Statistics and Its Interface Volume 3 (2010) 533–541

Double shrinkage empirical Bayesian estimation for unknown and unequal variances

Zhigen Zhao

In this paper, we construct a point estimator when assuming unequal and unknown variances by using the *empirical* Bayes approach in the classical normal mean problem. The proposed estimator shrinks both means and variances, and is thus called the double shrinkage estimator. Extensive numerical studies indicate that the double shrinkage estimator has lower Bayes risk than the estimator which shrinks the means alone, and the naive estimator which has no shrinkage at all. We further use a spike-in data set to assess different estimating procedures. It turns out that our proposed estimator performs the best and is thus strongly recommended for applications.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: James–Stein estimator, Log-normal model, Loss function.

1. INTRODUCTION

The shrinkage estimator has a long history dating back to the 1950s. Assume that $X_i \overset{\text{iid}}{\sim} N(\theta_i, \sigma_i^2)$ $(i = 1, 2, \ldots, p)$. When $p \geq 3$ and variances σ_i^2's are known and all equal to σ^2, James and Stein (1961) proposed an estimator

$$\hat{\theta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\sum_i X_i^2}\right) X \tag{1}$$

which dominates the estimator $\hat{\theta} = X$ when using squared error loss

$$L(\hat{\theta}, \theta) = \frac{1}{p} \sum_i (\hat{\theta}_i - \theta_i)^2. \tag{2}$$

Further, it is easily seen that the estimator (1) can be dominated by the positive part James–Stein estimator defined as

$$\hat{\theta}_{JS+} = \left(1 - \frac{(p-2)\sigma^2}{\sum_i X_i^2}\right)_+ X = \delta_{JS+} X. \tag{3}$$

When simultaneously estimating all the parameters θ_i's, it is beneficial to use X_i and the seemly unrelated observations $X_{(-i)}$, where the $X_{(-i)}$ consists of all the observations X but X_i, when estimating each individual θ_i. This is known as borrowing strength effect. Although, the shrinkage scaling $\delta_{JS+} = (1 - \frac{(p-2)\sigma^2}{\sum_i X_i^2})_+$ falls between zero and one, leading to biased estimators $\hat{\theta}_{JS+}$ for all the parameters θ_i's, such scaling reduces the variances of the estimators resulting in a lower risk compared with the naive estimator X_i's.

The positive part James–Stein's estimator is purely frequentist based. Lindley (1962) derived an estimator based on the *empirical* Bayesian approach described as following. Given a prior distribution of θ_i as $N(\mu, \tau^2)$, the posterior expectation of θ_i given X_i is $MX_i + (1-M)\mu$ where $M = \frac{\tau^2}{\tau^2 + \sigma^2}$ which minimizes the Bayes risk. Since μ and τ^2 are usually unknown in practice, one estimates M and μ from the data by using, for example, the method of moments. This resulted in the Lindley–James–Stein's estimator, abbreviated below as LJS in short,

$$\hat{\theta}_{LJS,i} = \hat{M} X_i + (1 - \hat{M}) \bar{X},$$

where $\hat{M}$ is a constant $(1 - \frac{(p-3)\sigma^2}{\sum_i (X_i - \bar{X})^2})$ for all θ_i's. Since M is never negative, it is natural to further take the positive part of $\hat{M}$ and obtain the positive part LJS estimator. This estimator is a shrinking-mean estimator. It pulls the observation X_i towards the arithmetic mean $\bar{X}$. It can also be shown that the positive part LJS has a smaller frequentist risk than that of LJS which in turn has smaller risk than X for every θ's.

The happy marriage of the James–Stein's idea and *empirical* Bayes approach brings a revolution in mathematical statistics. Statisticians use these ideas to produce different testing procedures, confidence intervals and others. To name a few, when assuming $\vec{X} \sim N(\vec{\theta}, \sigma^2 I)$, Casella and Hwang (1983) constructed a confidence set which dominates the naive confidence sphere centered at $\vec{X}$ with radius $c\sigma$, where $P(\chi_p^2 < c^2) = 1 - \alpha$. Morris (1983a), Morris (1983b), and He (1992) constructed different *empirical* Bayesian confidence intervals for θ_i's. Qiu and Hwang (2007) constructed confidence intervals for selected θ_i's when assuming a mixed prior. See Casella and Hwang (2010) for a review.

However, all the literature listed above either assume a known variance σ^2 or simply replace them by S_i^2 when σ_i^2's are unknown and unequal (See also Efron and Morris (1973); Morris (1983b)). This straightforward substitution results in a point estimator which only has the shrinking-mean effect for the heteroscedasticity case where the σ_i^2's are unequal and unknown. Fourdrinier et al. (2003) considered this type of estimators and proved that the frequentist risk under some weighted squared error of the new estimator is small. As we know, shrinking the means brings us much benefit, such as low risks in point estimators, shortness in intervals and powerfulness in testing procedures. What about shrinking the variances? Unlike shrinking the means, it is not until recently that researchers realize the advantage of the variance shrinkage and know how to shrink the variances in the context of estimating means.

Nowadays, in microarray experiments when the dimension p is very large, typically 10 thousands or more, some of the observations S_i^2's can be either extremely large or small. In the spike-in data set we analyze in Section (5), the smallest value of S_i^2's is 6.0611×10^{-5} while the largest one is 5.4160. Consequently, the testing procedure could either be of little power or detect much false significance. It seems advantageous to shrink the variances toward the common mean. This would enlarge the small variances and reduce the large variances.

Imposing an *inverse Gamma* prior of σ_i^2 with hyper parameter a and b (see Berger (1985)) along with the assumption that $\frac{S_i^2}{\sigma_i^2} \sim \frac{\chi_d^2}{d}$, Smyth (2004) developed a better testing procedure by using the *empirical* Bayes approach. In Smyth's procedure, the variance σ_i^2 is estimated by

$$\hat{\sigma}_i^2 = \frac{\frac{1}{\hat{b}} + \frac{dS_i^2}{2}}{\frac{d}{2} + \hat{a} - 1} \tag{4}$$

where $\hat{a}$ and $\hat{b}$ are estimated by a numerical algorithm. This estimator truncates the small value of S_i^2 to be at least $\frac{1/\hat{b}}{d/2+\hat{a}-1}$. When $S_i^2 \gg \frac{1}{\hat{b}}$, $\hat{\sigma}_i^2$ is at most $\frac{d/2}{d/2+\hat{a}-1} S_i^2$ which is smaller than S_i^2 given $\hat{a} > 1$. Consequently, Smyth's variance estimator shrinks the variances. However, there are no explicit formulas of $\hat{a}$ and $\hat{b}$, and it is hard to evaluate the property of the estimator analytically.

