Variance estimation for gene-expression microarray data

Tiejun Tong[1] and Yuedong Wang[2]

[1] *Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA*

[2] *Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA*

tiejun.tong@colorado.edu yuedong@pstat.ucsb.edu

October 23, 2007

Abstract

High-throughput microarrays allow scientists to simultaneously measure the expression of thousands of genes. Due to the cost and experimental difficulties in obtaining biological materials, it is common that thousands of genes were measured only with a small number of replicates. This poses challenges to traditional statistical methods. In particular, the gene-specific estimates of variances are unreliable which leads to low power in the corresponding gene-by-gene tests. Various methods have been proposed in the literature to improve the variance estimation. We review the shrinkage, Bayes and regression methods in this paper.

Key words and phrases: Bayes estimation; James-Stein estimator; Loss function; Measurement error; Permutation; Shrinkage; Simulation-extrapolation.

1. Introduction

The development of microarray technology has revolutionized biomedical research. Instead of working on a gene-by-gene basis, the microarray technology allows simultaneous monitoring of the whole genome (Brown and Botstein 1999, Nguyen, Arpat, Wang and Carroll 2002, Leung and Cavalieri 2003). The resulting high-dimensional data demand novel statistical approaches for the experimental design, data analysis and interpretation. One major objective of microarray experiments is to identify differentially expressed genes under different conditions or at different time points. In reality, due to the cost and experimental difficulties in obtaining biological materials, it is common that thousands of genes were measured only with a small number of replicates (Lönnstedt and Speed 2002, Kendziorski, Newton, Lan and Could 2003). As a consequence, we are facing a "large G, small n" paradigm, where G is total number of genes and n is the number of replicates.

For simplicity of exposition, we consider the one-sample comparison problem. Let Y_{ij} (e.g. the log-ratio of the two-channel intensities in a cDNA array) be the jth replicate of observed expression level of gene i. With proper data preprocessing, we assume that $Y_{ij} \overset{iid}{\sim} \mathrm{N}(\mu_i, \sigma_i^2)$ where $\mathrm{N}(\mu_i, \sigma_i^2)$ represents the normal distribution with mean μ_i and variance σ_i^2. Identifying differentially expressed genes is equivalent to testing the hypothesis $H_{i0}: \ \mu_i = 0$ against $H_{i1}: \ \mu_i \neq 0$. The standard one-sample t-statistic

$$T_i = \frac{\sqrt{n}\bar{Y}_{i\cdot}}{S_i}, \qquad i = 1, \ldots, G, \tag{1}$$

where $\bar{Y}_{i\cdot} = \sum_{j=1}^n Y_{ij}/n$ is the sample mean and $S_i^2 = \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i\cdot})^2/(n-1)$ is the sample variance. Extensions to two-sample comparison or to more general settings are straightforward, see for example Tusher, Tibshirani and Chu (2001), Pan (2002) and Cui and Churchill (2003).

Due to the small sample size n, the standard gene-specific estimates of variances, S_i^2, are unreliable. Consequently, the standard test (1) for detecting differentially expressed genes on a gene-by-gene basis has low power (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. Therefore, tests based on a common variance estimate for all genes are at the risk of generating misleading results (Cui and Churchill 2003, Cui, Hwang, Qiu, Blades and Churchill 2005).

A key to improving the variance estimation is to borrow information across genes, implicitly or explicitly, locally or globally. Two common approaches for borrowing information are shrinkage and Bayes methods. Regression methods have also been employed when a functional relationship between mean and variance is appropriate. We review the shrinkage, Bayes and regression methods in Sections 2, 3, and 4. Extended t-tests based on various estimates of variances are given in Section 5. We conclude the paper in Section 6. This paper is not, and is not intended to be, an exhaustive survey of methods for the variance estimation.

2. Shrinkage methods

One of the earliest methods to stabilize the variance estimation was proposed by Tusher et al. (2001). In order to avoid the undue influence of small variance estimates in the denominator of (1), Tusher et al. (2001) proposed to replace S_i by $S_i/2 + s_0/2$ where s_0 is a constant. Note that the factor 1/2 is added for illustration and the resulting test statistic is equivalent to the SAM (significance analysis of microarrays) test statistic in Tusher et al. (2001). It is clear that SAM shrinks the gene-specific standard deviation S_i toward s_0 with the shrinkage factor 1/2. One popular choice of the constant s_0 is the 90th percentile of all estimated standard deviation (Efron, Tibshirani, Storey and Tusher 2001). The choices of the constant s_0 and shrinkage factor in the SAM test are somewhat arbitrary.

