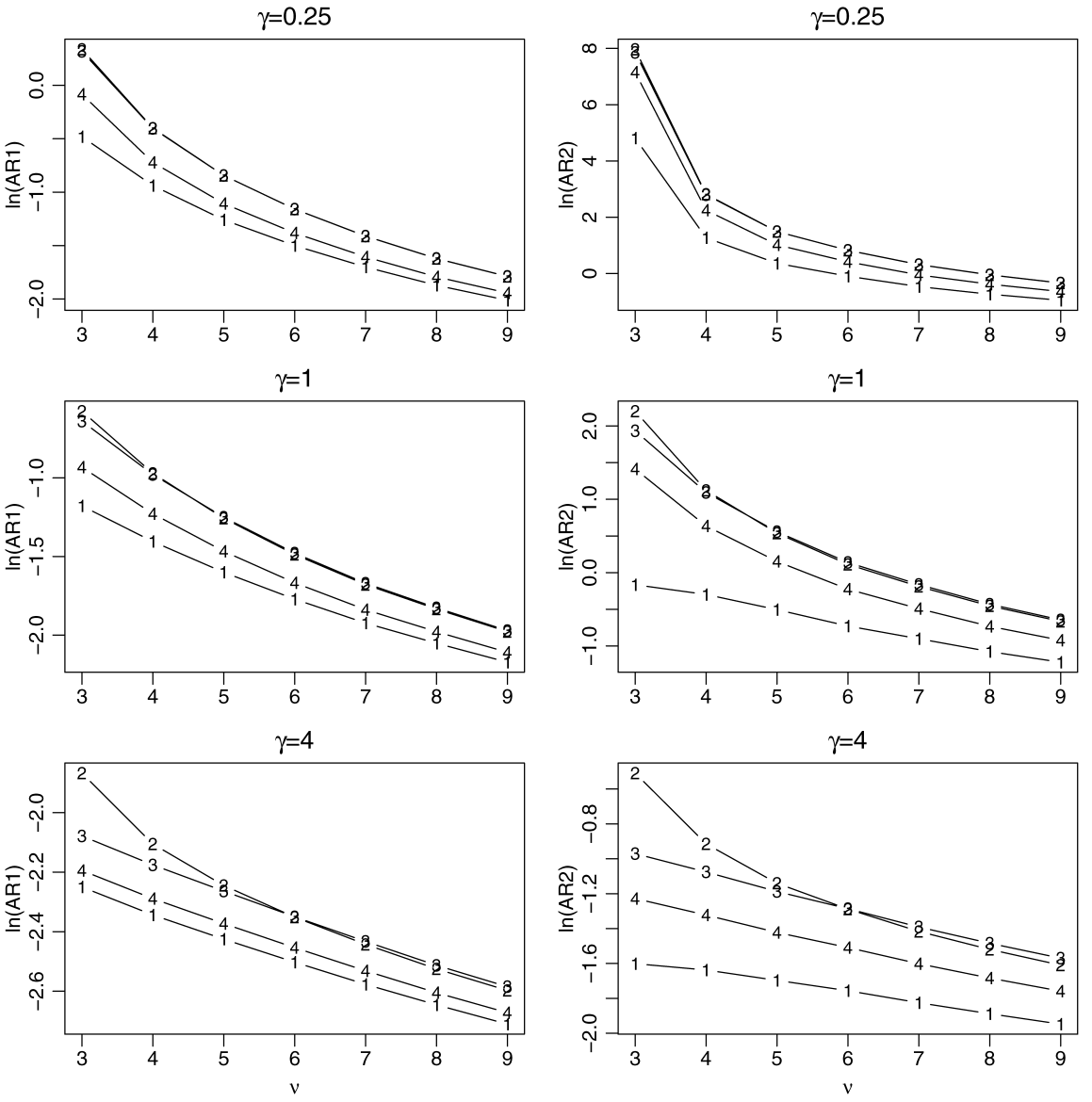

Figure 2. Plots of Log Average Risks for Estimating $(\sigma_g^2)^{-1}$ Under the Stein Loss Function (left) and the Squared Loss Function (right). The three rows correspond to the three shape parameters. The lines in each plot marked 1, 2, 3, and 4 correspond to $\hat{\sigma}_g^{-2}(\breve{\alpha}_1^*)$, $\hat{\sigma}_g^2(\breve{\alpha}_1^*)$, $\hat{\sigma}_g^2(\hat{\alpha}_0)$, and $\tilde{\sigma}_g^2(\hat{\alpha}_0)$, respectively.

tions 4.1 and 4.2. Comparative results remain the same. More insights on these estimators can be found in Tong and Wang (2005).