In 2005, Cui et al. (2005) proposed the exponential LJS estimator for the variance component σ_i^2's with an explicit form which shrinks the observation towards their geometric mean as explained in Section 4. They further argued that the testing procedure F_S based on this variance shrinkage estimator enjoys high power. The subscript S here means that the procedure has only one shrinkage factor – shrinking the variances.

In addition to modeling the true parameter θ_i's, Hwang and Liu (2010) further put a log-normal prior for the variances σ_i^2's. After approximating the $\frac{\chi_d^2}{d}$ by another log-normal random variable, they proposed the so-called *Log-Normal* model. Using the *empirical* Bayes approach, they derived another testing procedure F_{SS} where the subscript SS means that this procedure has double-shrinkage factor—shrinking both the means and variances. They have demonstrated that the average power of F_{SS} is higher than that of all the other tests, such as F_S, the shrinking variance alone test, and the T-test without any shrinkage. They have further concluded that it is better than the moderated T-test based on the variance shrinkage estimator (4) as in Smyth (2004).

Based on the same model, Hwang, Qiu and Zhao (2009) studied the *empirical* Bayes confidence interval with the double shrinkage effect. It turns out that this new construction dominates the naive t interval in terms of a sharper average length when guaranteeing the *empirical* Bayes coverage probability. They further argued that the confidence interval with double shrinkage is better than both the shrinking-mean-alone and the shrinking-variance-alone interval, which are better than intervals with no shrinkage.

It is interesting that no study aims at constructing estimators that shrink both the means and variances as far as the author is aware. This is what we do in this paper. We construct estimators for θ's which shrinks both the means and variances. The article is organized as following. In Section 2, we introduce the *general* Log-Normal model and derive a point estimator when assuming known hyperparameters. In Section 3, we estimate the hyper parameters from the data and derive the *empirical* Bayes estimator. We study the Bayes risk of the new estimator for a class of priors based on the loss function (2) by using extensive simulation studies and a real data analysis in Sections 4 and 5. We conclude in Section 6 that the point estimator with the double shrinkage is better than the estimator with one shrinkage, which is better than the estimator without any shrinkage.

2. ESTIMATOR WITH KNOWN HYPER PARAMETER

In this section, we define the canonical model over which we shall construct double shrinkage estimators. Firstly, assume that each observation X_i $(i = 1, \ldots, p)$ follows a normal distribution with mean θ_i and unknown variance σ_i^2 which differ across all the observations. The heteroscedasticity of σ_i^2 occurs often in application but causes lots of difficulties. We assume that there exists another statistic S_i^2 containing the information of the variance σ_i^2 which is independent of X_i. In general, it is assumed that $S_i^2|\sigma_i^2 \overset{\text{iid}}{\sim} \sigma_i^2 \frac{\chi_{d_i}^2}{d_i}$ where d_i represents the degrees of freedom corresponding to the i-th observation.

In modern application such as microarray technology, the dimension p is large, typically varying from several thousands to 50 thousands. Therefore, it is practical to put a prior distribution over θ_i. When assuming that all the σ_i^2's are equal and known as σ^2, Lindley (1962) assumes a normal prior $N(\mu, \tau^2)$ for θ_i and derived the well known Lindley James–Stein estimator of θ_i as

$$\hat{\theta}_i = \bar{X} + \left(1 - \frac{(p-3)\sigma^2}{\sum_i (X_i - \bar{X})^2}\right)(X_i - \bar{X}). \tag{5}$$

Similarly, in our model, we assume the same prior $N(\mu, \tau^2)$ for the true parameter θ_i.

There are many variances σ_i^2's, and it seems reasonable to assume a prior for the variances σ_i^2's as well. It is convenient to put an inverse gamma prior with the shape parameter a and scale parameter b (see Berger (1985)) for σ_i^2 because it is conjugate to the χ^2 random variable. Smyth (2004) took this approach and derived an *empirical* Bayes testing procedure. However, a disadvantage of such an approach is that there is no explicit formula for the estimator of the hyper parameters. Smyth (2004) introduced a numerical algorithm to estimate these two parameters.

In our model, we first approximate $\log \frac{\chi^2_{d_i}}{d_i}$ by $N(m_i, \sigma^2_{ch,i})$ where

$$m_i = E \log \frac{\chi^2_{d_i}}{d_i} = \psi\left(\frac{d_i}{2}\right) - \log \frac{d_i}{2},$$
$$\sigma^2_{ch,i} = Var\left(\log \frac{\chi^2_{d_i}}{d_i}\right) = \psi'\left(\frac{d_i}{2}\right),$$

where $\psi(x) = \frac{d}{dx}\log\Gamma(x)$, known as the *digamma* function. The two constants m_i and $\sigma^2_{ch,i}$ depend solely on the degrees of freedom d_i and can be evaluated easily. Consequently, we approximate the logarithm of $\frac{\chi^2_{d_i}}{d_i}$ by a normal random variable with the same first and second moments. As a result,

$$\log S_i^2 | \log \sigma_i^2 \overset{\text{iid}}{\sim} N(m_i + \log \sigma_i^2, \sigma^2_{ch,i}).$$

Furthermore, we assume that $\log \sigma_i^2$ is a normal random variable with a hyper mean μ_v and a variance τ_v^2. To distinguish with the prior distribution of the mean θ_i, we use the subindex v, since they are the hyper parameters corresponding to the variances.