Another simple approach to stabilize the variance estimation was introduced by Cui and Churchill (2003), where they proposed a hybrid test by replacing the denominator of (1) with the square root of the average of the gene-specific and the pooled estimates. Though no theoretical justification was provided on why it works, they illustrated in both simulations and real studies that the hybrid test performs better in many situations than the test using either the gene-specific or the pooled estimate.

Cui et al. (2005) constructed a different variance estimator by shrinking gene-specific variance estimates toward their bias-corrected geometric mean. Specifically, the following James and Stein (1961)'s shrinkage technique was used in their derivation. Let $\mathbf{X} = (X_1, \ldots, X_G)^T \sim \mathrm{N}(\boldsymbol{\theta}, I_{G\times G})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)^T$. The goal is to estimate the unknown mean $\boldsymbol{\theta}$. Under the squared loss function, $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \|^2$, James and Stein (1961) proved that when $G \geq 3$, the standard MLE estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}} = \mathbf{X}$, was dominated by the James-Stein estimator

$$\tilde{\theta}_i = \left(1 - \frac{G-2}{\| \mathbf{X} \|^2}\right) X_i, \quad i = 1, \ldots, G.$$

$\tilde{\boldsymbol{\theta}}$ shrinks each sample mean toward a common value (zero here). The James-Stein estimator can be further improved by (Baranchik 1970)

$$\tilde{\theta}_i = \left(1 - \frac{G-2}{\| \mathbf{X} \|^2}\right)_+ X_i, \quad i = 1, \ldots, G,$$

where $(a)_+ = \max(0, a)$.

For microarray data based on the assumption that $Y_{ij} \overset{iid}{\sim} \mathrm{N}(\mu_i, \sigma_i^2)$, we have $S_i^2 = \sigma_i^2 \chi_{i,\nu}^2 / \nu$, where for ease of notation, $\chi_{i,\nu}^2$ denote i.i.d. random variables which have the Chi-squared distribution with $\nu = n - 1$ degrees of freedom. Taking the log-transformation leads to

$$X_i = \ln \sigma_i^2 + \epsilon_i, \tag{2}$$

where $X_i = \ln(S_i^2) - m$, $\epsilon_i = \ln(\chi_{i,\nu}^2/\nu) - m$ and $m = \mathrm{E}\{\ln(\chi_{i,\nu}^2/\nu)\}$.

Treating X_i in (2) as normal random variables, Cui et al. (2005) applied the James-Stein shrinkage method to derive a shrinkage estimate for $\ln \sigma_i^2$. Transforming back to the original scale, their estimates of the variances are

$$\tilde{\sigma}_i^2 = B \left(\prod_{i=1}^{G} (S_i^2)^{1/G} \right) \exp \left[\left(1 - \frac{(G-3)V}{\sum (\ln S_i^2 - \overline{\ln S_i^2})^2} \right)_+ \times (\ln S_i^2 - \overline{\ln S_i^2}) \right], \tag{3}$$

where $\overline{\ln S_i^2} = \sum_{i=1}^{G} \ln(S_i^2)/G$ and $B = \exp(-m)$ is the bias correction factor such that $B \prod_{i=1}^{G} (S_i^2)^{1/G}$ gives an unbiased estimator of σ^2 when $\sigma_i^2 = \sigma^2$ for all i.

Since X_i in (2) can be far from normal when ν is small, it is natural to ask whether the shrinkage variance estimates (3) can be improved further. It is also important to investigate theoretical properties. The shrinkage variance estimation has a long history starting with the amazing inadmissibility result discovered by Stein (1964), where the sample variance is improved by a shrinkage estimator using information from the sample mean. Much research has been done since then (Maatta and Casella 1990, Kubokawa 1999). However, most research in the literature concerned with a single variance (Kubokawa 1999) which is not applicable to microarray data analysis since the assumption of homogeneity among the variances is unlikely to be true. Some research has devoted to the shrinkage estimator of the covariance matrix (Kubokawa and Srivastava 2003). However, all existing methods required that $n > G$ to ensure the non-singularity of the sample covariance matrix. Therefore, these methods broke down for microarray data analysis.

