Empirical Bayes confidence intervals for selected parameters in high dimension with application to microarray data analysis

J. T. Gene Hwang
Department of Mathematics, Department of Statistics
Cornell University
Ithaca, NY 14853, USA

Zhigen Zhao
Department of Statistics
Temple University
Philadelphia, PA 19122, USA

August 9, 2011

Abstract

Modern statistical problems often involve a large number of populations and hence a large number of parameters that characterize these populations. The interest often lies in studying and making inference regarding the parameters corresponding to the populations that have been selected, called the *selected parameters*. The current statistical practices either apply a traditional procedure assuming there was no selection, a practice not valid, or use the Bonferroni type of procedures which are valid but very conservative and often non-informative.

In this paper, the authors propose valid and sharp confidence intervals which allow scientists to select parameters and to make inference for the selected parameters using the same data. These type of confidence intervals allow the users to zero in on the most interesting selected parameters.

The validity of confidence intervals is defined in terms of an *empirical* Bayes coverage probability criterion. This requires that the Bayes coverage probabilities with respect to a class of distributions for the parameters to be no less than a nominal level. The *empirical* Bayes coverage probabilities are exactly the frequentist's coverage probabilities for a mixed model where the key parameters the intervals try to capture are the random effects.

In this paper, we assume that the observation are normally distributed with unequal and unknown variance. The parameters have been selected according to the magnitude of the t-values. We then construct the *empirical* Bayes confidence intervals for these selected parameters. Our interval has good *empirical* Bayes coverage probability, and has sharp average lengths. Our interval is also applied to the spike-in data and shown to be better than all the alternatives.

Key words: *Empirical* Bayes, selection, multiple

1 Introduction

Modern statistical problems often involve a large number of populations which, for instance, are characterized by parameters $\theta_1, \theta_2, \cdots, \theta_p$ (see Efron [2008]). A case in point is the microaray data analysis where the population is the gene and the parameter is the differential gene expression θ_i, for the i-th gene. The genes are often selected based on data. The interest lies in studying and making inference regarding the parameters of the selected genes using the same data. The focus of this paper is to construct confidence intervals for the parameters corresponding to the selected populations, called *selected parameters*. The notation and precise definition of these selected parameters are given in the last paragraph of Section 2.

Confidence intervals are important, since they can obviously be used to do the hypothesis testing and to address statistical significance. However, statistical significance alone is not enough. Genes with expression levels shown to be statistically significantly different from zero may have small differential expressions (small effect sizes), having no biological significance. While hypothesis testing can not be used to assess the biological significance, confidence intervals can, since they specify the range of parameters of these

expression levels. The importance of estimating sizes of the differential expressions has been argued forcefully in Montazeri et al. [2010]. Here we work for a more ambitious goal of constructing a confidence interval. A confidence interval, however, is better than a point estimator, since unlike the latter, the former provides an answer taking into account of estimation error with a guarantee of a high coverage probability. This can be used to assess the practical (or biological) significance.

Although the paper focuses on microarray, selection using data is prevalent in statistical applications especially in problems involving many parameters. See for example, two highly publicized researches on clinical trials: Giovannucci et al. [1995] and Rossouw et al. [2002]. The goal of this paper aims at providing the post-selection analysis, an area which is becoming active in modern statistics research, although many researchers, including e.g. Berk et al. [2010a] and Berk et al. [2010b], focus on variable selection in the regression context. The problem we work on and the technique we develop may appear to be unrelated to their regression problem, our technique can most likely be carried over to deal with their problem.

What is a statistically valid way of further analyzing the selected genes regarding their gene expression levels? One possible answer is to experiment again with the selected genes and take more data and use the new data to do statistical inference. This is statistically sound. It is however not always possible due to various reasons such as the constraint of budget or time. In such a case, it is very desirable to be able to perform the statistical inference based only on the data at hand.

Currently, scientists including Giovannucci et al. [1995] and Rossouw et al.

[2002], often proceed with a naive approach. Namely, one constructs a standard confidence interval to the selected parameters, pretending there were no selection. The naive approach has severe bias, and the corresponding confidence interval could have extremely low coverage probability. The alternative approach is to use Bonferroni type of procedures which are valid under any selection, but could be very conservative and non-informative. It is very important to develop statistical confidence intervals which are valid even after selection and which are not as conservative as the Bonferroni type procedures.

It is well known that for a single prior, the Bayes approach (and Bayes posterior coverage probability) is unaffected by the prior selection based on data. However, a single prior is unrealistic and hence it is important to consider a class of priors. It may seem that for a class of priors, the Bayes coverage probabilities can naturally be unaffected by the prior selection. However, this more realistic approach actually poses much technical difficulty.

In this paper, under any statistical selection, it is shown that it is possible to construct $1-\alpha$ *empirical* Bayes confidence intervals for selected parameters in the sense of Morris [1983]. Namely,

$$\text{the Bayes coverage probability} \geq 1-\alpha, \quad \text{for a class of priors } \pi \in \Pi. \tag{1.1}$$

Note the difference between the *Bayes* interval and the *empirical* Bayes interval. The Bayes interval has the guaranteed Bayes coverage probability for one *single* prior whereas the *empirical* Bayes interval guarantees good

Bayes coverage probability for a class of prior distributions which is more realistic.

The *empirical* Bayes criterion (1.1) is becoming increasingly appealing in modern applications which typically involve a large number of θ_i's. Facing many θ_i's, scientists often think and communicate in terms of the distribution of θ_i's. By pooling the observations from all the populations, it is possible to check whether the assumed probability models of θ_i and X_i fit the data. See, for example, Qiu and Hwang [2007]. Using the estimated model, one can then develop more efficient statistical inference tools valid for a class of models as what we are about to do.

Another compelling reason to deal with criterion (1.1) is that it is exactly the frequentist' criterion when θ_i's are the random effects of random effect or mixed effect models, where θ_i's are assumed to have a distribution π in a class Π. Assumption of such random effect or mixed effect models is especially appropriate if the θ_i's correspond to experiment units sampled from a population.

Earlier remarkable attempts on this problem include Cohen and Sackrowitz [1982], Venter [1988], and Venter and Steel [1991] who, under normal assumption, constructed confidence intervals with good frequentist coverage probabilities for the selected parameter corresponding to the largest observation.

For the practically important problem of dealing with various selected parameters, Qiu and Hwang [2007] constructed much shorter individual intervals and simultaneous intervals for the selected parameters. Their intervals are demonstrated to be $1-\alpha$ *empirical* Bayes confidence intervals, i.e.

satisfying (1.1). The class of priors assumed consists of any mixture of a normal prior with zero, which is shown to fit the data they considered.

Assume that X_i's are independent estimates of θ_i's and X_i is distributed according to $N(\theta_i, \sigma^2)$. Based on X_i's, they provided very sharp simultaneous confidence intervals for the selected parameters. Their intervals are certainly much sharper than the Bonferroni type intervals and even sharper than the naive intervals, a surprise due to the *empirical* Bayes effect. Furthermore, their intervals are based on shrinking the data. They perform well for all θ_i's and especially well for θ_i's that have a sparse structure, namely, most θ_i's being zero or nearly zero, which occurs often in microarray data and many other modern applications involving high dimension. It is no surprise that when applied to a real data set, the common length of these intervals for 89 selected genes are 18% shorter than the Bonferroni type interval for even a small $p = 1,285$.

However, the intervals proposed by Qiu and Hwang [2007] can only be applied to populations with identical variances, i.e., the variance of X_i being all equal. The data set discussed earlier in Qiu and Hwang [2007] fits the equal variance model well. However, typically, the variances depend on genes. Hence it is important to work on the case where the variances of X_i, denoted as σ_i^2, are not identical.

To relax such an assumption while providing a computationally simple solution proves to be extremely difficult. This is because one needs to consider a class of priors on (θ_i, σ_i^2). In contrast, Qiu and Hwang [2007] consider a prior on θ_i's only. As a parenthetical remark, computational simplicity is important since the data could be massive.

The confidence interval construction problem for the selected parameters is related to that for the unselected θ_i's which Morris [1983] considered where he assumed that σ_i^2's are known. See also Morris and Tang [2010] for a computationally less intensive procedure and Everson and Morris [2000] for a multivariate generalization. These procedures only shrink the means but not the variances. More recently, Hwang, Qiu, and Zhao [2009] constructed an $1-\alpha$ *empirical* Bayes confidence interval for θ_i when σ_i's are unknown but estimable. Zhao [2010] constructed an *empirical* Bayes estimator for θ_i's under the same model settings which is shown to dominate the naive estimator X_i's by extensive numerical calculations. These two papers proposed procedures which shrink both the means and variances. In order to construct an explicit solution for the case where σ_i^2's are unknown but estimable, we revisited Morris's procedure which aims at constructing intervals for the unselected parameter θ_i. In this paper, a novel notion on the second order correctness was introduced (see Section 3). This allows us to construct second order correct confidence interval (for known σ_i^2's) without employing a second stage prior on the hyper-parameters of the normal prior on the θ_i's as Morris had done. Lengthy algebraic calculation shows that Morris' interval is not second order correct. Numerical simulations (in Figure 1) also show that Morris' interval fails to have good coverage probabilities for some parameter settings. The second order calculation is used to modify Morris' interval which has good *empirical* Bayes coverage probabilities. The concept of second order correctness forms the basis for constructing *empirical* Bayes valid intervals for the selected parameters when σ_i^2's are unknown and estimable.