5. APPLICATIONS

Cui et al. (2005) demonstrated that the F test using the CHQBC estimator has the best or nearly the best power among several "information-sharing" statistics for detecting differentially expressed genes over a wide range of settings. For simplicity, we compare the performance of our shrinkage estimators with the CHQBC estimator, the modified CHQBC estimator, and the gene-specific estimator only. We introduce F-like statistics using shrinkage estimators in Section 5.1. In Section 5.2 we apply our method to microarray data, and conduct simulations to evaluate and compare the performances of several F-like statistics.

5.1 F-Like Statistics

For a fixed gene g, $g = 1, \ldots, G$, test statistics are usually based on the general linear mixed effects model (Kerr et al. 2002; Cui et al. 2005)

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta}_g + \mathbf{Z}_g\mathbf{b}_g + \boldsymbol{\epsilon}_g, \tag{11}$$

where $\mathbf{y}_g$ is the vector of all observations for gene g, $\mathbf{X}_g$ and $\mathbf{Z}_g$ are design matrices for the fixed effects $\boldsymbol{\beta}_g$ and the random effects $\mathbf{b}_g$, respectively, and $\boldsymbol{\epsilon}_g$ is the vector of random errors. We assume that

$$\begin{pmatrix}\mathbf{b}_g\\ \boldsymbol{\epsilon}_g\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \sigma_g^2\begin{pmatrix}\mathbf{G}_g & \\ & \mathbf{R}_g\end{pmatrix}\right),$$

where σ_g^2 is the error variance for gene g. Note that the general linear model is a special case with empty $\mathbf{Z}_g$ and $\mathbf{b}_g$. Denote $\hat{\boldsymbol{\beta}}_g$ and $\hat{\mathbf{b}}_g$ as the best linear unbiased predictor and denote its

variance–covariance matrix

$$\mathbf{C}_g = \sigma_g^2 \begin{pmatrix} \mathbf{X}_g^T\mathbf{R}_g^{-1}\mathbf{X}_g & \mathbf{X}_g^T\mathbf{R}_g^{-1}\mathbf{Z}_g \\ \mathbf{Z}_g^T\mathbf{R}_g^{-1}\mathbf{X}_g & \mathbf{Z}_g^T\mathbf{R}_g^{-1}\mathbf{Z}_g + \mathbf{G}_g^{-1} \end{pmatrix}^- \triangleq \sigma_g^2\mathbf{D}_g,$$

where the negative sign represents the generalized inverse of the matrix. The F statistic for testing the hypothesis H_0: $\mathbf{L}_g^T(\boldsymbol{\beta}_g^T, \mathbf{b}_g^T)^T = 0$ is (Littell, Milliken, Stroup, and Wolfinger 1996)

$$F = \frac{(\hat{\boldsymbol{\beta}}_g^T, \hat{\mathbf{b}}_g^T)\mathbf{L}_g(\mathbf{L}_g^T\hat{\mathbf{D}}_g\mathbf{L}_g)^{-1}\mathbf{L}_g^T(\hat{\boldsymbol{\beta}}_g^T, \hat{\mathbf{b}}_g^T)^T/\operatorname{rank}(\mathbf{L}_g)}{\hat{\sigma}_g^2}$$

$$\triangleq \Delta_g\hat{\sigma}_g^{-2}, \tag{12}$$

where $\hat{\mathbf{D}}_g$ is an estimator of $\mathbf{D}_g$ that can be calculated using the restricted maximum likelihood method (Searle, Casella, and McCulloch 1992) and $\hat{\sigma}_g^{-2}$ is an estimator of σ_g^{-2}.

Different estimators of σ_g^{-2} lead to different F-like statistics. We will consider four F-like statistics, F_1, F_2, F_3, and F_4, with $\hat{\sigma}_g^{-2}$ in (12) replaced by $\hat{\sigma}_g^{-2}(\breve{\alpha}_1^*)$, $\hat{\sigma}_g^{-2}(\breve{\alpha}_2^*)$, $1/\tilde{\sigma}_g^2(\hat{\alpha}_0)$, and the gene-specific estimator, respectively. The statistic F_3 is the same as F_s in Cui et al. (2005). Simulations (not shown) indicate that F-like statistics with $\hat{\sigma}_g^2$ in (12) replaced by $\hat{\sigma}_g^2(\hat{\alpha}_0)$, $\hat{\sigma}_g^2(\breve{\alpha}_1^*)$, or $\hat{\sigma}_g^2(\breve{\alpha}_2^*)$ have similar performance as F_3. To save space, these results are not presented. Because F-like statistics do not follow F distributions, we calculate p values by permutation as in Cui et al. (2005).