In summary, the canonical model is

$$(6) \qquad \begin{cases} X_i|\theta_i, \sigma_i^2 \overset{\text{iid}}{\sim} N(\theta_i, \sigma_i^2); \\ \theta_i \overset{\text{iid}}{\sim} N(\mu, \tau^2); \\ \log S_i^2 | \log \sigma_i^2 \overset{\text{iid}}{\sim} N(m_i + \log \sigma_i^2, \sigma^2_{ch,i}); \\ \log \sigma_i^2 \overset{\text{iid}}{\sim} N(\mu_v, \tau_v^2), \end{cases}$$

where $m_i = \psi(\frac{d_i}{2}) - \log \frac{d_i}{2}$, $\sigma^2_{ch,i} = \psi'(\frac{d_i}{2})$.

This model is called *Log-normal* model in Hwang and Liu (2010) and Hwang, Qiu and Zhao (2009) where they assume that the degrees of freedom d_i across all observations are the same. Hwang and Liu (2010) constructed a powerful testing procedure while Hwang, Qiu and Zhao (2009) constructed sharp *empirical* Bayesian confidence intervals based on the same model setting. In the data analysis part of Hwang, Qiu and Zhao (2009) when the degrees of freedom of each gene are either 2 or 3, they took a conservative approach and simply set all degrees of freedom to be 2. As illustrated in Section 4, if taking the same conservative approach when the degrees of freedom d_i's are different, the corresponding estimator has a slightly larger risk when compared with a new point estimation procedure based on this Log-Normal model, where d_i are not necessarily identical.

Having the model, we first derive the point estimator $\hat{\theta}$ when assuming that all the hyper-parameters $\mu, \tau^2, \mu_v, \tau_v^2$ are known. Since $X_i|\theta_i \sim N(\theta_i, \sigma_i^2)$ and $\theta_i \sim N(\mu, \tau^2)$, we know that $\theta_i|X, \sigma_i^2 \sim N(M_iX_i + (1-M_i)\mu, M_i\sigma_i^2)$ where $M_i = \frac{\tau^2}{\tau^2+\sigma_i^2}$. For the known variance σ_i^2 case, the natural estimator of θ_i is $M_iX_i + (1-M_i)\mu$, which is the posterior expectation of θ_i given X_i and σ_i^2. This estimator shrinks the observation X_i towards the hyper mean μ. The shrinkage scaling M_i equals $\frac{\tau^2}{\tau^2+\sigma_i^2}$, which depends on the variance σ_i^2 of the i-th observation.

However, since σ_i^2 is unknown, we need to substitute it by a variance estimator $\hat{\sigma}_i^2$ depending on the observation S_i^2 and hyper parameter μ_v and τ_v^2. One typical approach is to replace σ_i^2 by S_i^2, and estimate θ_i by

$$\hat{\theta}_i = \hat{M}_iX_i + (1-\hat{M}_i)\mu,$$

where $\hat{M}_i = \frac{\tau^2}{\tau^2+S_i^2}$. This estimator shrinks the observation X_i towards the common mean μ. However, there is no variance shrinkage.

Recall in model (6) one knows that,

$$\log S_i^2 | \log \sigma_i^2 \sim N(m_i + \log \sigma_i^2, \sigma^2_{ch,i}),$$

and

$$\log \sigma_i^2 \sim N(\mu_v, \tau_v^2).$$

A classical calculation indicates that

$$\log \sigma_i^2 | \log S_i^2 \sim N(M_{v,i}(\log S_i^2 - m_i) + (1-M_{v,i})\mu_v, M_{v,i}\sigma^2_{ch,i})$$

where $M_{v,i} = \frac{\tau_v^2}{\tau_v^2+\sigma^2_{ch,i}}$. There exists two natural estimators of the σ_i^2 based on the previous posterior density as

$$(7) \qquad \begin{aligned} \hat{\sigma}^2_{i,1} &= \exp(E \log \sigma_i^2 | S_i^2) \\ &= \exp(M_{v,i}(\log S_i^2 - m) + (1-M_{v,i})\mu_v), \end{aligned}$$

and

$$(8) \qquad \hat{\sigma}^2_{i,2} = E(\sigma_i^2|S_i^2) = \hat{\sigma}^2_{i,1} \exp\left(\frac{M_{v,i}\sigma^2_{ch}}{2}\right).$$

The estimator $\hat{\sigma}^2_{i,2}$ is based on the exact posterior expectation. When constructing a confidence interval for θ_i, Hwang, Qiu and Zhao (2009) prefers the estimator $\hat{\sigma}^2_{i,1}$ because it produces a shorter, in other words, more efficient interval than the other one when both guaranteeing the *empirical* Bayesian coverage probability. Here we will use the estimator $\hat{\sigma}^2_{i,1}$, written as $\hat{\sigma}_i^2$, to construct the estimator of θ_i later in this paper. Practically speaking, there is little difference for these two approaches in estimating θ_i's.

The exact posterior distribution of σ_i^2 given the observation (X_i, S_i^2) also depends on X_i and has no explicit form. We approximate this posterior by assuming that it depends solely on S_i^2. This approximation is practically and intuitively reasonable.

Having the variance shrinkage estimator, we now turn to the estimation of θ_i. Recall that

$$E(\theta_i | X_i, \sigma_i^2) = M_i X_i + (1 - M_i)\mu.$$

Then we can estimate θ_i by

$$\hat{\theta}_i = \hat{M}_i X_i + (1 - \hat{M}_i)\mu, \tag{9}$$

where $\hat{M}_i = \frac{\tau^2}{\tau^2 + \hat{\sigma}_i^2}$.

The above estimator is not the exact Bayes estimator $E(\theta_i | X_i, S_i^2)$ which minimizes the Bayes risk when using the loss function (2). However, (9) has the advantage of having instantaneous computation. It would be very interesting to derive some analytic results regarding the relations between (9) and the exact Bayes estimator.

3. ESTIMATING THE HYPER-PARAMETER

In Section 2, we have proposed the point estimator of θ when assuming known hyper parameters μ, τ^2, μ_v and τ_v^2, which, in practice, are unknown. To avoid any subjective choice, we incorporate the *empirical* Bayes approach by estimating the parameters through the data. The estimation resembles the calculation in Hwang, Qiu and Zhao (2009) where the method of moments is used.