Define the Stein and squared loss functions as

$$\begin{aligned} L_1(\sigma^2, \hat{\sigma}^2) &= \hat{\sigma}^2/\sigma^2 - \ln\left(\hat{\sigma}^2/\sigma^2\right) - 1, \text{ and} \\ L_2(\sigma^2, \hat{\sigma}^2) &= (\hat{\sigma}^2/\sigma^2 - 1)^2, \end{aligned}$$

respectively. Stein (1964) showed that S_i^2 are inadmissible under both loss functions. Therefore, S_i^2 are inadmissible under average loss functions $\sum_{i=1}^{G} L_k(\sigma_i^2, \hat{\sigma}_i^2)/G$, $k = 1, 2$.

Since variances appear in the denominator of (1), t-tests using estimates of the reciprocal of the standard deviations can be more powerful and robust than using the reciprocal of estimates of the standard deviations (Tong and Wang 2007). This insight led Tong and Wang (2007) to consider the estimation of $(\sigma_i^2)^t$ for any power $t \neq 0$. Note that σ_i and $1/\sigma_i$ are special cases with $t = 1/2$ and $t = -1/2$.

Let $S_i^{2t} = (S_i^2)^t$, $S_{pool}^{2t} = \prod_{i=1}^{G}(S_i^2)^{t/G}$ and

$$h_n(t) = (\frac{\nu}{2})^t \left(\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu}{2} + \frac{t}{n})}\right)^n, \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function. Tong and Wang (2007) proposed the following family of shrinkage estimators for $(\sigma_i^2)^t$,

$$\hat{\sigma}_i^{2t} = \left(h_G(t) S_{pool}^{2t}\right)^{\alpha} \left(h_1(t) S_i^{2t}\right)^{1-\alpha}, \quad 0 \leq \alpha \leq 1, \tag{5}$$

where $h_1(t)S_i^{2t}$ is an unbiased estimator of σ_i^{2t}, and $h_G(t)S_{pool}^{2t}$ is an unbiased estimator of σ^{2t} when $\sigma_i^2 = \sigma^2$ for all i. When $t = 1$, $\hat{\sigma}_i^2$ is a simple modification of the estimator in (3) (Tong and Wang 2007). The shrinkage parameter α controls the degree of shrinkage from the gene-specific variance estimate $h_1(t)S_i^{2t}$ toward the bias-corrected geometric mean $h_G(t)S_{pool}^{2t}$. There is no shrinkage when $\alpha = 0$, and all variance estimates are shrunken to the pooled variance when $\alpha = 1$. The goal is to find the optimal shrinkage parameter α in (5).

Let $\sigma^2_{pool} = \prod_{i=1}^G (\sigma_i^2)^{1/G}$, $\boldsymbol{\sigma}^{2t} = (\sigma_1^{2t}, \cdots, \sigma_G^{2t})$ and $\hat{\boldsymbol{\sigma}}^{2t} = (\hat{\sigma}_1^{2t}, \cdots, \hat{\sigma}_G^{2t})$. The average risks of the shrinkage estimators $\hat{\boldsymbol{\sigma}}^{2t}$ under the Stein and squared loss functions are

$$\begin{aligned} R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) &\triangleq \frac{1}{G}\sum_{i=1}^G \mathrm{E}\left(L_1(\sigma_i^{2t}, \hat{\sigma}_i^{2t})\right) \\ &= \frac{h_G^\alpha(t) h_1^{1-\alpha}(t)}{h_1^{G-1}\left(\frac{\alpha t}{G}\right) h_1\left((1-\alpha+\frac{\alpha}{G})t\right)} (\sigma^2_{pool})^{\alpha t} \frac{1}{G}\sum_{i=1}^G (\sigma_i^2)^{-\alpha t} \\ &\quad - \ln\left(h_G^\alpha(t) h_1^{1-\alpha}(t)\right) - t\Psi(\frac{\nu}{2}) + t\ln(\frac{\nu}{2}) - 1, \quad t > -\nu/2, \qquad (6) \\ R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) &\triangleq \frac{1}{G}\sum_{i=1}^G \mathrm{E}\left(L_2(\sigma_i^{2t}, \hat{\sigma}_i^{2t})\right) \\ &= \frac{h_G^{2\alpha}(t) h_1^{2(1-\alpha)}(t)}{h_1^{G-1}\left(\frac{2\alpha t}{G}\right) h_1\left(2(1-\alpha+\frac{\alpha}{G})t\right)} (\sigma^2_{pool})^{2\alpha t} \frac{1}{G}\sum_{i=1}^G (\sigma_i^2)^{-2\alpha t} + 1 \\ &\quad - \frac{2h_G^\alpha(t) h_1^{1-\alpha}(t)}{h_1^{G-1}\left(\frac{\alpha t}{G}\right) h_1\left((1-\alpha+\frac{\alpha}{G})t\right)} (\sigma^2_{pool})^{\alpha t} \frac{1}{G}\sum_{i=1}^G (\sigma_i^2)^{-\alpha t}, \quad t > -\nu/4, (7) \end{aligned}$$