We note that the method described here applies to general designs, and usually $\mathbf{X}_g$, $\mathbf{Z}_g$, $\mathbf{G}_g$, $\mathbf{R}_g$, and $\mathbf{L}_g$ are independent of g. The most commonly used design in practice is perhaps the simple k sample comparison with

$$y_{gij} = \mu_{gi} + \epsilon_{gij}, \qquad i = 1, \ldots, k; j = 1, \ldots, n,$$

where μ_{gi} is the mean of sample i and gene g, and $\epsilon_{gij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_g^2)$. The standard F statistic for testing the hypothesis H_0: $\mu_{g1} = \cdots = \mu_{gk}$ is $F = \Delta_g\hat{\sigma}_g^{-2}$, where $\Delta_g = n\sum_{i=1}^k(\bar{y}_{gi\cdot} - \bar{y}_{g\cdot\cdot})^2/(k-1)$, $\hat{\sigma}_g^2 = \sum_{i=1}^k\sum_{j=1}^n(y_{gij} - \bar{y}_{gi\cdot})^2/k(n-1)$, $\bar{y}_{gi\cdot} = \sum_{j=1}^n y_{gij}/n$, and $\bar{y}_{g\cdot\cdot} = \sum_{i=1}^k\sum_{j=1}^n y_{gij}/kn$. Shrinkage estimates of σ_g^{-2} are obtained with $Z_g = \hat{\sigma}_g^2$.

5.2 Case Study

Cui et al. (2005) described an experiment that compared two human colon cancer cell lines, CACO2 and HCT116, and three human ovarian cancer cell lines, ES2, MDAH2774, and OV1063. Five samples were arranged in a loop and no reference sample was used. Fluorescent dye labeled cDNA targets were hybridized to cDNA microarrays that contained 9,600 human cDNA clones. To simplify the analysis, as in Cui et al. (2005), the duplicated spots for the same gene on each array were averaged at the original signal level. Observations were transformed and normalized.

To each gene we fitted the model (Cui et al. 2005)

$$y_{ij} = \mu + A_i + D_j + S_{k(i,j)} + \epsilon_{ij},$$
$$i = 1, \ldots, 10; j = 1, 2; k = 1, \ldots, 5, \tag{13}$$

where μ is the gene mean, A_i is the array effect, D_j is the dye effect, $S_{k(i,j)}$ is the sample effect, and ϵ_{ij} is the random error. In general, the terms μ, D_j, and $S_{k(i,j)}$ are fixed effects and A_i are random effects. Cui et al. (2005) demonstrated that the array variance has little impact on the F tests. Therefore, as in Cui et al. (2005), we treat A_i as fixed effects. All analyses were conducted using the newest version of R/MAANOVA (Wu and Churchill 2005). At a nominal significance level of .01, F_1, F_2, F_3, and F_4 detected 1,859, 1,849, 1,823, and 1,439 significant genes, respectively. To overcome the problem of multiple comparisons, we applied the Benjamini and Hochberg (1995) procedure to control false discovery rate (FDR). At a FDR level of .05, F_1, F_2, F_3, and F_4 detected 1,823, 1,806, 1,769, and 1,002 significant genes, respectively.

To study the false positive and successful detection rates for these four F-like tests, we simulated 100 datasets using the same design as the real data. Each simulated data set contains 1,000 constant genes and 1,000 differentially expressed genes. Because the successful detection rate of a test depends on the magnitude of the overall treatment effect, $\Theta = \sum_{k=1}^5(S_k - \bar{S}_\cdot)^2/4$, where S_k is defined in (13), we considered $\sqrt{4\Theta}$ as a parameter. Specifically, we generated $S_k = \sqrt{4\Theta}(Q_k - \bar{Q})/\sqrt{\sum_{k=1}^5(Q_k - \bar{Q})^2}$, where $Q_k \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, with $\sqrt{4\Theta} = .1, .2, \ldots, 1$ representing 10 different levels of treatment effects. The fixed effects μ and D_j were drawn randomly from the normal distributions $\mathrm{N}(0, .65^2)$ and $\mathrm{N}(0, .35^2)$, respectively, and were held constant across all simulations. For each simulation, independent A_i and ϵ_{ij} were drawn randomly from $\mathrm{N}(0, .6^2)$ and $\mathrm{N}(0, \sigma_g^2)$, respectively, where σ_g^2 are sampled randomly without replacement from the 9,600 estimates of residual variances of the real data. As in Cui et al. (2005), the variability of the residual variances was controlled by a parameter τ through the formula $\sigma_{g,\tau}^2 = (\sigma_g^{2\tau}/\sigma_{\text{pool}}^{2\tau})\sigma_{\text{pool}}^2$. We considered three choices of τ, $\tau = .5$, 1, and 2, which correspond to CV = .63, 1.84, and 10.60, respectively. Random errors were generated such that they are mutually independent.

Average false positive rates at the significance level .05 and their standard errors are listed in Table 1. False positive rates for all four F-like statistics are under the nominal value, and are smaller for tests based on shrinkage variance estimators. Average false positive rates at other significance levels behave similarly.