A previous major development in the interval estimation for the selected parameters is the work of Benjamini and Yekutieli [2005], where a new criterion, False Coverage Rate (FCR) has been proposed. Assume that R parameters are selected and the corresponding intervals are constructed. Let V be the number of the constructed intervals failing to cover the true parameters. The false coverage rate is defined as the expectation of $\frac{V}{R}I(R > 0)$, where $I(R > 0)$, an indicator function, equals zero or one depending on whether R is positive or zero. They proposed and constructed intervals with controlled FCR under a frequentist paradigm. Such intervals are abbreviated as BY intervals in this paper and are sharper than the Bonferroni intervals in average length.

In 2010, Zhao and Hwang [2010] considered the Bayes FCR with respect to a prior π, which is defined as $E_\pi FCR$, the expectation of FCR against the prior π. A typical Bayesian approach aims at controlling this Bayes FCR corresponding to a single prior π. However, a more realistic criterion called the *empirical* Bayes criterion, is to control

$$E_\pi FCR \leq \alpha \quad \text{for a class of priors } \pi \in \Pi. \tag{1.2}$$

Note that when θ_i's are random effects (such as in an ANOVA model) distributed according to π belonging to Π, the *empirical* Bayes FCR criterion is the frequentist FCR criterion. When (1.2) holds, the *empirical* Bayes FCR is said to be controlled at the α-level.

The proposed individual $1 - \alpha$ *empirical* Bayes intervals constructed for the selected parameters are shown to have an *empirical* Bayes FCR

controlled at α-level with respect to a class of normal distribution, and are sharper than the intervals constructed by Benjamini and Yekutieli [2005].

2 Models and mathematical formulation

Assume a canonical form where observation (X_i, S_i^2) is a sufficient statistic from the i-th population, and is an unbiased estimator of (θ_i, σ_i^2). It is assumed that

$$X_i \sim N(\theta_i, \sigma_i^2), \theta_i \sim N(0, \tau^2), \tag{2.1}$$

and S_i^2 is an unbiased estimator for σ_i^2, where the distribution of S_i^2 and σ_i^2 are to be specified below. If an ANOVA model is assumed, X_i is typically taken to be the ANOVA unbiased estimator of θ_i. When θ_i's are the random effects in a mixed ANOVA model, the mean of θ_i is typically assumed to be zero as in (2.1). See a remark in the Appendix concerning a nonzero mean.

To select the most significant populations (or genes), traditionally one selects the largest t-statistics ($\frac{|X_i|}{S_i}$). Consider first just one selected parameter, before considering simultaneous intervals for several selected parameters later. Hence, this population produces the largest t-statistic among all populations. Let $\theta_{(p)}$ be the θ corresponding to this selected population. Hence $\theta_{(p)}$ is a function of θ_j's and the data $\frac{|X_i|}{S_i}$. Precisely, $\theta_{(p)} = \theta_j$ where j is such that $\frac{|X_j|}{S_j} = \max_i(\frac{|X_i|}{S_i})$. Note that $\theta_{(p)}$ may not equal the maximum of θ_i's. In this proposal, $\theta_{(p)}$ is called the *selected parameter* corresponding to the population with the largest $|X_i|/S_i$. Similarly, $\theta_{(p-1)}$ is the θ_j where $|X_j|/S_j$ corresponds to the second largest t statistic. The rest of the selected parameters, such as $\theta_{(p-2)}, \cdots, \theta_{(p-k)}$ can be similarly defined. Construct-

ing confidence interval for $\theta_{(p)}$ proves to be a very difficult problem. After three years of extensive research, the authors finally found a key to this problem. It is based on the following theorem, describing a concept called *second order correctness*.

3 Second order calculation

3.1 Known but unequal variances- second order correctness calculations

In this section, intervals are first constructed for θ_i, which is the parameter without prior selection. Assume that X_i and θ_i satisfy the canonical form (2.1), where σ_i^2 is assumed to be known. This problem has been addressed and a solution has been provided in Morris [1983] and He [1992]. Morris considered a class of normal priors with variance τ^2. To estimate τ^2, Morris introduced another prior on τ^2. Here the authors treat τ^2 as a parameter and use observations $X_1, \cdots, X_p$ to estimate τ^2 without introducing a prior. The solution can be more readily generalized to the case where σ_i's are unknown and where it is necessary to put a prior on σ_i^2. The approach then leads to shrinking the estimate of σ_i^2. Without such shrinkage of variances (as well as means), the resultant procedure will not work well for the selected parameters for unknown σ_i^2. This is because the selection procedure tends to choose S_i^2 which are small, making the intervals too narrow. The concept of the second order correctness also plays a key role in understanding why some approaches work and some do not, in both selected and unselected cases.

Under the classical model (2.1), the classical Bayesian calculation shows that

$$\theta_i|\vec{X} \sim N(M_i X_i, M_i\sigma_i^2), \tag{3.1}$$

where $\vec{X} = (X_1, \cdots, X_p)'$ and $M_i = \frac{\tau^2}{\sigma_i^2+\tau^2}$. A reasonable estimator for M_i is

$$\hat{M}_i = \frac{\hat{\tau}^2}{\sigma_i^2 + \hat{\tau}^2}, \quad \text{where} \quad \hat{\tau}^2 = \max(\frac{\sum_i (X_i^2 - \sigma_i^2)}{p}, 0). \tag{3.2}$$

From (3.1),

$$\theta_i - \hat{M}_i X_i|\vec{X} = \sqrt{M_i}\sigma_i Z_1 + (M_i - \hat{M}_i)X_i, \tag{3.3}$$

where Z_1 is $N(0,1)$, independent of X_i. The following theorem can be derived using (3.3) and the obvious formula below:

$$(M_i - \hat{M}_i)X_i = \frac{(\tau^2 - \hat{\tau}^2)\sigma_i^2}{(\sigma_i^2 + \tau^2)(\sigma_i^2 + \hat{\tau}^2)} X_i. \tag{3.4}$$

Theorem 3.1. *Assume that $0 < \sigma_i^2 < B$ for all i and finite number B. Then conditioning on X_i, the distribution of $\theta_i - \hat{M}_i X_i$ has the same distribution as*

$$A_{1,i}Z_1 + A_{2,i}Z_2 + O_P(\frac{\ln p}{p}), \tag{3.5}$$

where

$$A_{1,i} = \sqrt{M_i}\sigma_i, \quad \textit{and} \quad A_{2,i} = \frac{(1-M_i)^2}{\sigma_i^2} X_i \sqrt{2\sum_i (\sigma_i^2 + \tau^2)^2/p^2}, \tag{3.6}$$

and "O_P" is the probability big O as $p \to \infty$. Also, Z_1 and Z_2, independent of X_i, are two independent standard normal random variables.

Some comments follow. Note that in (3.5), the first term is of $O_P(1)$ and the second term $O_P(\frac{1}{\sqrt{p}})$. If one uses only the first term to construct the interval,

$$|\theta_i - \hat{M}_i X_i| \leq z A_{1,i}. \tag{3.7}$$

where z is such that $P(|Z_1| < z) = 1 - \alpha$, the coverage probability of (3.7), conditioning on X_i, would be of order $1 - \alpha + O_P(\frac{1}{\sqrt{p}})$, only first order correct.

To do better, one needs to use both leading terms of (3.5) to construct the interval

$$|\theta_i - \hat{M}_i X_i|^2 \leq z^2 (A_{1,i}^2 + A_{2,i}^2), \tag{3.8}$$

where the right hand side of (3.8) after omitting z is the variance of the sum of two leading terms (conditioning on X_i). *The coverage probability of (3.8), conditioning on X_i, equals $1 - \alpha + O_P(\frac{\ln p}{p})$, as $p \to \infty$.* The appendix provides the proof based on the Berry-Esseen Theorem.

In reality though, $A_{1,i}$ and $A_{2,i}$ involve unknown τ^2 which need to be estimated. An obvious choice is to estimate τ^2 by $\hat{\tau}^2$. Replacing τ^2 by $\hat{\tau}^2$, the interval (3.7) has poor coverage probabilities at times as was well-known in the work of Morris [1983] and Casella and Hwang [1983], the latter albeit being for the $p-$dimensional confidence set problem for $\vec{\theta} = (\theta_1, \cdots, \theta_p)$. Indeed, both papers come up with ingenious ways to enlarge the radii to improve on the coverage probabilities. Morris introduced an extra prior on τ^2. Later, for constructing confidence intervals, He [1992] took a decision theory approach similar to that of Casella and Hwang [1983] in order to overcome the low probability problem.