Firstly, we estimate the hyper parameters μ_v, τ_v^2 corresponding to the variances component. In model (6), it is assumed that

$$\log S_i^2 - m_i | \log \sigma_i^2 \sim N(\log \sigma_i^2, \sigma_{ch,i}^2),$$

and

$$\log \sigma_i^2 \sim N(\mu_v, \tau_v^2).$$

Consequently, $E(\log S_i^2 - m_i) = \mu_v$. We estimate μ_v by

$$\hat{\mu}_v = \frac{1}{p} \sum_i (\log S_i^2 - m_i).$$

Further, $E(\log S_i - m_i)^2 = \mu_v^2 + \tau_v^2 + \sigma_{ch,i}^2$. We thus estimate τ_v^2 by

$$\hat{\tau}_v^2 = \left(\frac{1}{p} \left(\sum_i (\log S_i^2 - m_i)^2 - \sigma_{ch,i}^2 - \hat{\mu}_v^2 \right) \right)_+,$$

and

$$\hat{M}_{v,i} = \frac{\hat{\tau}_v^2}{\hat{\tau}_v^2 + \sigma_{ch,i}^2}.$$

Providing with the estimation of the hyper parameter corresponding to the variances and formula (7), we derive an *empirical* Bayes estimator of σ_i^2 as

$$\hat{\sigma}_{EB,i}^2 = \exp(\hat{M}_{v,i}(\log S_i^2 - m_i) + (1 - \hat{M}_{v,i})\hat{\mu}_v). \tag{10}$$

When the degrees of freedom $d_i = d$ $(i = 1, \ldots, p)$, the estimator (10) can be written as

$$\hat{\sigma}_{EB,i}^2 = \exp(\hat{M}_v(\log S_i^2 - m) + (1 - \hat{M}_v)\hat{\mu}_v),$$

where

$$m = \psi(d/2) - \log(d/2), \sigma_{ch}^2 = \psi'(d/2)$$

and

$$\hat{M}_v = \left(1 - \frac{(p-3)\sigma_{ch}^2}{\sum_i (\log S_i^2 - \overline{\log S_\cdot^2})^2} \right)_+$$

which is the exponential Lindley–James–Stein's estimator introduced in Cui et al. (2005).

Note that the *empirical* Bayes estimator (10) can be written as

$$\hat{\sigma}_{EB,i}^2 = \left(\left(\prod_i \frac{S_i^2}{e^{m_i}} \right)^{1/p} \right)^{1 - \hat{M}_v} \left(\frac{S_i^2}{e^{m_i}} \right)^{\hat{M}_v}.$$

This indicates that $\hat{\sigma}_{EB,i}$ shrinks the observation $S_i^2 / \exp(m_i)$ towards their geometric mean $(\prod_i S_i^2 / \exp(m_i))^{1/p}$, resulting in a variance shrinkage estimator.

The next step is to estimate the hyper parameters μ and τ^2 of the means θ_i's. Since

$$X_i | \sigma_i^2 \sim N(\mu, \sigma_i^2 + \tau^2),$$

we estimate μ by the weighted average as

$$\hat{\mu} = \sum \frac{X_i / \hat{\sigma}_{EB,i}^2}{\sum 1 / \hat{\sigma}_{EB,i}^2}.$$

Further, since $E(X_i - \mu)^2 | \sigma_i^2 = \sigma_i^2 + \tau^2$, Hwang, Qiu and Zhao (2009) estimated τ^2 as

$$\hat{\tau}^2 = \left(\frac{\sum (X_i - \hat{\mu})^2 - \hat{\sigma}_{EB,i}^2}{p} \right)_+.$$

However, the estimator $\hat{\sigma}_{EB,i}^2$ is not an unbiased estimator of σ_i^2, resulting in an inconsistent estimator of τ^2 as $p \to \infty$. In order to remedy this, we estimate τ^2 by using

$$\hat{\tau}^2 = \left(\frac{\sum (X_i - \hat{\mu})^2 - S_i^2 \exp(-m_i - \sigma_{ch,i}^2 / 2)}{p} \right)_+,$$

due to the fact that

$$E S_i^2 | \log \sigma_i^2 = \sigma_i^2 \exp\left(m_i + \frac{\sigma_{ch,i}^2}{2} \right).$$

When assuming that $\frac{S_i^2}{\sigma_i^2} \sim \frac{\chi_{d_i}^2}{d_i}$, other than the log-normal assumption, we remove the term $\exp(-m_i - \sigma_{ch,i}^2/2)$ when estimating τ^2 in order to obtain a consistent estimator of τ^2.

With the estimators of all the hyper parameters available, we propose the estimator for θ_i as

$$\hat{\theta}_{SS,i} = \hat{M}_{EB,i} X_i + (1 - \hat{M}_{EB,i})\hat{\mu}, \tag{11}$$

where $\hat{M}_{EB,i} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2_{EB,i}}$.

It is worthy noting that the estimator (11) is a shrinking-mean estimator for it shrinks the observation X_i towards the weighted average $\hat{\mu}$. Additionally, the estimator $\sigma^2_{EB,i}$, as defined in (10), is a variance shrinkage estimator as it shrinks the observation $S_i^2/\exp(m_i)$'s towards their geometric mean. Therefore, we call the estimator (11) as double shrinkage estimator $\hat{\theta}_{SS}$. When estimating θ_i, especially the hyperparameters μ, μ_v, τ^2, and τ_v^2, we borrow the strength from the seemly unrelated observations X_j, and S_j^2 where $j \neq i$.

4. SIMULATION STUDY

In Section 3, we have proposed the double shrinkage estimator $\hat{\theta}_{SS}$ of θ which shrinks both the means and variances. Alternatively, if replacing the variance σ_i^2 simply by S_i^2 and through replicating the procedure above, one can propose an alternative estimator $\hat{\theta}_{SM,i}$ as

$$\hat{\theta}_{SM,i} = \hat{M}'_{EB,i} X_i + (1 - \hat{M}'_{EB,i})\hat{\mu}', \tag{12}$$

where $\hat{M}'_{EB,i}$ and $\hat{\mu}'$ are derived similarly as in Section 3 with $\sigma^2_{EB,i}$ replaced by S_i^2. Such an estimator is called shrink-mean-alone estimator for it shrinks X_i towards the weighted average $\hat{\mu}$ and has no variance shrinkage. Like $\hat{\theta}_{SS,i}$, this estimator also has the borrowing strength effect.