All tests based on shrinkage variance estimators, F_1–F_3, have much larger powers than those of the gene-specific test F_4. Improvements of the new F-like tests come with moderate to large treatment effect. For simplicity, in Figure 3, we plot the successful detection rates for F_1, F_2, and F_3 only with $.4 \le \sqrt{4\Theta} \le 1$. Standard errors of these detection rates (not shown) range from .00001 to .00249. Most of the differences in detection rates are

Table 1. Average False Positive Rates and Standard Errors (inside parentheses) of Four F-Like Statistics at the Significance Level .05

Parameter	F_1	F_2	F_3	F_4
CV = .63	$.028_{(.0008)}$	$.029_{(.0008)}$	$.029_{(.0007)}$	$.048_{(.0009)}$
CV = 1.84	$.041_{(.0009)}$	$.041_{(.0009)}$	$.043_{(.0009)}$	$.048_{(.0009)}$
CV = 10.6	$.046_{(.0009)}$	$.045_{(.0008)}$	$.047_{(.0009)}$	$.048_{(.0009)}$
$\rho = 0$	$.036_{(.0009)}$	$.036_{(.0009)}$	$.039_{(.0009)}$	$.048_{(.0009)}$
$\rho = 1/3$	$.034_{(.0014)}$	$.034_{(.0014)}$	$.037_{(.0015)}$	$.046_{(.0015)}$
$\rho = 2/3$	$.035_{(.0026)}$	$.034_{(.0026)}$	$.037_{(.0028)}$	$.046_{(.0027)}$
$\rho = 1$	$.037_{(.0047)}$	$.036_{(.0047)}$	$.039_{(.0051)}$	$.047_{(.0052)}$

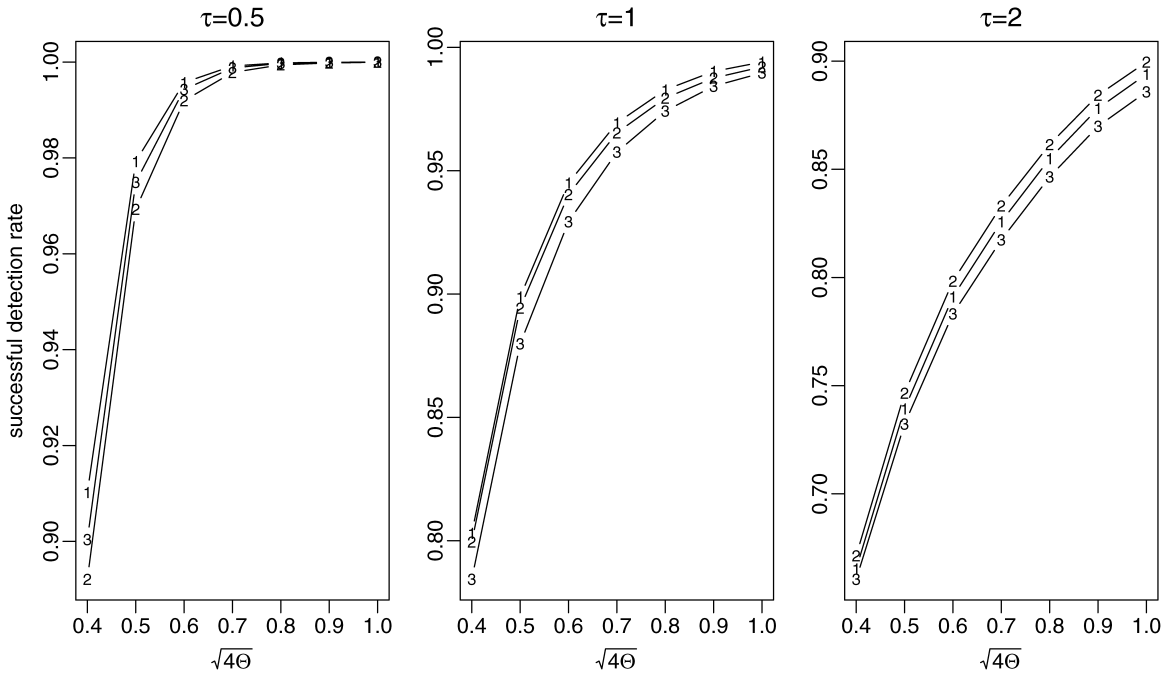

Figure 3. Plots of the Average Successful Detection Rates for Simulations Where Observations Are Independent. The lines in each plot marked 1, 2, and 3 correspond to F_1, F_2, and F_3, respectively.

statistically significant. The statistic F_1 outperforms F_3 in all situations and also performs better than F_2 when the heterogeneity of variance is not large (CV = .63 and CV = 1.84). When the heterogeneity of variance is large (CV = 10.6), F_2 performs better than F_1. For microarray data in their typical range of CV, we recommend F_1.