where $\Psi(t) = \Gamma'(t)/\Gamma(t)$ is the digamma function (Abramowitz and Stegun 1972).

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

(i) $(\partial/\partial\alpha) R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=0} < 0$*;*

(ii) $(\partial/\partial\alpha) R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=1} \geq 0$*, where the equality holds if and only if $\sigma_i^2 = \sigma^2$ for all i;*

(iii) $(\partial^2/\partial\alpha^2) R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) > 0$ *for all $\alpha \in [0, 1]$.*

Theorem 1 indicates that for any fixed G, ν and non-zero $t > -\nu/2$, there exists a unique α_1^* in $(0, 1]$ which minimizes $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$.

Theorem 2. *For any fixed G, ν and non-zero power $t > -\nu/4$, we have*

(i) $(\partial/\partial\alpha) R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=0} < 0$*;*

(ii) $(\partial/\partial\alpha)R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})|_{\alpha=1} > 0$.

By Theorem 2 and the fact that $R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}) \geq 0$, 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 in $[0, 1]$. Therefore, α_2^* may not be unique.

Theorem 3. *For any fixed G and non-zero $t > -\nu/2$ $(t > -\nu/4)$, as $\nu \to \infty$,*

(i) $\alpha_1^* \to 0$ $(\alpha_2^* \to 0)$ when σ_i^2 are not all the same;

(ii) $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t})$ $(R_2(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}))$ approaches to a constant function of α when $\sigma_i^2 = \sigma^2$ for all i.

Theorem 3 indicates that it is unnecessary to borrow information from other genes when the sample size is large.

The optimal shrinkage parameters α_1^* and α_2^* cannot be used directly in practice since they depend on unknown parameters $\boldsymbol{\sigma}^{2t}$. They need to be estimated from data. Since both α_1^* and α_2^* depend on $\boldsymbol{\sigma}^{2t}$ through the unknown quantity

$$b(\boldsymbol{\sigma}^2, \eta) = (\sigma_{pool}^2)^{\eta} \frac{1}{G} \sum_{i=1}^{G} (\sigma_i^2)^{-\eta}, \tag{8}$$

where $\eta = \alpha t$ or $\eta = 2\alpha t$, we need to estimate $b(\boldsymbol{\sigma}^2, \eta)$. 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 (6) and (7) replaced by $b(\mathbf{Z}, \eta)$. The following theorem shows that, as $\nu \to \infty$, $\sigma_i^{2t}(\hat{\alpha}_1^*)$ and $\sigma_i^{2t}(\hat{\alpha}_2^*)$ are asymptotically optimal, and $\hat{\alpha}_1^*$ and $\hat{\alpha}_2^*$ are consistent.

Theorem 4. *For any fixed G and non-zero t, when $\nu \to \infty$, we have*

(i) $b(\mathbf{Z}, \alpha t) \overset{a.s.}{\to} b(\boldsymbol{\sigma}^2, \alpha t)$ uniformly for $\alpha \in [0, 1]$;

(ii) $R_k(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\hat{\alpha}_k^*(\nu))) - R_k(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha_k^*(\nu))) \overset{a.s.}{\to} 0$, $k = 1, 2$;

(iii) $\hat{\alpha}_1^*(\nu) \overset{a.s.}{\to} 0$ and $\hat{\alpha}_2^*(\nu) \overset{a.s.}{\to} 0$ when σ_i^2 are not all the same.