In addition, one can estimate θ simply by $\hat{\theta}_{NS} = X$ which neither shrinks nor borrows strength from other observations. In this section, we use simulation studies to calculate the Bayes risk under various parameter settings and model settings. The loss function is defined in (2).

In Figures 1 and 3, random numbers are generated according to the genuine *Log-Normal* model. We have simulated the Bayes risk of the estimators $\hat{\theta}_{SS}$ as in (11), $\hat{\theta}_{SM}$ as in (12), and $\hat{\theta}_{NS} = X$ with the dimension p being $2,000$. Their risks are represented by curves with markers as Diamonds, Circles, and Crossings respectively.

In Figure 1, the degrees of freedom d_i's are randomly selected among 2, 3, 4 and 5. The hyper parameters $\mu = \mu_v = 0$ and τ_v^2 varies among 0, 0.25, 0.5, and 1 from the top to the bottom. The Bayes risk is plotted against $M = \frac{\tau^2}{\tau^2 + \exp(\mu_v + \sigma_v^2/2)}$, varying from 0 to ∞. In other words, the hyper parameter τ^2 goes from 0 to ∞. In Figure 3, all the degrees of freedom d_i's equal 2.

From these two figures, it is seen that $\hat{\theta}_{SS}$ always dominates both $\hat{\theta}_{SM}$ and $\hat{\theta}_{NS}$ for different hyper parameter settings. Both the shrinkage estimators substantially improve

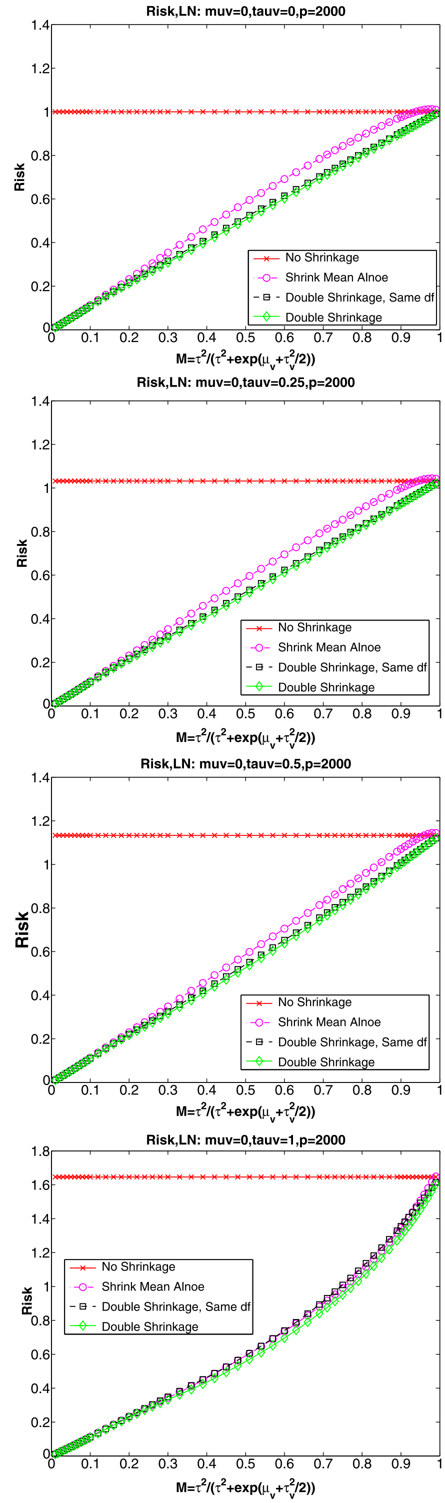

Figure 1. These figures are the Bayes risks of four point estimators with the dimension $p = 2,000$. The random numbers are generated according to the genuine Log-Normal model. The degrees of freedom are randomly chosen from 2 to 5. The hyper parameter setting are $\mu = 0$, $\mu_v = 0$. The τ_v^2 varies from 0, 0.25, 0.5, to 1 from the top to the bottom. We plot the risk against $M = \frac{\tau^2}{\tau^2 + \exp(\mu_v + \tau_v^2)}$ which goes from 0 to 1.

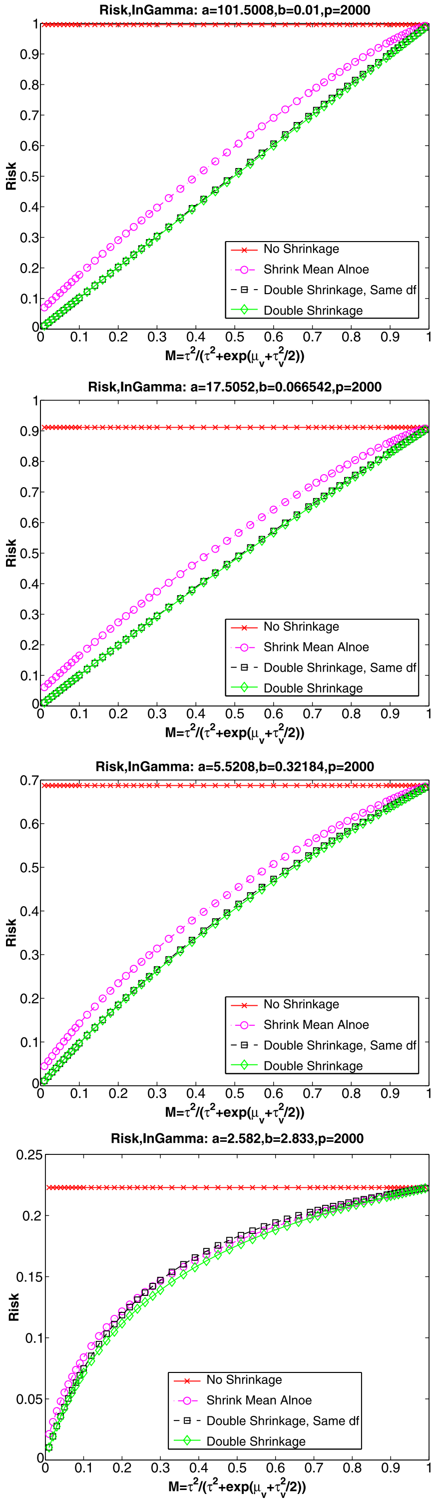

Figure 2. These figures are the Bayes risks of four point estimators with the dimension $p = 2,000$. The degrees of freedom are randomly chosen among 2, 3, 4, and 5. The random numbers are generated according to the inverse gamma model. The hyper parameters a and b are chosen according to (14). We plot the risk against M which goes from 0 to 1. The hyper parameter μ is 0.