Here, (3.8) is designed for the case that θ_i has zero prior mean μ and below the Morris' interval analytically compared with is the one corresponding to such a case also. Morris [1983] uses a weighted average approach in estimating τ^2 by $\hat{\tau}^2_W$ which is equivalent to

$$\sum \frac{X_i^2}{\sigma_i^2 + \tau^2} = p, \tag{3.9}$$

which obviously shows that there is a unique solution for τ^2 when τ^2 is restricted to the region that $\sigma_i^2 + \tau^2 > 0$ for every i, or equivalently,

$$\tau^2 > -\min_i \sigma_i^2. \tag{3.10}$$

The above restriction does not affect the statistical procedure at all since when the solution is negative, the estimator $\hat{\tau}_W$ is taken to be zero. We state the theorem.

Theorem 3.2. *Assume that* $\tau^2 > 0$ *and* $\max_i \sigma_i^2 < B$, *where* B *is finite. Then conditioning on* X_i, *the distribution of* $\theta_i - \hat{M}^W X_i$ *has the same distribution as*

$$A_{1,i} Z_1 + A^W_{2,i} Z_2 + O_p(\frac{\ln p}{p}), \tag{3.11}$$

where $A_{1,i} = \sqrt{M_i}\sigma_i$ *and*

$$A^W_{2i} = \sqrt{\frac{2}{p}} \sigma_i^2 W_i^2 X_i / \bar{W}, W_i = (\tau^2 + \sigma_i^2)^{-1}, \bar{W} = \frac{\sum W_i}{p}. \tag{3.12}$$

Also, Z_1, Z_2 *are standard normal random variables independent of* X_i.

Hence the interval based on the second order formula (3.11) is

$$|\theta_i - \hat{M}_i^W X_i|^2 \leq z^2 Var, \tag{3.13}$$

where $Var = A_{1,i}^2 + (A_{2,i}^W)^2$. In comparison, Morris [1983] interval uses

$$Var^M = A_{1,i}^2 + C_i^2, \tag{3.14}$$

and replace the τ^2 by $\hat{\tau}_W^2$. In the equation above, the first term agrees with Var, but the second term C_i does not. We give C_i in the appendix and show that

$$(\frac{C_i}{A_{2i}})^2 = \frac{\bar{W}}{W_i}. \tag{3.15}$$

This implies that Morris' variance is second order correct if and only if (3.15) equals to 1, which is equivalent to the case when σ_i^2's are identical. For the unequal σ_i^2 case, when σ_i is small and hence W_i is large, the length of Morris' interval is short compared to the second order formula. Figure 1 shows that "Morris interval, known length", i.e., the interval described in (3.13) with Var replaced by Var^M of (3.14) and (3.15) has poor coverage probabilities when τ^2 is small. The phenomenon occurs more often for θ_i such that σ_i is the smallest. However, Figure 1 shows that σ_i does not have to be the smallest- it focuses on the coordinate with the 10 percentile of σ_i's. The second-order corrected version of Morris' interval (3.13) has good coverage probabilities for all τ.

Figure 1 also graphs the coverage probabilities of the Morris' interval, denoted as "Morris" in the box. Here, the interval is exactly the inter-

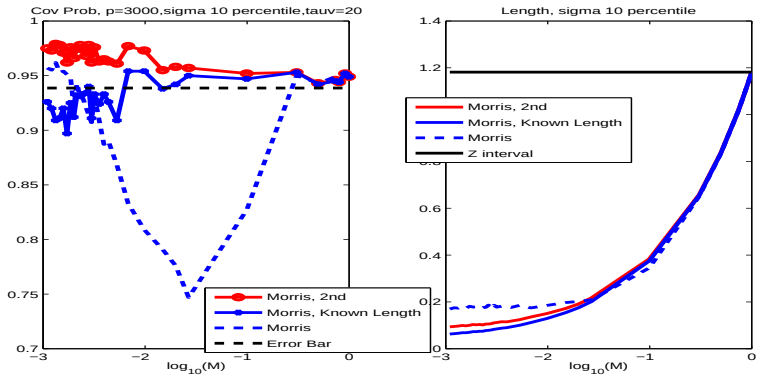

Figure 1: "Morris, Known Length" corresponds to (3.13) with Var replaced by Var^M defined in (3.14) and (3.15). The corresponding solid curve shows low coverage probabilities when τ^2 is small. The corresponding second order correct interval (3.13), denoted as "Morris, 2nd", boosts the coverage probabilities to be above the nominal level $1-\alpha = 0.95$. Morris [1983] shrinking towards a common mean has poor coverage probabilities as plotted by the dashed blue curve.

val shrinking toward an estimated common mean in Morris [1983], namely, his interval (4.3) with z_i in (1.10) taken to be the vector of ones. See Everson and Morris [2000] for a more explicit formula. The low coverage problem is now more serious when the τ^2 in the length is estimated, as in Morris [1983], by solving equation (3.10) with p replaced by $p-1$. Since the result seems so surprising, the computer output is compared in agreement with another step-by-step calculation using the matlab command mode.

To relate the above description to the ordered parameters such as $\theta_{(p)}$, note that, for any selection rule based on $\vec{X}$.

$$\theta_{(p)}|\vec{X} \sim N(M_{(p)}X_{(p)}, M_{(p)}\sigma^2_{(p)}). \tag{3.16}$$

Note here $M_{(p)}, X_{(p)}$, and $\sigma_{(p)}$ are defined to be M_j, X_j, and σ_j if the selection

rule after observing $\vec{X}$ chooses the j-th population, similar to the definition of $\theta_{(p)} = \theta_j$. With these replacement, (3.16) follows from (3.1). Because of this property, one expects the interval derived in (3.7) should have good *empirical* Bayes coverage probability for $\theta_{(p)}$ if all the i's in $\hat{M}_i, X_i, A_{1,i}$, and $A_{2,i}$ are replaced by (p) (or (i) for estimating $\theta_{(i)}$). Below when an interval for $\theta_{(i)}$ is constructed from an interval for θ_i, one makes the similar substitution by adding parentheses, called *parentheses adding adjustment*.

Figure 1 shows that the Morris' interval with the parentheses adding adjustment does not always have coverage probabilities above the nominal level. The blue curve plots the coverage probabilities of Morris adjusted interval and it shows that in many cases they are below the error bar, the 5% lower percentile of $\hat{p}$, the unbiased estimator of p for a binomial experiment with $p = 0.95$. In contrast, the second order correct interval has good coverage probabilities all above the error bar except in one case. See the red curve. The failure in one case does not prove that any of the coverage probabilities are less than 0.95. Even if all the true coverage probabilities are above the green line, the percentage of failure is 5%, which is larger than one out 31 cases. The second order formula proves to be very useful in constructing confidence intervals for $\theta_{(p)}$. It also explains why the second order correct interval has good coverage probabilities whereas Morris' interval does not. The length of the interval using the second order correctness formula is slightly longer but virtually identical to that of Morris' (See the second panel of Figure 1).

4 The unknown and unequal but estimable variances

For the case of unknown and unequal but estimable variances σ_i^2, it is necessary to consider a prior distribution for σ_i^2, so that there is a hope to construct a valid interval for parameters selected based on S_i's (and X_i's).

A conventional Bayes model assumes that

$$K_i \equiv \frac{S_i^2}{\sigma_i^2} \sim \frac{\chi_{df}^2}{df}, \text{ and } \sigma_i^2 \text{ has an inverse gamma distribution.} \tag{4.1}$$

Hwang and Liu [2010] models instead, $\log K_i \sim N(m, \sigma_K^2)$, where m and σ_k^2 denote the mean and the variance of $\log(\chi_{df}^2/df)$, two constants depending solely on the degrees of freedom df. They also assume that $\log \sigma_i^2$ has a $N(\mu_v, \tau_v^2)$ distribution. Putting all these together leads to

$$\log S_i^2 - m \sim N(\log \sigma_i^2, \sigma_K^2), \log \sigma_i^2 \sim N(\mu_v, \tau_v^2), \tag{4.2}$$

which were used to derive the powerful and practically very useful test, F_S test, of Cui et al. [2005]. The model, basically a normal-normal Bayesian model, leads to very simple estimate of the hyper-parameters μ_v and τ_v^2. It was found in both papers that the shrinkage test derived is quite robust with respect to the misspecification of the likelihood function and prior distribution, since the derived F_S test works well in simulating from (4.1) or some real data. Simulations involving with the real data are very likely different from (4.2).