For microarray data, ν is small and G is large. Therefore, it is more appropriate to investigate asymptotic properties as $G \to \infty$ for a fixed ν. Consider σ_i^2 as random variables and assume that $\sigma_i^2 \overset{iid}{\sim} F$, $i = 1, \ldots, G$. Let

$$w(\alpha t) = (\nu/2)^{\alpha t} h_1(-\alpha t) \exp\left[-\alpha t \Psi(\nu/2)\right].$$

It can be shown that for any fixed non-zero t with $\nu > 2t$, $\mathrm{E}(\sigma_1^2)^{-t} < \infty$ and $\mathrm{E}(\ln(\sigma_1^2)) < \infty$,

$$w(\alpha t) b(\mathbf{Z}, \alpha t) - b(\boldsymbol{\sigma}^2, \alpha t) \overset{a.s.}{\to} 0 \text{ uniformly for } \alpha \in [0,1] \text{ as } G \to \infty.$$

$w(\alpha t)$ acts as a bias correction factor for estimating $b(\boldsymbol{\sigma}^2, \alpha t)$. For a fixed t, let $H_k(\boldsymbol{\sigma}^2, \alpha, G) = R_k(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha))$, $H_k(\mathbf{Z}, \alpha, G)$ 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 $\alpha_k^*(G) = \underset{\alpha \in [0,1]}{\text{argmin}} H_k(\boldsymbol{\sigma}^2, \alpha, G)$ and $\breve{\alpha}_k^*(G) = \underset{\alpha \in [0,1]}{\text{argmin}} H_k(\mathbf{Z}, \alpha, G)$.

Theorem 5. *For any fixed non-zero t,*

(i) when $\nu > 2|t|$, $\mathrm{E}(\sigma_1^2)^{-t} < \infty$ and $\mathrm{E}(\ln(\sigma_1^2)) < \infty$, we have $R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\breve{\alpha}_1^*(G))) - R_1(\boldsymbol{\sigma}^{2t}, \hat{\boldsymbol{\sigma}}^{2t}(\alpha_1^*(G))) \overset{a.s.}{\to} 0$ and $\breve{\alpha}_1^*(G) - \alpha_1^*(G) \overset{a.s.}{\to} 0$ as $G \to \infty$;

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

Note that there is no corresponding consistent result for α_2^* since it may not be unique. Theorem 5 does not apply for small ν. The following two-step procedure may be used when ν is small: (i) substitute $b(\boldsymbol{\sigma}^2, \eta)$ in (6) and (7) by $b(\mathbf{Z}, \eta)$, and compute temporary optimal shrinkage parameters and the corresponding shrinkage estimators, say $\hat{\boldsymbol{\sigma}}_-^2$; (ii) substitute $b(\boldsymbol{\sigma}^2, \eta)$ in (6) and (7) by $b(\hat{\boldsymbol{\sigma}}_-^2, \eta)$ to get the final optimal shrinkage parameters.

3. Bayes methods

Baldi and Long (2001) assumed the following conjugate prior for (μ_i, σ_i^2),

$$p(\mu_i, \sigma_i^2|\alpha) = \mathrm{N}(\mu_i; \mu_0, \sigma_i^2/\lambda_0)\mathcal{I}(\sigma_i^2; \nu_0, \sigma_0^2),$$

where $\alpha = (\mu_0, \lambda_0, \nu_0, \sigma_0^2)$, $\mathrm{N}(x; a, b^2)$ represents the normal density function with mean a and variance b^2, $\mathcal{I}(x; a, b^2)$ represents the scaled inverse gamma density with degrees of freedom a and scale b. The posterior density for (μ_i, σ_i^2) has the same functional form as the prior density

$$p(\mu_i, \sigma_i^2|\boldsymbol{Y_i}, \alpha) = \mathrm{N}(\mu_i; \mu_n, \sigma_i^2/\lambda_n)\ \mathcal{I}(\sigma_i^2; \nu_n, \sigma_n^2),$$

where $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{in})$,

$$\begin{aligned}
\mu_n &= \frac{\lambda_0}{\lambda_0 + n}\mu_0 + \frac{n}{\lambda_0 + n}\bar{Y}_{i\cdot}, \\
\lambda_n &= \lambda_0 + n, \\
\nu_n &= \nu_0 + n, \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)S_i^2 + \frac{\lambda_0 n}{\lambda_0 + n}(\bar{Y}_{i\cdot} - \mu_0)^2.
\end{aligned}$$