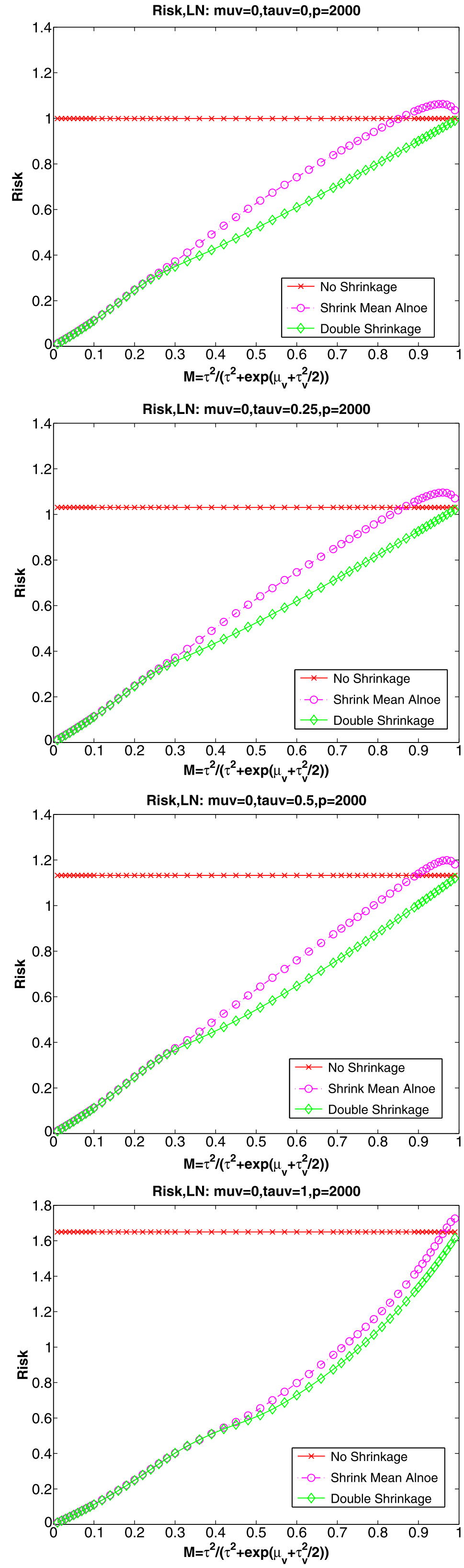

Figure 3. The parameter settings are the same as that in Figure 1. The only difference is that the degrees of freedom d_i's are the same and equal to 2.

$\hat{\theta}_{NS}$ when τ^2 is close to 0. This indicates that shrinking-mean is important when all the true means are close. On the other hand, when τ^2 goes to infinity, the Bayes risk of double shrinkage estimator converges to the risk of $\hat{\theta}_{NS}$ from below. Surprisingly, the Bayes risk of $\hat{\theta}_{SM}$ exceeds the level of that of the no shrinkage estimator X for large τ^2 and small degrees of freedom. We further notice that $\hat{\theta}_{SS}$ dominates $\hat{\theta}_{SM}$ under every case. The improvement is significant especially for small τ^2 when the variances σ_i^2's are close to each other. When τ_v^2 is large, $\hat{\theta}_{SM}$ performs nearly the same as $\hat{\theta}_{SS}$.

In conclusion, the simulation results show that $\hat{\theta}_{SS}$ dominates both $\hat{\theta}_{SM}$ and $\hat{\theta}_{NS}$ under the log-normal model.

In Figures 2 and 4, we have generated the random number according to the *inverse gamma* model with the last two equation of model (6) being replaced by

$$\begin{cases} S_i^2|\sigma_i^2 \sim \sigma_i^2 \frac{\chi^2_{d_i}}{d_i}; \\ \sigma_i^2 \sim \text{ Inverse Gamma}(a, b). \end{cases} \tag{13}$$

In other words, $(\sigma_i^2)^{-1}$ has a *Gamma* distribution with parameters a and b. See Berger (1985).

In these simulations, the dimension $p = 2,000$. The degrees of freedom d_i's are randomly chosen among 2, 3, 4, and 5 in Figure 2 and set to be 2 in Figure 4. The hyper parameters a and b are chosen such that

(14)
$$E\sigma_i^2 = E(\exp(N(\mu_v, \tau_v^2))), Var(\sigma_i^2) = Var(\exp(N(\mu_v, \tau_v^2)))$$

where $\mu_v = 0$, τ_v^2 varies among 0, 0.25, 0.5, and 1 from the top to the bottom in each figure.

In all these studies, the Bayes risk of $\hat{\theta}_{SS}$ is smaller than that of $\hat{\theta}_{SM}$, which is smaller than that of $\hat{\theta}_{NS}$. The improvement of $\hat{\theta}_{SM}$ over $\hat{\theta}_{NS}$ is very substantial for small τ^2. When $\tau^2 \to \infty$, in other words, $M \to 1$, the Bayes risk of the shrinkage estimators converge to the risk of no shrinkage estimator from below. For small τ_v^2, the double shrinkage estimator improves shrink-mean-alone estimator especially for small degrees of freedom. In Figure 5, we have simulated the Bayes risk of the estimators based on the *inverse gamma* model with $\tau_v^2 = 0$ and equal degrees of freedom d, which varies among 2, 6, 10, and 20. The discrepancy between $\hat{\theta}_{SS}$ and $\hat{\theta}_{SM}$ gets smaller when the degrees of freedom increases. Nevertheless, $\hat{\theta}_{SS}$ always dominates $\hat{\theta}_{SM}$.

In both Figures 1 and 2 when the degrees of freedom d_i's are different across the observations, we have plotted the risk of the double shrinkage estimator when simply putting all the degrees of freedom to be the $\min_{1\le i\le p} d_i$. This approach was taken by Hwang, Qiu and Zhao (2009) in constructing the confidence interval for each parameter θ_i. The Bayes risk of this estimator is represented by the lines with Squares in these figures. It turns out that it is dominated by the new estimator $\hat{\theta}_{SS}$.