The preceding simulations assume that the genes are mutually independent, which is unlikely to hold in practice. To evaluate and compare the performance of various tests under more realistic situations, we simulated another 100 datasets using the same design as the real data, model (13), and estimated covariance matrix. Again, each simulated dataset contains 1,000 constant genes and 1,000 differentially expressed genes. We assumed model (13), where μ, A_i, D_j, and S_k are generated in the same manner as the previous simulation. Random errors from different genes are now correlated. To generate correlated random errors, we first computed the sample covariance matrix Σ, based on residuals from analysis of the real data with 9,600 genes. Then we treated these 9,600 genes as the population and treated the 2,000 genes in each simulation replication as a random sample from the population. After 2,000 genes are chosen, the covariance matrix of these 2,000 genes is a submatrix of Σ, say Σ_1, with rows and columns that correspond to the selected genes. To investigate the effect of different levels of correlation, we considered the covariance matrix $\Sigma_1(\rho)$ with off-diagonal elements of Σ_1 multiplied by a number ρ. We considered four different values, $\rho = 0, 1/3, 2/3$, and 1, where $\rho = 0$ corresponds to the independent situation in the previous simulation. Finally, for each fixed i and j, we generated the vector of random errors for all 2,000 genes according to a multivariate normal distribution with mean zero and covariance matrix $\Sigma_1(\rho)$.

Average false positive rates at the significance level .05 and their standard errors are also listed in Table 1. Again, all four F-like statistics have false positive rates that are under the nominal value. The F-like statistics based on shrinkage variance estimators have smaller false positive rates than those based on the gene-specific test. Detection rates are shown in Figure 4. Standard errors of these detection rates (not shown) range from .0006 to .011, and most of the differences in detection rates are statistically significant. We conclude that the correlation has little effect on power when the treatment effect is large. When the treatment effect is moderate, the power decreases slightly as the correlation increases for all shrinkage variance estimator based tests. Nevertheless, the comparative results remain the same: all shrinkage variance estimator based tests have much larger powers than those of the gene-specific test, and F_1 and F_2 outperform F_3 in all situations.

For simulations under both independent and correlated situations, we apply the Benjamini and Hochberg procedure with various FDR levels. All four F-like statistics have actual FDRs that are under the nominal value, and F-like statistics based on shrinkage variance estimators have smaller actual FDRs than those based on the gene-specific test. The performance differences among four F-like statistics remain the same.

6. DISCUSSION

One major challenge in microarray data analysis is the relatively small number of replications for each gene compared to the large number of genes. In this article we propose a family of shrinkage variance estimators that borrow information across genes by shrinking each gene-specific variance estimator toward the bias corrected geometric mean of variance estimators for all genes. The amount of optimal shrinkage depends on the variability of the individual variances. We have shown that the standard sample variance is inadmissible under either the Stein or the squared loss function. Our optimal shrinkage estimators compare favorably with the CHQBC estimator in terms of both

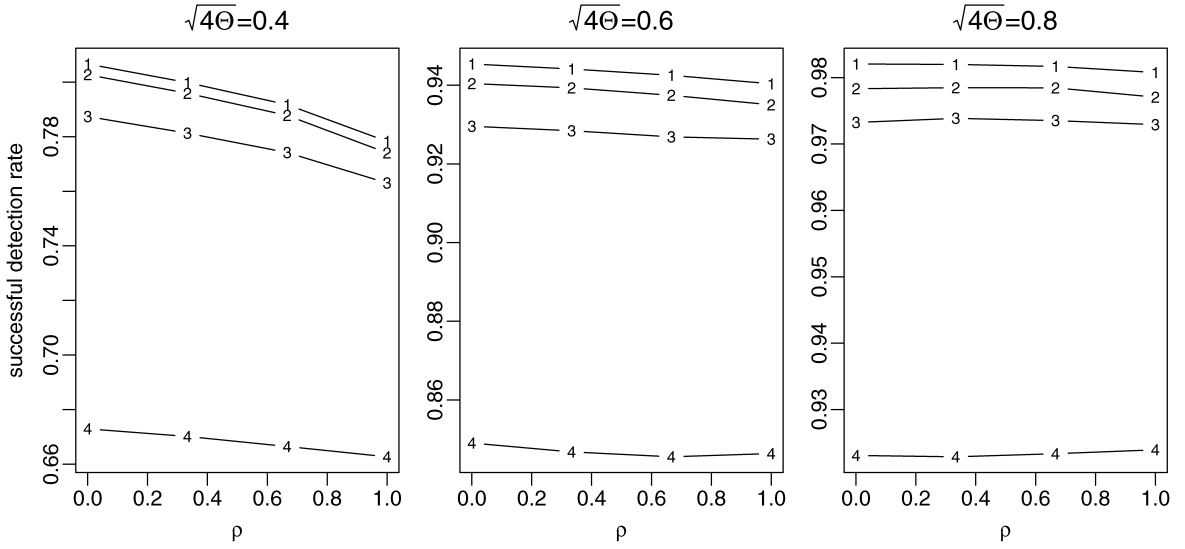

Figure 4. Plots of the Average Successful Detection Rates for Simulations Where Observations Are Correlated. The lines in each plot marked 1, 2, 3, and 4 correspond to F_1, F_2, F_3, and F_4, respectively.

estimation and hypothesis test. One key is to use an estimator of σ_g^{-2} directly in the F statistic instead of using the reciprocal of an estimator of σ_g^2. We note that the optimal shrinkage variance estimators may also be used for other purposes such as constructing confidence intervals. We recommend the estimator under the Stein loss function for microarray data analysis.