Now redefine $\hat{M}_i = \frac{\hat{\tau}^2}{S_i^2+\hat{\tau}^2}$, which implies

$$\theta_i - \hat{M}_i X_i | (X_i, \sigma_i, S_i^2) \sim \sqrt{M_i}\sigma_i Z_1 + (M_i X_i - \hat{M}_i X_i). \tag{4.3}$$

Here Z_1 is a standard normal random variable independent of X_i, S_i, and σ_i. Also

$$M_i X_i - \hat{M}_i X_i = \frac{(\tau^2 - \hat{\tau}^2)\sigma_i^2 X_i}{(\sigma_i^2 + \tau^2)(S_i^2 + \hat{\tau}^2)} + \frac{(S_i^2 - \sigma_i^2)\tau^2 X_i}{(\sigma_i^2 + \tau^2)(S_i^2 + \hat{\tau}^2)}. \tag{4.4}$$

The first term of the right hand side is similar to (3.4) and, therefore, could be similarly dealt with. Numerical calculation of coverage probabilities shows that one could simply use a term, $\hat{A}_{2,i}$ similar to the second order term $A_{2,i}$ in (3.6) except that σ_i^2 is estimated by S_i^2. By using the conditional distribution of σ_i^2 given S_i^2, the second term on the right hand side of (4.4) can be shown to behave like

$$A_{3,i}(Z_3) = \frac{(1 - \exp(Z_3\sqrt{M_v \sigma_K^2}))\tau^2 X_i}{(1 + \frac{\tau^2}{\sigma_i^2})(S_i^2 + \hat{\tau}^2)}, \tag{4.5}$$

where $M_v = \frac{\tau_v^2}{\sigma_K^2 + \tau_v^2}$, and Z_3 is a standard normal random variable independent of Z_1. Putting all these together, leads to the approximated confidence interval

$$|\theta_i - \hat{M}_i X_i| \leq V_i, \tag{4.6}$$

where V_i is the cutoff point such that

$$P(|\sqrt{\hat{A}_{1,i}^2 + \hat{A}_{2,i}^2} Z_2 + \hat{A}_{3,i}(Z_3)| \leq V_i) = 1 - \alpha, \tag{4.7}$$

where $\hat{A}_{1,i}, \hat{A}_{2,i}$, and $\hat{A}_{3,i}$ are defined in (3.6) and (4.5) with σ_i^2 replaced by S_i^2, and Z_2 and Z_3 are independent standard normal random variables. Here V_i can be obtained by simulating Z_2 and Z_3. However, the computation could be very time consuming especially if one needs, for example, 100 intervals to calculate in repeat simulations. Instead, a computationally simpler intervals can be constructed based on the following inequality.

Theorem 4.1. *Suppose z_2 be such that $P(|Z| < z_2, |Z_3| < z_2) = 1 - \alpha$, (Hence z_2 is such that $P(|Z| < z_2) = \sqrt{1-\alpha}$.) Let $D = aZ + (1 - \exp(cZ_3))b$, where a, b, and c are all positive numbers. Then*

$$P(D \in (L, U)) \geq 1 - \alpha, \quad \textit{where} \tag{4.8}$$

$$L = -az_2 + (1 - \exp(cz_2))b, \quad \textit{and} \quad U = az_2 + (1 - \exp(-cz_2))b. \tag{4.9}$$

This theorem recommends the confidence interval (L_i, U_i), defined as in (4.9) with $a = \sqrt{\hat{A}_{1,i}^2 + \hat{A}_{2,i}^2}$, $b = \frac{\tau^2 X_i}{(1+\frac{\tau^2}{S_i^2})(S_i^2+\hat{\tau}^2)}$, and $c = \sqrt{M_v \sigma_K^2}$. Although the theorem requires using z_2, we used z_1 instead of z_2 where $P(|Z| < z_1) = 1 - \alpha$ in all our numerical calculations below. The corresponding interval is narrower but it still has good coverage probabilities.

At this point, τ^2 appeared in a and b is still not estimated. It is somewhat tricky in estimating τ^2. There is yet another step needed. In all the constructions of confidence intervals, centered at shrinkage estimators, there is an issue of truncation on $\hat{\tau}^2$ defined in (3.2) with σ_i^2 replaced by S_i^2. The coverage probabilities are usually lower than the nominal level when τ^2 is small. Most likely this is due to the fact that $\hat{\tau}^2$ can be zero which

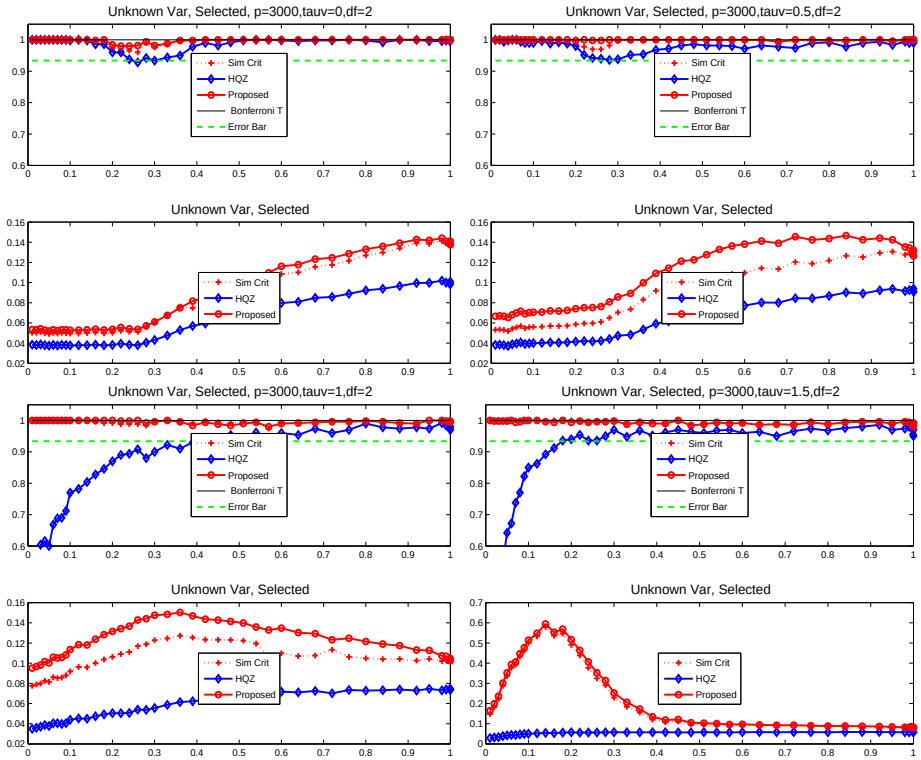

Figure 2: (K=1). The coverage probabilities for $\theta_{(p)}$, $p = 3,000$ of four intervals are simulated and plotted in the graphs in the first and the third rows. The average ratios of the lengths of the same intervals to the lengths of the Bonferroni t-intervals are shown in the graphs of the second and the fourth rows. Note that in many cases, the average ratios are less than 16% for the three alternative intervals: "sim crit" as in (4.7), "proposed" as in (4.10), and HQZ. All these graphs are plotted against M. Unlike HQZ, the proposed interval and the "sim crit" interval are shown to have the *empirical* Bayes probabilities no less than $1 - \alpha = 95\%$. The degrees of freedom are two. The σ_i^2 are generated from a log normal distribution with the variance τ_v^2 ranging from 0, 0.5, 1, and 1.5.

reduces (L_i, U_i) to a zero length interval. Hence one needs to replace $\hat{\tau}^2$ by $\max(\hat{\tau}^2, \tau_0^2)$, where τ_0^2 is a positive number. He [1992], Qiu and Hwang [2007] and Hwang, Qiu, and Zhao [2009] all employed some kind of truncation. In particular, the last two papers propose a systematic way to derive the truncation point τ_0^2. The same choice of τ_0^2 in (6.3) of Hwang et al. [2009] is used here, except $\hat{\sigma}^2_{EB,i}$ is replaced by S_i^2.

Let $\hat{\tau}_* = \max(\tau_0, \hat{\tau})$. The interval proposed is

$$(\inf_{0<\tau<\hat{\tau}_*} L_i, \sup_{0<\tau<\hat{\tau}_*} U_i). \tag{4.10}$$

When $\hat{\tau}$ is large, the supremum(sup) and infimum(inf) occur at $\hat{\tau}$. However for small $\hat{\tau}$, L and U are not monotonic in τ which motivates the taking of inf and sup. At this point, this is done numerically by taking the minimum and maximum over 60 evenly spaced points starting at $\tau = 0$ and ending at $\hat{\tau}_*$. Although one may propose a computationally simpler method, the proposed method is not demanding in computation.

Finally, τ_v, M_v, and μ_v need to be estimated, but this is a much easier problem. It can be resolved by replacing them with intuitive estimators without truncation as in Hwang, Qiu, and Zhao [2009]. With the substitution, (4.10) is the proposed interval.

After generating the random observations, the top parameter (out of $p = 3,000$) corresponding to the largest t-values in magnitude are selected. We then constructed the simultaneous confidence intervals for this selected parameter and calculated the Bayes coverage probability based on the proposed approach (4.10), the Bonferroni t-interval, and the HQZ intervals,

proposed by Hwang et al. [2009]. Also calculated are the average ratios of the lengths of various intervals to the lengths of the Bonferroni t-intervals. See Figure 2. The Morris intervals are not included and are not expected to have good coverage probabilities since they do not shrink the variances.