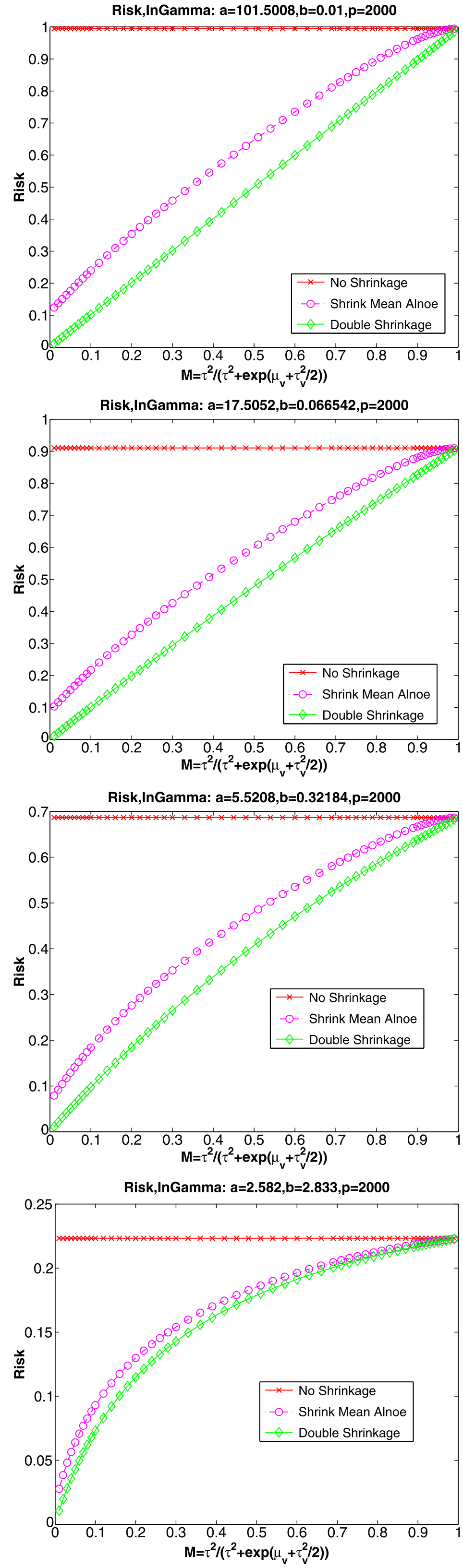

Figure 4. The parameter settings are the same as that in Figure 2. The only difference is that the degrees of freedom d_i's are the same and equal to 2.

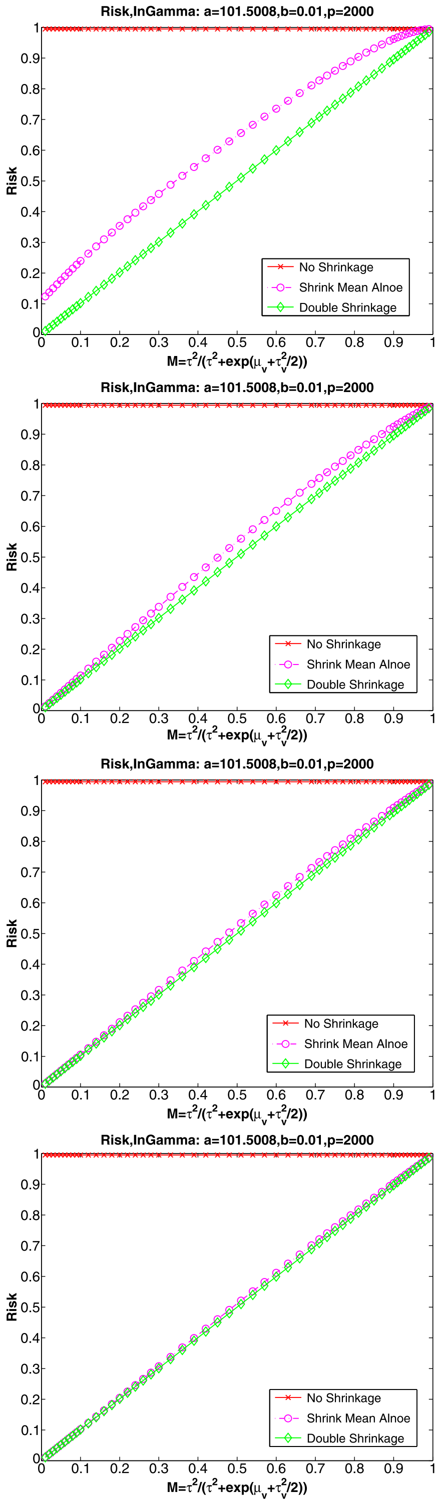

Figure 5. In this study, the data are generated according to the inverse Gamma model with p being 2,000. The hyper parameters $\mu = \mu_v = 0, \tau_v = 0$. The parameters a and b are chosen according to (14). In each graph, the degrees of freedom are the same and equal to 2, 5, 10, and 20 from the top to the bottom. The discrepancy between the risk of $\hat{\theta}_{SS}$ and $\hat{\theta}_{SM}$ increases when d decreases.

We have used sum of squared error loss (2) in the simulation. One of the referees pointed out that the same conclusion holds when we are only interested in estimating one parameter, saying θ_1, with the one-dimensional squared error loss $L(\theta_1, \hat{\theta}_1) = (\hat{\theta}_1 - \theta_1)^2$. By claiming domination of the one-dimensional loss, the estimator we proposed is better than others under any weighted squared error loss.

All these simulation studies indicate that the double shrinkage estimator dominates the shrink-mean-alone estimator and no shrinkage estimator. In addition, $\hat{\theta}_{SS}$ always performs the best for these two different model settings. This demonstrates that the new procedure is robust in some sense. Thus, the double shrinkage estimator $\hat{\theta}_{SS}$ is strongly recommended.

5. REAL DATA ANALYSIS

We apply different estimators to an Affymetrix Control data set, the golden spike-in data set of Choe et al. (2005). All the parameters in this data set are pre-chosen and known. Therefore, it can be used to check different statistical procedures, such as the performance of confidence intervals in Hwang, Qiu and Zhao (2009) and point estimators as stated in this article.