On the logarithm scale, our shrinkage estimator (7) is a weighted average of the gene-specific variance and the bias corrected geometric mean. One of our future research topic is to consider a weighted average of the gene-specific variance and the arithmetic mean (Baldi and Long 2001). Our method shrinks all gene-specific variances to a unique common variance. One may also borrow information contained in the sample mean (Stein 1964). Better shrinkage estimators may be constructed when additional information becomes available. For example, instead of shrinking to the overall geometric mean, one may borrow information from neighboring genes (Baldi and Long 2001; Jain et al. 2003).

APPENDIX: PROOFS

Proof of Lemma 2

Because $s(\alpha) = (\sigma_g^2)^{-\alpha t}$ is a convex function of α, then $(\sigma_g^2)^{-\alpha t} = s(\alpha) \le (1-\alpha)s(0) + \alpha s(1) \le 1 + (\sigma_g^2)^{-t}$. By theorem 16(a) in Ferguson (1996) and the fact that $\mathrm{E}(\sigma_g^2)^{-t} < \infty$,

$$\frac{1}{G}\sum_{g=1}^{G}(\sigma_g^2)^{-\alpha t} \xrightarrow{\text{a.s.}} \mathrm{E}(\sigma_1^2)^{-\alpha t}$$

uniformly for $\alpha \in [0,1]$ as $G \to \infty$. (A.1)

Similarly, by the fact that $\mathrm{E}(Z_1)^{-\alpha t} = \mathrm{E}[\mathrm{E}(Z_1^{-\alpha t}|\sigma_1^2)] = \mathrm{E}(\sigma_1^2)^{-\alpha t}/h_1(-\alpha t)$,

$$\frac{h_1(-\alpha t)}{G}\sum_{g=1}^{G}(Z_g)^{-\alpha t} \xrightarrow{\text{a.s.}} \mathrm{E}(\sigma_1^2)^{-\alpha t}$$

uniformly for $\alpha \in [0,1]$ as $G \to \infty$. (A.2)

Combining (A.1) and (A.2) yields

$$\frac{h_1(-\alpha t)}{G}\sum_{g=1}^{G}(Z_g)^{-\alpha t} - \frac{1}{G}\sum_{g=1}^{G}(\sigma_g^2)^{-\alpha t} \xrightarrow{\text{a.s.}} 0$$

uniformly for $\alpha \in [0,1]$ as $G \to \infty$. (A.3)

By the strong law of large numbers, $\ln(\sigma_{\text{pool}}^2) = \sum_{g=1}^{G}\ln(\sigma_g^2)/G \xrightarrow{\text{a.s.}} \mathrm{E}\ln(\sigma_1^2)$. Thus $[\sigma_{\text{pool}}^2\exp(-\mathrm{E}\ln(\sigma_1^2))]^t \xrightarrow{\text{a.s.}} 1$ for any fixed t. Following the arguments similar to those in Tong and Wang (2005) and using the fact that $\mathrm{E}(\ln(Z_1)) = \mathrm{E}[\mathrm{E}(\ln(Z_1)|\sigma_1^2)] = \mathrm{E}\ln(\sigma_1^2) + \Psi(\nu/2) - \ln(\nu/2)$, as $G \to \infty$, we have

$$(\sigma_{\text{pool}}^2)^{\alpha t} \xrightarrow{\text{a.s.}} \exp(\alpha t\,\mathrm{E}\ln(\sigma_1^2)) \quad \text{uniformly for } \alpha \in [0,1], \tag{A.4}$$

and

$$\left(\frac{\nu}{2}\right)^{\alpha t}\exp\left(-\alpha t\Psi\left(\frac{\nu}{2}\right)\right)(Z_{\text{pool}})^{\alpha t} \xrightarrow{\text{a.s.}} \exp(\alpha t\,\mathrm{E}\ln(\sigma_1^2))$$

uniformly for $\alpha \in [0,1]$. (A.5)

Combining (A.4) and (A.5), as $G \to \infty$, gives

$$\left(\frac{\nu}{2}\right)^{\alpha t}\exp\left(-\alpha t\Psi\left(\frac{\nu}{2}\right)\right)(Z_{\text{pool}})^{\alpha t} - (\sigma_{\text{pool}}^2)^{\alpha t} \xrightarrow{\text{a.s.}} 0$$

uniformly for $\alpha \in [0,1]$. (A.6)