The coverage probability of the proposed interval (4.10) is above a nominal level $1-\alpha$, taken to be 95%. Also the average length is shorter than that of the t-interval using the Bonferroni correction, called the Bonferroni t-interval, but is longer than that of HQZ by Hwang, Qiu, and Zhao [2009] which have poor *empirical* Bayes coverage probabilities at times. See Figure 2.

It is worthy to mention that second order term derived from $A_{2,i}$ in (3.6) remains important. Dropping this term leads to an interval that has coverage probabilities below 95%. Having the second order term is one of the main reasons why the proposed interval works well.

5 Simultaneous Intervals

All the discussions thus far focus on the interval construction for only the top observation. In practice, one often selects several significance parameters,

$$\theta_{(p)}, \theta_{(p-1)}, \cdots, \theta_{(p-R+1)}. \tag{5.1}$$

Consequently, one would like to have simultaneous confidence intervals for these parameters that satisfy criteria such as having large simultaneous coverage probabilities.

To do so, one could construct the proposed interval for each of (5.1) at level $1-\frac{\alpha}{R}$. Since each interval is expected to have *empirical* Bayes coverage probabilities above $1-\frac{\alpha}{R}$, Bonferroni inequality would imply the simultaneous coverage probabilities to be above $1-\alpha$.

Note that the proposed simultaneous intervals are different from the Bonferroni t-intervals because the coverage probability for a single selected parameter is adjusted to be $1-\alpha/R$ rather than $1-\alpha/p$ as in the Bonferroni t-intervals. Note that R could be 100 and is much less than p, typically ten thousands or more in a microarray experiment. Hence the proposed intervals are much shorter.

HQZ intervals can also be adapted for constructing simultaneous intervals for (5.1) based on the same adjustment discussed above. These *empirical* Bayes intervals shrink both the means and variances, having good coverage probabilities for the unselected parameters. However, for (5.1), HQZ intervals have poor coverage probabilities or large *empirical* Bayes FCR. See Figures 3 and 4.

Similar to the simulation in Section 4, we simulated σ_i^2 from either the Log-Normal model or the Inverse-Gamma model. The results lead to similar conclusions. Hence only the graphs based on the Inverse-Gamma model are reported.

Unlike the study in the previous section which focuses on one-selected population, the top 50 parameters (out of $p = 3,000$) corresponding to the largest t-values in magnitude are selected. We then constructed the simultaneous confidence intervals for these 50 selected parameters and calculated the simultaneous Bayes coverage probabilities based on the proposed ap-

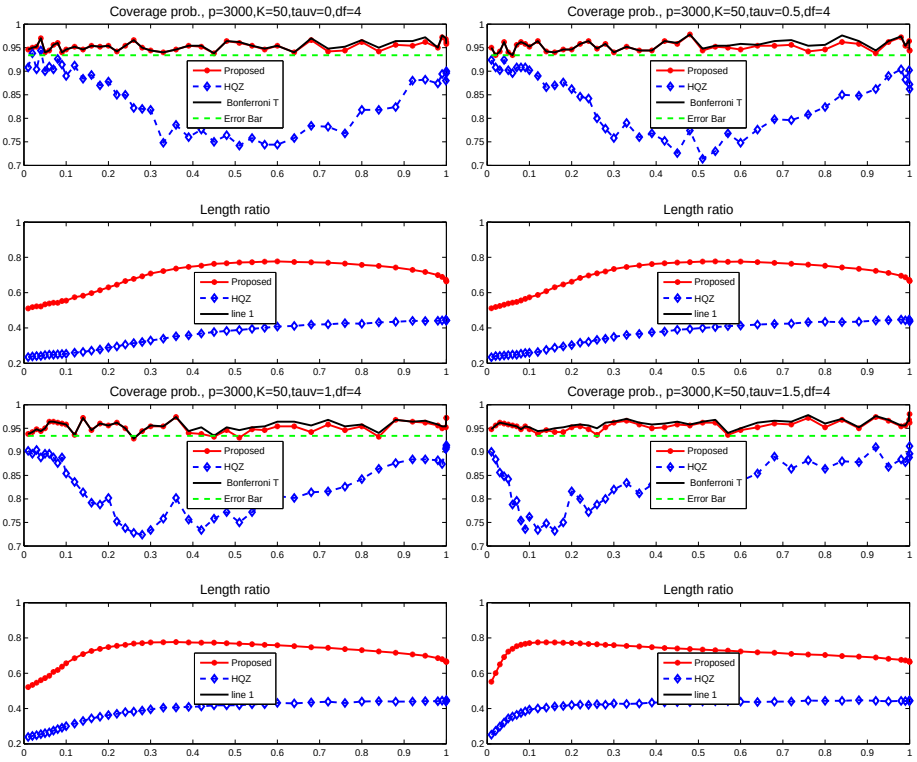

Figure 3: (R=50). Simultaneous coverage probabilities for 50 selected parameters are plotted in the graphs in the first and the third rows. Also, the average ratios of the lengths of various intervals to those of Bonferroni t-interval are plotted in the graphs in the second and the fourth rows. The proposed simultaneous intervals (4.10) perform the best, having *empirical* Bayes coverage probabilities no less than $1-\alpha$ and a smaller average length compared to that of the Bonferroni t-intervals. Simulation settings are similar to Figure 2 except the degrees of freedom are four.

proach, the Bonferroni t-interval, and the HQZ intervals. Also calculated are the average ratios of the lengths of various intervals to the lengths of the Bonferroni t-intervals. See Figure 3.

In Figure 3, both the proposed intervals and the Bonferroni t-intervals have the simultaneous coverage probabilities above the nominal level, 95%. However, HQZ intervals fail drastically, even though they enjoy great length reduction when compared with the Bonferroni t-intervals. The proposed intervals have the coverage probabilities above nominal level and their average ratios of the lengths to those of the Bonferroni intervals are always less than 80%. In some cases, the average ratios can be as small as 60%. This simulation studies indicate that both the Bonferroni t-intervals and the HQZ intervals are either too conservative or too liberal and the proposed intervals attain a good balance of having short lengths and good *empirical* Bayes coverage probabilities.

Now some discussions of Bayes FCR are in order. Under the same simulation setting, Figure 4 graphs the Bayes FCR and the average ratios of the lengths of the HQZ intervals, and the proposed intervals to those of the BY intervals. It is clearly seen that the proposed intervals control the *empirical* Bayes FCR at a satisfactory level. While BY intervals also control the *empirical* Bayes FCR, the proposed intervals have, on average, only 40% to 80% of the lengths of the B-Y intervals.

Note that the proposed intervals used to calculate the FCR are individual $1-\alpha$ intervals and are different from and shorter than the proposed $1-\alpha$ *empirical* simultaneous intervals described in paragraph 2 of this section. Here each interval has the $1-\alpha$ and not $1-\alpha/R$ *empirical* Bayes coverage

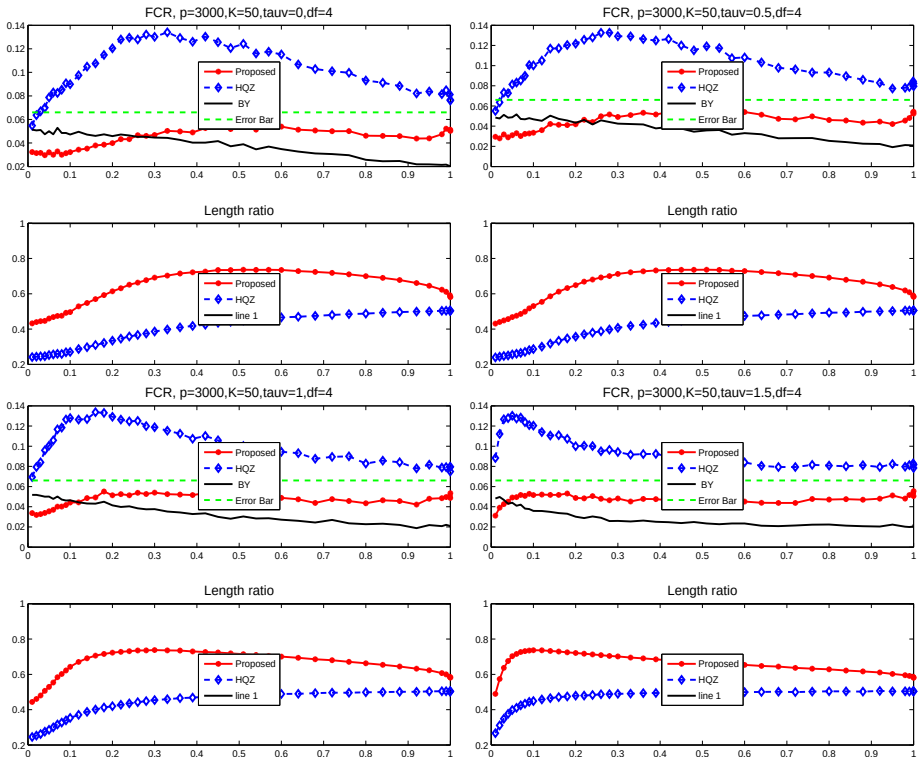

Figure 4: (R=50). Graphs and simulation settings are similar to Figure 3, except here the Bayes FCR are plotted in the graphs in the first and the third rows. Also, graphs in the second and the fourth rows show the average ratios of the lengths of various intervals relative to BY intervals. The graphs show that the proposed intervals, like BY intervals, have *empirical* Bayes FCR controlled while being only 40% to 75% as long as the BY intervals.

probabilities. In contrast, the approach of Benjamini and Yekutieli [2005] uses the usual t-interval each at $1 - \frac{\alpha R}{p}$ level. Hence, our intervals can be proved analytically to be shorter than B-Y interval.