Note that the posterior mean μ_n is a weighted average of the prior mean μ_0 and the sample mean $\bar{Y}_{i\cdot}$. Though a prior mean μ_0 may be used to improve estimation of the mean, Baldi and Long (2001) claimed that it suffices to use $\mu_0 = \bar{Y}_{i\cdot}$. With $\mu_0 = \bar{Y}_{i\cdot}$, the posterior means of μ_i and σ_i^2 are

$$\begin{aligned}
\hat{\mu}_i &= \mu_n = \bar{Y}_{i\cdot}, \\
\hat{\sigma}_i^2 &= \frac{\nu_n}{\nu_n - 2}\sigma_n^2 = \frac{\nu_0\sigma_0^2 + (n-2)S_i^2}{\nu_0 + n - 2}.
\end{aligned}$$

The posterior modes have the same form as above with $n-2$ replaced by $n-1$. It is clear that both posterior mean and mode of σ_i^2 are shrinkage estimators. The background variance σ_0^2 is estimated by pooling together all the neighboring genes contained in a window of a certain size. The parameter ν_0 represents the degree of confidence in the background variance σ_0^2 versus the gene-specific sample variance. In principle, the smaller the sample size n, the larger ν_0 ought to be. See Baldi and Long (2001) for more detail on selecting the hyperparameters.

Other methods in the Bayesian framework are summarized as follows. Lönnstedt and Speed (2002) proposed a posterior odds of differential expression in a replicated two-color experiment using an empirical Bayes approach that combines information across genes. Kendziorski et al. (2003) extended the empirical Bayes method using the hierarchical gamma-gamma and lognormal-normal models. Smyth (2004) developed a hierarchical model in the context of general linear models.

4. Regression methods

In microarray data analysis, it is commonly observed that the variance increases proportionally with the intensity level (Chen, Dougherty and Bittner 1997, Rocke and Durbin 2001, Strimmer 2003, Weng, Dai, Zhan, He, Stepaniants and Bassett 2006). Instead to borrowing information from the whole genome, it is attractive to borrow information from genes with similar variances.

The regression approach assumes a functional relationship between the mean and the variance: $\sigma_i^2 = g(\mu_i)$. The goal is to estimate the variance-mean function g. The function g can be modeled parametrically and nonparametrically. Popular parametric models include the constant coefficient of variation model (Chen et al. 1997), $g(\mu) = \theta\mu^2$, and the quadratic model (Rocke and Durbin 2001, Chen, Kamat, Dougherty, Bittner, Meltzer and Trent 2002), $g(\mu) = \theta_1 + \theta_2\mu^2$. While it is adequate for genes with high expression levels, the constant of variation model is inaccurate when the signal is weak in comparison to the background. The quadratic model was proposed to overcome this problem. In general, denote $g(\mu, \boldsymbol{\theta})$ as the parametric variance function with parameters $\boldsymbol{\theta}$. Nonparametric regression approach may be used when it is undesirable to assume a parametric model for g. Any one of the nonparametric regression approaches such as smoothing spline and local polynomial could be used to model g nonparametrically.

The estimation, however, is not as simple as it seems. There are several subtle issues involved in the estimation. First, for the purpose of estimating the variance function, the means μ_i represent a large number of unknown nuisance parameters. Direct application of likelihood (or quasi-likelihood) is likely to lead to inconsistent estimates. This is a Neyman-Scott type problem where care needs to be taken to derive consistent estimates.

If μ_i were observable, it is not difficult to construct consistent estimators, $\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})$ and $\hat{g}(\boldsymbol{\mu})$, for $\boldsymbol{\theta}$ and g, where the dependence on $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_G)$ is expressed explicitly. Unfortunately, $\boldsymbol{\mu}$ are not observable. Since $\bar{\boldsymbol{Y}}. = (\bar{Y}_{1\cdot}, \cdots, \bar{Y}_{G\cdot})$ are unbiased estimates of $\boldsymbol{\mu}$, a naive approach is to replace $\boldsymbol{\mu}$ by $\bar{\boldsymbol{Y}}.$, i.e. to use estimates $\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{Y}}.)$ and $\hat{g}(\bar{\boldsymbol{Y}}.)$. Ignoring the sampling error in $\bar{\boldsymbol{Y}}.$, $\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{Y}}.)$ and $\hat{g}(\bar{\boldsymbol{Y}}.)$ are in general inconsistent (Wang, Ma and Carroll 2007, Carroll and Wang 2007). This is the second subtle issue involved in the estimation.