In this section, we will calculate the risks of estimators $\hat{\theta}_{SS}$, $\hat{\theta}_{SM}$, and $\hat{\theta}_{NS}$. We download the data from `http://www.elwood9.net/spike`. After taking the $\log_2$ transformation, we fit the data to a one-way ANOVA model with the number of genes p being 14,010. There are 6 replicates for each gene, three from each of the control and treatment group. Let

$$X_i = \bar{Y}_{i1} - \bar{Y}_{i2}, S_i^2 = \sqrt{s_{1i}^2/3 + s_{2i}^2/3}.$$

The degrees of freedom are calculated according to Satterthwaite approximation. In each study, we randomly sample 2,000 observations among all genes with replacement and then estimate the true parameters by different estimators and calculate corresponding losses. We replicate this study 2,000 times and calculate the risk by taking the average of the losses. (See Table 1.) The risk of $\hat{\theta}_{SS}$ is about 92.8% of that of $\hat{\theta}_{SM}$, and 31.9% of that of $\hat{\theta}_{NS}$.

We have also calculated the standard deviation of the difference of the losses between an estimator $\hat{\theta}$ and the double shrinkage estimator $\hat{\theta}_{SS}$ and displayed it in the last row of Table 1. Clearly, the double shrinkage estimator improves shrinking-mean-alone estimator significantly which improves $\hat{\theta}_{NS}$ significantly.

Along with the simulation studies we have presented in Section 4, we can state that the double shrinkage estimator $\hat{\theta}_{SS}$ is better than the shrinking-mean-alone estimator $\hat{\theta}_{SM}$, which is better than the estimator $\hat{\theta}_{NS}$ without any shrinkage.

The code for the double shrinkage estimator can be downloaded from `http://astro.temple.edu/~zhaozhg/publications.html`.

Table 1. The risk comparison of the estimators for golden spike-in data set. Within this table, $DL = L(\hat{\theta}, \theta) - L(\hat{\theta}_{SS}, \theta)$, the difference between the losses of any estimator $\hat{\theta}$ and the double shrinkage estimator $\hat{\theta}_{SS}$

Estimator	$\hat{\theta}_{NS}$	$\hat{\theta}_{SM}$	$\hat{\theta}_{SS}$
Risk	0.3243	0.1115	0.1035
E(DL)	0.2208	0.0080	0
Std(DL)	0.0107	0.0032	0

6. CONCLUSION AND DISCUSSION

In this article, we have constructed a new estimator when assuming the observation X_i follows a normal distribution with an unknown and unequal variance σ_i^2. The estimator is based on the model (6), a general form of *Log-Normal* model firstly proposed by Hwang and Liu (2010) and further studied in Hwang, Qiu and Zhao (2009). In these two papers, they have constructed the double shrinkage testing procedure and confidence interval by using the *empirical* Bayes approach. We adopt the *empirical* Bayes approach to construct a point estimator for multiple parameters which shrinks both the means and variances. We call this estimator $\hat{\theta}_{SS}$ the double shrinkage estimator.

We further analyze the performance of $\hat{\theta}_{SS}$, comparing with the shrinking-mean-alone estimator $\hat{\theta}_{SM}$ and the estimator $\hat{\theta}_{NS} = X$ with no shrinkage. Both extensive simulation studies and a real data analysis indicate that $\hat{\theta}_{SS}$ performs uniformly better than the other two. We thus strongly recommend the new approach.

This article proposes a new methodology in estimating under the condition of *heteroscedasticity*. However, much work is needed. For instance, we would like to know how $\hat{\theta}_{SS}$ relates to the exact Bayes estimator in terms of the relative savings loss introduced in Efron and Morris (1973). However, the proof of any analytic results will be very difficult and heavily involved due to the unknown and unequal variances. We leave this for the future research.

ACKNOWLEDGMENTS

Thanks go to Harrison Huibin Zhou for inviting me to write the paper. I am especially grateful for the ideas, discussions, and encouragements of my advisor, J. T. Gene Hwang. The author also thanks the referees for their valuable comments which improved the paper.

Received 29 January 2010

REFERENCES

Berger., J. O. *Statistical decision theory and Bayesian analysis*. Springer-Verlag, New York, second edition, 1985. MR0804611

Casella, G. and Hwang, J. T. Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *Journal of the American Statistical Association*, 78 (383): 688–698, 1983. ISSN 01621459. MR0721220

Casella, George and Hwang, J. T. Shrinkage confidence procedures. *Statistical Science*, 2010. To appear.

Choe, S. E., Bouttros, M., Michelson, A. M., Chruch, G. M. and Halfon, M. S. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*, 6 (2): R16.1–16, 2005.

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

Efron, B. and Morris, C. N. Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68: 117–130, 1973. ISSN 0162-1459. MR0388597

Fourdrinier, Dominique, Strawderman, William E. and Wells, Martin T. Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *Journal of Multivariate Analysis*, 85 (1): 24–39, 2003. ISSN 0047-259X. MR1978175

He, K. Parametric empirical Bayes confidence intervals based on James-Stein estimator. *Statist. Decisions*, 10 (1–2): 121–132, 1992. ISSN 0721-2631. MR1165708

Hwang, J. T. and Liu., P. Optimal tests shrinkage both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology*, 9 (1): Article 36, 2010.

Hwang, J. T., Qiu, J. and Zhao, Z.. Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71 (1): 265–285, 2009.

James, W. and Stein, C. Estimation with quadratic loss. *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 361–379, 1961. MR0133191

Lindley, D. V. Discussion on Professor Stein's paper. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 24: 285–287, 1962.

Morris, C. N. Parametric empirical Bayes confidence intervals. *Scientific inference, data analysis, and robustness (Madison, Wis., 1981)*, 48: 25–50, 1983a. MR0772762

Morris, C. N. Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.*, 78 (381): 47–65, 1983b. ISSN 0162-1459. With discussion. MR0696849

Qiu, J. and Hwang, J. T. Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63 (3): 767–776, 2007. MR2395714

Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 (1): Article 3, 2004. MR2101454

Zhigen Zhao
346 Speakman Hall
Temple University
Philadelphia, 19122
USA