6 Data Analysis

We apply various intervals to an Affymetrix Control data set, the golden spike-in data set of Choe et al. [2005]. All the parameters in this data set is pre-chosen and known. Therefore, it can be used to assess different statistical procedures. In Hwang et al. [2009] and Zhao and Hwang [2010], they have compared the performance of various confidence intervals construction. In Zhao [2010], he used this dataset to calculate the risk of various point estimators. The conclusion reported below is consistent with the simulation results reported in Figures 2-4.

We download the data from `http://www.elwood9.net/spike`. There are 13,991 genes and there are 6 replicates for each gene, three from each of the control and treatment group. Define

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

The degrees of freedom are calculated according to Satterthwaite approximation. As similarly argued in Zhao and Hwang [2010], we take the bias in consideration when calculating the coverage probabilities and FCRs. Otherwise, the coverage probabilities of all approaches are extremely low. In the following studies, we have calculated the selected confidence intervals based

Table 1: Coverage probabilities and lengths comparison of various intervals based on the observation with the largest t value in magnitude out of 3,000 genes. We are aiming at controlling the coverage probability at 5% level.

	Proposed	HQZ	Bonferroni T
Cov	93.4%	84.6%	1
Leng	0.334	0.204	4.34

on various selection rules.

Note that the proposed intervals used to calculate the FCR are different from and shorter than the proposed $1-\alpha$ *empirical* Bayes simultaneous intervals described in paragraph 2 of this section. Here each interval has the $1-\alpha$ and not $1-\frac{\alpha}{R}$ *empirical* Bayes coverage probability. In constrast, the approach of Benjamini and Yekutieli [2005] uses the usual t-intervals, each at $1-\frac{\alpha R}{p}$ probability level.

Case 1 (Table 1 - one selected parameter.) For each simulation, we randomly sample 3,000 observations among all genes with replacement and then select the top parameter with the largest T-value in magnitude. We then construct the various confidence intervals for the selected parameter $\theta_{(p)}$. We replicate this simulation study 1,000 times and report the relative frequency of coverages, an estimator of coverage probability. We have also reported the average half length comparison as shown in Table 1.

In this study, we are aiming at controlling the Bayes coverage probability at the 5% level. The proposed approach has the coverage probabilities 93.4%, close to the nominal level. As expected, the Bonferroni intervals are so extremely conservative that the coverage probability is one. The HQZ intervals have short average length, shroter than the proposed intervals, but their coverage probability is 84.6%, significantly smaller than the nominal

Table 2: FCR calculation based on the golden spike-in dataset. Out of 3,000 genes, K observations are selected for the simultaneous confidence interval construction where $K = 50$ and 100. The FCR and average length of various approaches are reported in the table.

	K=50	HQZ	Proposed	B-Y	K=100	HQZ	Proposed	B-Y
FCR		0.0864	0.0340	0.0115		0.1052	0.0363	0.0143
Length		0.3301	0.7759	1.5443		0.3579	0.8392	1.2962

level 95%. The proposed interval has good coverage probability 93.4% and short average length (0.334), less than 8% of the average length of the Bonferroni t-intervals.

Case 2 (Table 2 - selected fixed R parameters. Similarly as in Case 1, we randomly sample 3,000 observations among all genes with replacement. We then select the top R parameter with the largest T-value in magnitude. We construct the confidence intervals using various approaches for these selected parameters

$$\theta_{(p)}, \theta_{(p-1)}, \cdots, \theta_{(p-R+1)}$$

and calculate the total number of non-coverage V and the false coverage ratio as $\frac{V}{R}$. We replicate this simulation study 1,000 times and take the average of the false coverage ratio as an estimator of FCR. We have also reported the average half length comparison as shown in Table 2.

The proposed approach and B-Y's method controls the FCR at the nominal level 5% while the HQZ approach fails even though it has the shortest average length. The length of our approach is about 50% of that of B-Y's approach when $K = 50$. Therefore, the proposed approach attains a good balance between the precision and accuracy of the confidence interval.

Table 3: FCR and average length comparisons based on the Golden Spike-in dataset when the parameters are selected by using the BH's FDR controlling procedure with the FDR level 5%.

	HQZ	Proposed	B-Y
FCR	0.1148	0.0250	0.0171
length	0.3842	0.8461	0.9557

Case 3 (Table 3 - random R). In Case 2, the number of parameters that are selected for interval construction are pre-chosen as 50 or 100. In some application, scientists might use the data-driven selection procedure such as the FDR controlling approach by Benjamini and Hochberg [1995], an approach used in Table 3. In this study, we first randomly draw 3,000 genes out of 13,997 as in Case 1. BH's procedure with 5% FDR level is applied to select the parameters of interest. We thus construct the simultaneous intervals for these selected intervals and report the corresponding FCR and average length as in Table 3. It is clearly seen that the proposed approach enjoys the balance of having controlled FCR at the nominal level 5% and small average length.

7 Discussion

In this paper, having considered the construction confidence interval and simultaneous intervals after selecting the parameters by using statistical procedures. The post inference after the selection is very difficult under the frequentist framework. The Bayes statistics is immune to the selection issue but requires the exact prior information. We therefore employ the *empirical* Bayes criterion.

The *empirical* Bayes inference is especially appropriate for the modern large p problem. The existence of a large number of parameters motivates a statistician to model it by using some prior distribution characterized by a few hyper-parameters. Some statisticians might use another hierarchical prior for these hyperparameters. However, by not introducing another stage of prior, our approach is simpler and is more readily applicable to the unknown estimable variances.

The confidence intervals constructed in this article has demonstrated good properties. However, there are still much to do to enhance the procedure. For instance, the dependence structure of X_i's needs to be addressed.

8 Acknowledgements

The authors thank Song Chang at Cornell who at the earlier stage of this research, helped us with simulations. Also we thank Professors Carl N. Morris and Philip Everson for explaining their work to the authors.

A Remarks and Technical Proofs

Remark on nonzero prior mean of θ_i.

Assume that θ_i are i.i.d. $N(\mu, \tau^2)$, where unlike (2.1), μ is not necessarily zero. When μ is known, one can obvious replace X_i by $X_i' = X_i - \mu$. Letting $\theta_i' = \theta_i - \mu$ leads to $X_i' \sim N(\theta_i, \sigma_i^2)$ and $\theta_i' \sim N(0, \tau^2)$ where $\theta_i' = \theta_i - \mu$, which satisfies model (2.1). Hence the interval (4.10) which has good coverage probabilities can be applied and have the same coverage probabilities. Hence

we replace all X_i by X_i' and we could come up with an interval for θ_i', i.e.,

$$L' \leq \theta_i' - M_i' X_i' \leq U', \tag{A.1}$$

where L', M_i' and U', are respectively L, M_i and U expect that X_i is replaced by X_i'. The interval then leads to

$$L' \leq \theta_i - (M_i'(X_i - \mu) + \mu) \leq U'. \tag{A.2}$$

For the selected parameter $\theta_{(i)}$, we only need to apply the parentheses adding adjustment to (A.2). The resultant interval would obviously have good coverage probabilities even for the selected parameter as demonstrated in Figure 5.