Using (A.2), (A.3), (A.4), and (A.6), we have

$$\begin{aligned} & w(\alpha t)b(\mathbf{Z},\alpha t) - b(\boldsymbol{\sigma}^2,\alpha t) \\ &= \left[\left(\frac{\nu}{2}\right)^{\alpha t}\exp\left(-\alpha t\Psi\left(\frac{\nu}{2}\right)\right)(Z_{\text{pool}})^{\alpha t} - (\sigma_{\text{pool}}^2)^{\alpha t}\right] \\ &\quad \times \frac{h_1(-\alpha t)}{G}\sum_{g=1}^{G}(Z_g)^{-\alpha t} \\ &\quad - (\sigma_{\text{pool}}^2)^{\alpha t}\left[\frac{h_1(-\alpha t)}{G}\sum_{g=1}^{G}(Z_g)^{-\alpha t} - \frac{1}{G}\sum_{g=1}^{G}(\sigma_g^2)^{-\alpha t}\right] \\ &\xrightarrow{\text{a.s.}} 0 \quad \text{uniformly for } \alpha \in [0,1] \text{ as } G \to \infty. \end{aligned}$$

Proof of Theorem 6

(a) By Lemma 2, it is not difficult to check that

$$H_1(\mathbf{Z}, \alpha, G) - H_1(\boldsymbol{\sigma}^2, \alpha, G) \xrightarrow{\text{a.s.}} 0$$

$$\text{uniformly for } \alpha \in [0, 1] \text{ as } G \to \infty. \quad \text{(A.7)}$$

Following arguments similar to those in Tong and Wang (2005) and using the fact that $\max_{\alpha\in[0,1]} \lim_{G\to\infty} C_1(\alpha) < \infty$, where $C_1(\alpha) = h_G^{\alpha}(t)h_1^{1-\alpha}(t)/(h_1^{G-1}(\alpha t/G)h_1((1-\alpha+\alpha/G)t))$, we have $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\breve{\alpha}_1^*(G))) - R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha_1^*(G))) \xrightarrow{\text{a.s.}} 0$.

Now we show for each pair $(\boldsymbol{\sigma}^2, \mathbf{Z})$ that satisfies (A.7), that $\breve{\alpha}_1^*(G) - \alpha_1^*(G) \to 0$. Because that there exists a pair $(\boldsymbol{\sigma}^2, \mathbf{Z})$ for which (A.7) holds and $\breve{\alpha}_1^*(G) - \alpha_1^*(G) \nrightarrow 0$. Because α belongs to the compact interval $[0, 1]$, there exists a subsequence $\{G_n\}$ such that $|\breve{\alpha}_1^*(G_n) - \alpha_1^*(G_n)| \to \beta > 0$. It can be shown that (Tong and Wang 2005)

$$\underline{\lim}_{G\to\infty} \frac{\partial^2 H_1(\boldsymbol{\sigma}^2, \alpha, G)}{\partial\alpha^2} \geq \frac{\Gamma(\nu/2 + (1-\alpha)t)}{\Gamma^{\alpha}(\nu/2)\Gamma^{1-\alpha}(\nu/2+t)} t^2 \Psi'\left(\frac{\nu}{2} + (1-\alpha)t\right).$$

Because the nonzero $t > -\nu/2$, it is not difficult to show that $\delta \triangleq \min_{\alpha\in[0,1]}\{\lim_{G\to\infty} C_1(\alpha)D'(\alpha)\} > 0$, where $D'(\alpha) = ((G-1)t^2(\Psi'(\nu/2+\alpha t/G) + (G-1)\Psi'(\nu/2+(1-\alpha+\alpha/G)t))/G^2$. Then for an arbitrary $0 < \epsilon < \delta\beta^2/16$, there exists an $N_1 > 0$ such that for any $G_n > N_1$, $|\breve{\alpha}_1^*(G_n) - \alpha_1^*(G_n)| > \beta/2$. Consequently,

$$\begin{aligned} &H_1(\boldsymbol{\sigma}^2, \breve{\alpha}_1^*(G_n), G_n) - H_1(\boldsymbol{\sigma}^2, \alpha_1^*(G_n), G_n) \\ &= \frac{1}{2}\frac{\partial^2 H_1(\boldsymbol{\sigma}^2, \alpha, G_n)}{\partial\alpha^2}\bigg|_{\alpha=\xi} (\breve{\alpha}_1^*(G_n) - \alpha_1^*(G_n))^2 \\ &\geq \frac{\delta\beta^2}{8}. \end{aligned}$$