Since $\bar{\boldsymbol{Y}}.$ are error-prone unbiased measures of $\boldsymbol{\mu}$ with non-constant variation, the problem can be viewed as one with heteroscedastic measurement error. Thus the following SIMEX (simulation extrapolation) method developed in the heteroscedastic measurement error model framework may be used (Carroll, Ruppert, Stefanski and

Crainiceanu 2006):

1. Generate $Z_{bij} \overset{iid}{\sim} N(0,1)$, $i = 1, \cdots, G$, $j = 1, \cdots, n$, $b = 1, \cdots, B$. Let
$$c_{bij} = \frac{Z_{bij} - \bar{Z}_{bi\cdot}}{\sqrt{\sum_{j=1}^{n}(Z_{bij} - \bar{Z}_{bi\cdot})^2}}.$$

2. For $i = 1, \cdots, G$, $j = 1, \cdots, n$, $b = 1, \cdots, B$, let $W_{ij} = Y_{ij}$ and
$$W_{bi}(\zeta) \quad = \quad \bar{W}_{i\cdot} + (\frac{\zeta}{n})^{1/2} \sum_{j=1}^{n} c_{bij} W_{ij}.$$

3. Estimate $\boldsymbol{\theta}$ and g by replacing $\boldsymbol{\mu}$ in $\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})$ and $\hat{g}(\boldsymbol{\mu})$ with $\boldsymbol{W}_b(\zeta) = (W_{b1}, \cdots, W_{bG})$ for each b and then average over b.

4. Extrapolate back to $\zeta = -1$.

Except for some special cases, the above direct application of the SIMEX method still does not lead to consistent estimates (Wang et al. 2007, Carroll and Wang 2007). In general, SIMEX-type methods and indeed most measurement error methods require nondifferential measurement error, i.e., the measurement error is independent of the response. However, this is not the case here: the "response" S_i^2 is not independent of $W_{bi}(\zeta)$, and hence the measurement error in the SIMEX steps is differential. This is the third subtle issue involved in the estimation.

To "break" the connection between the response and the measurement errors, and to force nondifferential error, the following permutation SIMEX methods were proposed in Wang et al. (2007) and Carroll and Wang (2007):

1. Do $j = 1, \cdots, n$,

(a) Generate $Z_{bik} \overset{iid}{\sim} N(0,1)$, $i = 1, \cdots, G$, $k = 1, \cdots, n-1$, $b = 1, \cdots, B$. Let

$$c_{bik}^{(j)} = \frac{Z_{bik} - \bar{Z}_{bi\cdot}}{\sqrt{\sum_{k=1}^{n-1}(Z_{bik} - \bar{Z}_{bi\cdot})^2}}.$$

(b) For $i = 1, \cdots, G$, $k = 1, \cdots, n-1$, $b = 1, \cdots, B$, let

$$\begin{aligned} W_{ik}^{(j)} &= \begin{cases} Y_{ik} & 1 \le k \le j-1, \\ Y_{i(k+1)} & j \le k \le n-1, \end{cases} \\ W_{bi}^{(j)}(\zeta) &= \overline{W}_{i\cdot}^{(j)} + (\frac{\zeta}{n-1})^{1/2} \sum_{k=1}^{n-1} c_{bik}^{(j)} W_{ik}^{(j)}, \\ S_i^{(j)} &= \{Y_{ij} - W_{bi}^{(j)}(\zeta)\}^2. \end{aligned}$$

2. Estimate $\boldsymbol{\theta}$ and g by replacing $\boldsymbol{\mu}$ in $\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})$ and $\hat{g}(\boldsymbol{\mu})$ with $\boldsymbol{W}_b^{(j)}(\zeta) = (W_{b1}^{(j)}, \cdots, W_{bG}^{(j)})$ for each combination of j and b, and then average over all j and b.

3. Extrapolate to $\zeta = -1$.

The permutation SIMEX method requires $n \ge 3$. It does lead to consistent estimates (Wang et al. 2007, Carroll and Wang 2007).

5. Extended t-tests

When plugged in (1), different estimates of variances lead to different test statistics. These are referred to as the extended t-tests. Unlike the sample variance, the improved estimates of variances do not follow Chi-squared distributions. Therefore, the extended t-tests do not follow t distributions. The null distribution is usually calculated through permutation. The key to developing a permutation strategy is to identify the exchangeable units under the null hypothesis. For two-color arrays with null hypotheses that the genes are all constantly expressed, the red and green channel intensities are regarded to be exchangeable. Often there is not enough permutations to establish the null distribution due to small sample size. To reduce the granularity of the gene-specific null distribution, Cui et al. (2005) suggested to use a common null distribution by using all shuffled statistics over index i.