When μ is unknown, we replace μ by $\bar{X} = (\sum X_i)/p$ and calculate numerically the coverage probabilities. This produces a graph similar to Figure 5. For the R selection mean case, we anticipate a similar modification works. □

Proof of Theorem 3.1:

Since $\theta_i | X_i \sim N(M_i X_i, M_i \sigma_i^2)$, then

$$\begin{aligned} & \theta_i - \hat{M}_i X_i | X_i = (M_i - \hat{M}_i X_i) + \sqrt{M_i}\sigma_i Z_1 \\ = \quad & \frac{(\tau^2 - \hat{\tau}^2)\sigma_i^2}{(\sigma_i^2 + \tau^2)(\sigma_i^2 + \hat{\tau}^2)} X_i + \sqrt{M_i}\sigma_i^2 Z_1. \end{aligned}$$

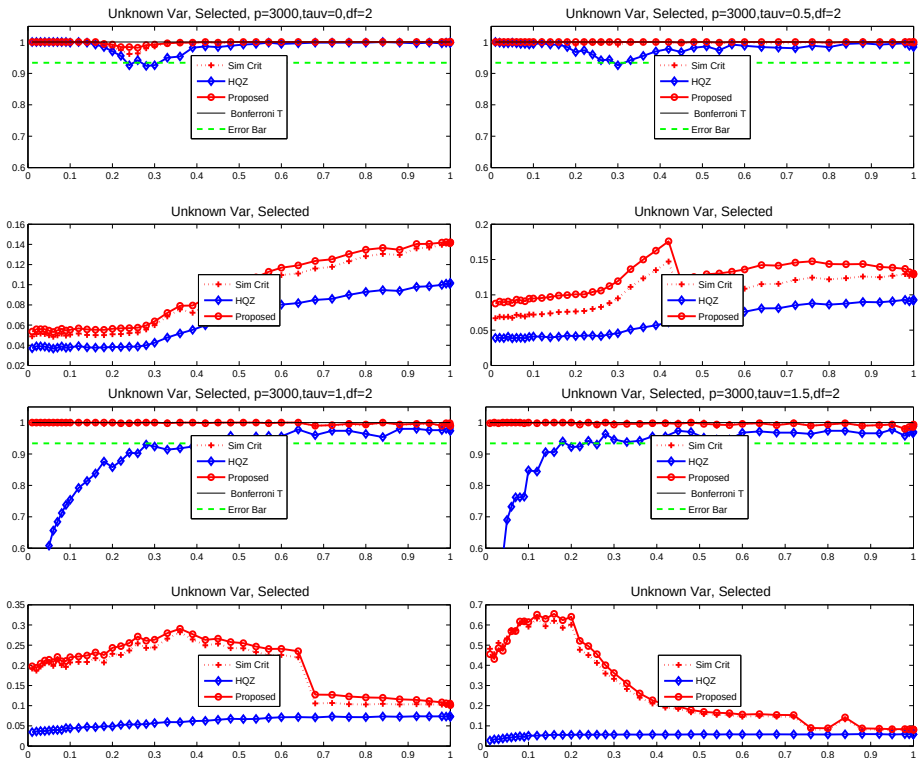

Figure 5: (K=1). In this study, we assume that the hyper mean μ of the parameter of θ_i's is equal to 1. The rest of the simulation setting are the same as that of Figure 2. The coverage probabilities for $\theta_{(p)}$, $p = 3,000$ of four intervals are simulated and plotted in the graphs in the first and the third rows. The average ratios of the lengths of the same intervals to the lengths of the Bonferroni t-intervals are shown in the graphs of the second and the fourth rows. The conclusion we previous obtained still holds for the estimated mean intervals (A.2).

According to the definition of $\hat{\tau}^2$, one knows that

$$\hat{\tau}^2 - \tau^2 | X_i = \frac{\sum_{j \neq i}(X_j^2 - \sigma_j^2 - \tau^2)}{p} + \frac{X_i^2 - \sigma_i^2 - \tau^2}{p}.$$

Direct calculation shows that

$$Var(\frac{\sum_{j \neq i}(X_j^2 - \sigma_j^2 - \tau^2)}{p}) = \frac{2\sum_{j \neq i}(\sigma_j^2 + \tau^2)^2}{p^2}.$$

Let $Z_2^* = \frac{\frac{1}{p}\sum_{j \neq i}(X_j^2 - \sigma_j^2 - \tau^2)}{\sqrt{\frac{2\sum_{j \neq i}(\sigma_j^2 + \tau^2)^2}{p^2}}}$. Then

$$\theta_i - \hat{M}_i X_i | X_i = \frac{\sigma_i^2}{(\sigma_i^2 + \tau^2)^2} X_i \sqrt{\frac{2\sum_j (\sigma_j^2 + \tau^2)^2}{p^2}} Z_2^* + \sqrt{M_i}\sigma_i Z_1 + O_p(1/p). \tag{A.3}$$

According to Berry-Esseen's theorem, one knows that $F_{Z_2^*}(x) = \Phi(x) + O(1/\sqrt{p})$. Take $U \sim U(0,1)$. Note that

$$Z_2^* \stackrel{d}{=} F_{Z_2^*}^{-1}(U), \tag{A.4}$$

Since $F_{Z_2^*}(x) = \Phi(x) + O(\frac{1}{\sqrt{p}})$. For $(\ln p)^{-1+\delta} \leq u \leq 1 - (\ln p)^{-1+\delta}$, consider $F_{Z_2^*}^{-1}(u)$. Since $F_{Z_2^*}$ is a cumulative distribution function of a continuous random variable, it is strictly increasing. Consequently, there exists an unique x, such that $u = F_{Z_2^*}(x)$.

$$\begin{aligned} & F_{Z_2^*}^{-1}(u) - \Phi^{-1}(u) \\ = \; & x - \Phi^{-1}(F_{Z_2^*}(u)) = \Phi^{-1}(\Phi(x)) - \Phi^{-1}(F_{Z_2^*}(x)) \\ = \; & |\Phi(x) - F_{Z_2^*}(x)| |\frac{1}{\phi(\Phi^{-1}(\xi))}|, \end{aligned}$$

where ξ is a value between $\Phi(x)$ and u. Consequently, $|\xi - u| = O(\frac{1}{\sqrt{p}})$. Without loss of generality, assume that $\Phi^{-1}(\xi)$ is positive. According to Mill's ratio, for any $x \geq 0$,

$$\frac{\phi(x)}{1 - \Phi(x)} \geq \frac{1}{\sqrt{2\pi}}.$$

This inequality holds if we replace x by $\Phi^{-1}(\xi)$ and becomes

$$\frac{\phi(\Phi^{-1}(\xi))}{1 - \phi(\Phi^{-1}(\xi))} = \frac{\phi(\Phi^{-1}(\xi))}{1 - \xi} \geq \frac{1}{\sqrt{2\pi}}.$$

As a result,

$$\frac{1}{\phi(\Phi^{-1}(\xi))} \leq \frac{\sqrt{2\pi}}{1 - \xi} = O(\ln p^{1-\delta}).$$

In summary, one knows that $F^{-1}_{Z_2^*}(u) - \Phi^{-1}(u) = O(\frac{\ln p^{1-\delta}}{\sqrt{p}})$ for $(\ln p)^{-1+\delta} \leq u \leq 1 - (\ln p)^{-1+\delta}$.

Next consider the random variable $F^{-1}_{Z_2^*}(U) - \Phi^{-1}(U)$. Assume that $(\ln p)^{-1+\delta} \leq u \leq 1 - (\ln p)^{-1+\delta}$. Then for any $\epsilon > 0$, when p is sufficiently large,

$$\frac{\sqrt{p}}{\ln p} |F^{-1}_{Z_2^*}(u) - \Phi^{-1}(u)| = O(\ln p^{-\delta}) < \epsilon.$$

Therefore

$$\begin{aligned} & P(\frac{\sqrt{p}}{\ln p} |F^{-1}_{Z_2^*}(u) - \Phi^{-1}(u)| > \epsilon) \\ \leq \quad & P(U < (\ln p)^{-1+\delta}) + P(U > 1 - (\ln p)^{-1+\delta}) < 2(\ln p)^{-1+\delta} \to 0. \end{aligned}$$

This proves that

$$F^{-1}_{Z_2^*}(U) = \Phi^{-1}(U) + O_p(\frac{\ln p}{\sqrt{p}}) = Z_2 + O_p(\frac{\ln p}{\sqrt{p}}).$$

As a result,

$$\begin{aligned} & \theta_i - \hat{M}_i X_i | X_i \\ \stackrel{d}{=} \quad & \frac{\sigma_i^2}{(\sigma_i^2+\tau^2)^2} X_i \sqrt{\frac{2\sum_j(\sigma_j^2+\tau^2)}{p^2}}(Z_2 + O_p(\frac{\ln p}{\sqrt{p}})) + \sqrt{M_i}\sigma_i Z_1 + O_p(1/p) \\ \stackrel{d}{=} \quad & A_{1,i} Z_1 + A_{2,i} Z_2 + O_P(\ln p/p), \end{aligned}$$

which completes the proof. □

Proof of Theorem 3.2:

As in Theorem 4.1,

$$\theta_i - \hat{M}_i^W X_i | X_i = \sqrt{M_i}\sigma_i Z_1 + (M_i - \hat{M}_i^W) X_i, \tag{A.5}$$

and

$$(M_i - \hat{M}_i^W) X_i = \frac{(\tau^2 - \hat{\tau}_W^2)\sigma_i^2}{(\sigma_i^2 + \tau_W^2)(\sigma_i^2 + \hat{\tau}_W^2)} X_i. \tag{A.6}$$

The major difficulty to overcome is to calculate the asymptotic distribution of (A.6). The calculation is based on Lemma A.1, to be proved below. Using this lemma and (A.6), we arrive the asymptotic expression (3.11), establishing Theorem 3.2.