For the same ϵ, by (A.7), there exists another $N_2 > 0$ such that for any $G_n > N_2$, $\sup_{\alpha\in[0,1]}|H_1(\mathbf{Z}, \alpha, G_n) - H_1(\boldsymbol{\sigma}^2, \alpha, G_n)| < \epsilon$. Therefore, for any $G_n > \max\{N_1, N_2\}$, we have

$$\begin{aligned} H_1(\mathbf{Z}, \breve{\alpha}_1^*(G_n), G_n) &\geq H_1(\boldsymbol{\sigma}^2, \breve{\alpha}_1^*(G_n), G_n) - \epsilon \\ &\geq H_1(\boldsymbol{\sigma}^2, \alpha_1^*(G_n), G_n) + \delta\beta^2/8 - \epsilon \\ &\geq H_1(\mathbf{Z}, \alpha_1^*(G_n), G_n) + \delta\beta^2/8 - 2\epsilon \\ &> H_1(\mathbf{Z}, \alpha_1^*(G_n), G_n), \end{aligned}$$

which contradicts the fact that $H_1(\mathbf{Z}, \breve{\alpha}_1^*(G_n), G_n)$ is the minimum of $H_1(\mathbf{Z}, \alpha, G_n)$.

The proof of Theorem 6(b) is skipped because it is similar.

[Received May 2005. Revised September 2006.]

REFERENCES

Abramowitz, M., and Stegun, I. (1972), *Handbook of Mathematical Functions*, New York: Dover.

Baldi, P., and Long, A. D. (2001), "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized *t*-Test and Statistical Inferences of Gene Changes," *Bioinformatics*, 17, 509–519.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000), "Microarray Expression Profiling Identifies Genes With Altered Expression in hdl-Deficient Mice," *Genome Research*, 10, 2022–2029.

Cui, X., and Churchill, G. A. (2003), "Statistical Tests for Differential Expression in cDNA Microarray Experiments," *Genome Biology*, 4, Art. no. 210.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005), "Improved Statistical Tests for Differential Gene Expression by Shrinking Variance Components Estimates," *Biostatistics*, 6, 59–75.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, London: Chapman and Hall.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., and Lee, J. (2003), "Local-Pooled Error Test for Identifying Differentially Expressed Genes With a Small Number of Replicated Microarrays," *Bioinformatics*, 19, 1945–1951.

James, W., and Stein, C. (1961), "Estimation With Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley, CA: University of California Press, pp. 361–379.

Kamb, A., and Ramaswami, A. (2001), "A Simple Method for Statistical Analysis of Intensity Differences in Microarray-Derived Gene Expression Data," *BMC Biotechnology*, 1, 1–8.

Kendziorski, C. M., Newton, M. A., Lan, H., and Could, M. N. (2003), "On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles," *Statistics in Medicine*, 22, 3899–3914.

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, G. A. (2002), "Statistical Analysis of a Gene Expression Microarray Experiment With Replication," *Statistica Sinica*, 12, 203–218.

Kubokawa, T. (1999), "Shrinkage and Modification Techniques in Estimation of Variance and the Related Problems: A Review," *Communications in Statistics A. Theory and Methods*, 28, 613–650.

Kubokawa, T., and Srivastava, M. S. (2003), "Estimating the Covariance Matrix: A New Approach," *Journal of Multivariate Analysis*, 86, 28–47.

Leung, Y., and Cavalieri, D. (2003), "Fundamentals of cDNA Microarray Data Analysis," *Trends in Genetics*, 11, 649–659.

Lin, Y., Nadler, S. T., Lan, H., Attie, A. D., and Yandell, B. S. (2003), "Adaptive Gene Picking With Microarray Data: Detecting Important Low Abundance Signals," in *The Analysis of Gene Expression Data: Methods and Software*, eds. G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, New York: Springer-Verlag, pp. 291–312.

Littell, R. C., Milliken, G., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute.

Lönnstedt, I., and Speed, T. (2002), "Replicated Microarray Data," *Statistica Sinica*, 12, 31–46.

Maatta, J. M., and Casella, G. (1990), "Developments in Decision-Theoretic Variance Estimation," *Statistical Science*, 5, 90–101.

Nguyen, D. V., Arpat, A. B., Wang, N., and Carroll, R. J. (2002), "DNA Microarray Experiments: Biological and Technological Aspects," *Biometrics*, 58, 701–717.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.

Stein, C. (1964), "Inadmissibility of the Usual Estimator for the Variance of a Normal Distribution With Unknown Mean," *Annals of the Institute of Statistical Mathematics*, 16, 155–160.

Storey, J., and Tibshirani, R. (2003), "SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays," in *The Analysis of Gene Expression Data: Methods and Software*, eds. G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, New York: Springer-Verlag, pp. 272–290.

Tong, T., and Wang, Y. (2005), "Optimal Shrinkage Estimation of Variances With Applications to Microarray Data Analysis," Technical Report 404, Department of Statistics and Applied Probability, University of California, Santa Barbara, available at *http://www.pstat.ucsb.edu/faculty/yuedong*.

Wu, H., and Churchill, G. A. (2005), "R/MAANOVA: An Extensive R Environment for the Analysis of Microarray Experiments," available at *http://www.jax.org/staff/churchill/labsite/software/Rmaanova/maanova.pdf*.