Some extended t-tests including the SAM t-statistic (Tusher et al. 2001), the regularized t-statistic (Baldi and Long 2001) and the posterior odds statistic (Lönnstedt and Speed 2002) are implemented in the software package **Limma** for the R computing environment (Smyth, Ritchie, Thorne and Wettenhall 2006). Limma is part of the Bioconductor project at http://www.bioconductor.org. Comparison results among these tests can also be found in Kooperberg, Aragaki, Strand and Olson (2005).

Cui et al. (2005) demonstrated that their shrinkage-based test has the best or nearly the best power among several "information-sharing" statistics including the regularized t-test, the SAM t-test, the posterior odds statistic and the hybrid test by Cui and Churchill (2003). The optimal shrinkage parameters can further improve the power of shrinkage-based test (Tong and Wang 2007). Further research is necessary to compare the practical performance of extended t-tests including many methods which are not reviewed in this paper.

6. Discussion

One major limitation of microarray experiments is that the number of samples is relative small compared to the large number of genes. This poses challenges to traditional statistical methods. In this paper we review various methods for estimating variances in the "large G, small n" paradigm. In addition to constructing more powerful tests, the improved variance estimation has other applications. For example, it can be used to develop new discrimination methods (Pang, Tong and Zhao 2007).

Most existing methods assume that the genes are mutually independent which is unlikely to hold in practice. Some methods may still work well in practice (Tong and Wang 2007). Nevertheless, further research is necessary to assess the impact, if any, of correlation between genes on the estimation of variances and the power of tests.

Acknowledgements

Tiejun Tong's research was supported by a start-up grant from University of Colorado at Boulder. Yuedong Wang's research was supported by a grant from the National Science Foundation (DMS-0706886).

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of mathematical functions*, Dover, New York.

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.

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution, *Annals of Mathematical Statistics* **41**: 642–645.

Brown, P. and Botstein, D. (1999). Exploring the new world of the genome with dna microarrays, *Nat Genet.* **21**: 33–37.

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.

Carroll, R. J. and Wang, Y. (2007). Nonparametric variance estimation in the analysis of microarray data: a measurement error approach, *Submitted*.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed.*, Chapman & Hall, New York.

Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-biased decisions and the quantitative analysis of cDNA microarray images, *Journal of Biomedical Optics* **2**: 364–374.

Chen, Y., Kamat, V., Dougherty, E. R., Bittner, M. L., Meltzer, P. S. and Trent, J. M. (2002). Ratio statistics of gene expression levels and applications to microarray data analysis, *Bioinformatics* **18**: 1207–1215.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments, *Genome Biology* **4**: 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.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.* **1**: 361–379.

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.

Kooperberg, C., Aragaki, A., Strand, A. D. and Olson, J. M. (2005). Significance testing for small microarray experiments, *Statistics in Medicine* **24**: 2281–2298.

Kubokawa, T. (1999). Shrinkage and modification techniques in estimation of variance and the related problems: A review, *Comm. Statist. A–Theory 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.

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.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics* **18**: 546–554.

Pang, H., Tong, T. and Zhao, H. (2007). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data, *Submitted*.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays, *Journal of Computational Biology* **8**: 557–569.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiment, *Statistical Applications in Genetics and Molecular Biology* **3**: 1.

Smyth, G. K., Ritchie, M., Thorne, N. and Wettenhall, J. (2006). Limma: Linear models for microarray data, user's guide, *Melbourne: Walter and Eliza Hall institute of Medical Research*.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean, *Ann. Inst. Statist. Math.* **16**: 155–160.

Strimmer, K. (2003). Modeling gene expression measurement error: a quasi-likelihood approach, *BMC Bioinformatics* **4**: 10.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis, *Journal of the American Statistical Association* **102**: 113–122.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significant analysis of microarray applied to transcriptional responses to ionizing radiation, *Proceedings National Academic Science* **98**: 5116–5121.

Wang, Y., Ma, Y. and Carroll, R. J. (2007). Variance estimation in the analysis of microarray data, *Submitted*.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B. and Bassett, D. E. (2006). Rosetta error model for gene expression analysis, *Bioinformatics* **22**: 1111–1121.