Lemma A.1.

$$\sqrt{p}(\hat{\tau}^2 - \hat{\tau}_W^2)|X_i \sim N(0, \frac{2}{(\bar{W})^2}), \tag{A.7}$$

where $\bar{W} = \frac{1}{p}\sum W_i$ *and* $W_i = (\sigma_i^2 + \tau^2)^{-1}$.

Proof: In the argument below, we would just prove the unconditional version of Lemma A.1. Conditioning on X_i merely introduces a term like X_i/p which can be ignored.

Below, we shall simplify the notation by using V_i and A to denote σ_i^2 and τ^2. Also, we write $\hat{\tau}_W^2 = \max(0, \hat{A})$ where $\tau^2 = \hat{A}$ is defined as the unique solution to (3.9). Note that $\hat{A}$ may be negative. Using the simplified notation, we shall now prove that

$$\sqrt{p}(A - \hat{A}) \sim N(0, 2/(\bar{W})^2), \tag{A.8}$$

where $\bar{W} = \sum W_i/p$ and $W_i = (V_i + A)^{-1}$. Also $\hat{W}_i$ denotes $(V_i + \hat{A})^{-1}$.

Equation (A.8) and the assumption that $A > 0$ imply that (A.7) holds if $\hat{A}$ is replaced by $\hat{\tau}_W^2$, because by (A.8),

$$P(\hat{A} = \hat{\tau}_W^2) = P(\hat{A} > 0) \to 1, \text{as} \quad p \to \infty. \tag{A.9}$$

To prove (A.8), let $D = 1 - \frac{1}{p}\sum W_i X_i^2$. Note that (3.9) implies that $\frac{1}{p}\sum_i \hat{W}_i X_i^2 = 1$ and

$$D = \frac{1}{p}\sum(\hat{W}_i X_i^2 - W_i X_i^2) = \frac{1}{p}(A - \hat{A})\sum X_i^2 \hat{W}_i W_i, \tag{A.10}$$

where we use the identity

$$\hat{W}_i - W_i = (A - \hat{A})(\hat{W}_i - W_i). \tag{A.11}$$

Now

$$|D| \geq |A - \hat{A}|(\frac{1}{p}\sum X_i^2 W_i)\frac{1}{(B+1)+\hat{A}}, \tag{A.12}$$

by employing the assumption that $V_i < B < B + 1$ for all i. Consider any subsequence of $\hat{A}$ as p changes with a limit A^*. If $A^* = \infty$, by the law of large numbers, the right hand side of (A.12) approaches to 1 where as the left side approaches zero, a contradiction. We next show that A^* can not be $-\infty$ either. For $\hat{A} < 0$, (A.12) implies that

$$|D| \geq |A - \hat{A}|(\frac{1}{p}\sum X_i^2 W_i)/(B+1). \tag{A.13}$$

The right hand side converges to ∞ if $\hat{A} \to A^* = -\infty$ and the left side to zero, leading to a contradiction.

Finally, we show $A^* = A$. We only need to consider the case where A^* is finite, hence (A.12) implies that

$$0 \geq |A - A^*|\frac{1}{(B+1)+A^*}, \tag{A.14}$$

where $(B+1)+A^* > (B+1) - \min V_i > 1$. The last inequality implies that $A^* = A$.

We have now proved that any subsequence of $\hat{A}$ with a limit must have

the limit A. Hence we conclude that

$$\hat{A} \to A \quad \text{almost surely.} \tag{A.15}$$

Consequently, $P(\hat{A} > A/2) \to 1$ and hence from now on, we may assume without loss of generality that $\hat{A} > \frac{A}{2} > 0$. From (A.10),

$$(A - \hat{A}) = \frac{D}{\frac{1}{p}\sum(X_i^2 \hat{W}_i W_i)}. \tag{A.16}$$

Now we claim that

$$\frac{\sum(X_i^2 \hat{W}_i W_i)}{\sum X_i^2 W_i^2} \to 1, a.s. \tag{A.17}$$

Note that

$$\frac{\sum X_i^2 W_i^2}{\sum X_i^2 \hat{W}_i W_i} - 1 = \frac{\sum X_i^2 W_i (W_i - \hat{W}_i)}{\sum X_i^2 \hat{W}_i W_i} = (\hat{A} - A)\frac{\sum X_i^2 W_i^2 \hat{W}_i}{\sum X_i^2 \hat{W}_i W_i},$$

where the last equation follows from (A.11). However, $\frac{\sum X_i^2 W_i^2 \hat{W}_i}{\sum X_i^2 \hat{W}_i W_i} \leq \max W_i = \frac{1}{A}$. Hence (A.15) implies (A.17),

Putting (A.17) and (A.16) together yields

$$A - \hat{A} \stackrel{d}{=} D(\frac{1}{p}\sum X_i^2 W_i^2)^{-1} = D(\frac{1}{p}\sum W_i Z_i^2)^{-1}. \tag{A.18}$$

where $Z_i = \sqrt{W_i} X_i \sim N(0,1)$. One can establish that

$$\frac{\sum W_i Z_i^2}{\sum W_i} \to 1, \tag{A.19}$$

by using the inequality

$$\frac{\sum W_i Z_i^2}{\sum W_i} - 1 = \frac{\frac{1}{p}\sum W_i(1 - Z_i^2)}{\frac{1}{p}\sum Z_i^2 W_i} = O_p(\frac{1}{\sqrt{p}}).$$

Now (A.17) and (A.19) imply

$$A - \hat{A} \stackrel{d}{=} D(\bar{W})^{-1} = (1 - (\sum Z_i^2)/p)/\bar{W}.$$

This obviously implies Lemma A.1.

Proof of (3.15). Morris' interval for the case that θ_i are i.i.d. $N(0, \tau^2)$ reduces to an interval using (3.14) as the variance estimator with

$$\begin{aligned} C_i^2 &= \frac{2}{p}(1 - M_i)^2 X_i^2 (\tau^2 + \frac{\sum W_i \sigma_i^2}{\sum W_i}) W_i \\ &= \frac{2}{p}(1 - M_i)^2 X_i^2 \frac{W_i}{\bar{W}}. \end{aligned}$$

Since $(A_{2i}^W)^2 = \frac{2}{p}\sigma_i^4 W_i^4 X_i^2/(\bar{W})^2$, (3.15) follows easily.

Proof of Theorem 4.1.

Note that $D \in (L, U)$ if and only if

$$-az_2 + (1 - \exp(cz_2)b) \leq aZ + (1 - \exp(cZ_3))b \leq az_2 + (1 - \exp(-cz_2))b.$$

The last inequality is obviously implied by the two inequalities:

$$-z_2 \leq Z \leq z_2 \quad \text{and} \quad -z_2 \leq Z_3 \leq z_2, \tag{A.20}$$

which simultaneously hold with probability $1 - \alpha$ by the assumption. Hence

the probability that $D \in (L, U)$ is at least $1 - \alpha$, establishing (4.8).

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, 100 (469):71–93, 2005. With comments and a rejoinder by the authors.

R. Berk, L. D. Brown, L. Zhao, A. Buja, and K. Zhang. Valid statistical inference after model selection. 2010a. preprint.

Richard Berk, Lawrence Brown, and Linda Zhao. Statistical inference after model selection. *Journal of Quantitative Criminology*, 26:217–236, 2010b.

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

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

A. Cohen and H. B. Sackrowitz. Estimating the mean of the selected population,. *In Statistical Decision Theory and Related Topics III (S. S. Gupta, J. O. Berger, eds.)*, 1:243–270, 1982.

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

Bradley Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008.

P. J. Everson and C. N. Morris. Inference for multivariate normal hierarchical models. *Journal Of The Royal Statistical Society Series B*, 62(2): 399–412, 2000.

Edward Giovannucci, Alberto Ascherio, Eric B. Rimm, Meir J. Stampfer, Graham A. Colditz, and Walter C. Willett. Intake of Carotenoids and Retino in Relation to Risk of Prostate Cancer. *Journal of the National Cancer Institute*, 87(23):1767–1776, 1995.

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

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

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

Zahra Montazeri, Corey M. Yanofsky, and David R. Bickel. Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by

estimation in high-dimensional biology: Applications to differential gene expression. *Statistical Applications in Genetics and Molecular Biology*, 9 (1):23, 2010.

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

C. N. Morris and R. Tang. Estimating random effects via adjustment for density maximization. 2010.

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

J. E. Rossouw, G. L. Anderson, R. L Prentice, and et al. Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. *JAMA*, 288(3):321–333, July 2002. ISSN 0098-7484.

J. H. Venter. Confidence bounds based on the largest treatment mean. *South African Journal of Science*, 84, 1988.

J. H. Venter and S. J. Steel. Estimation of the mean of the population selected from k populations. *J. Statist. Comput. Simulation*, 38(1-4):1–14, 1991. ISSN 0094-9655.

Z. Zhao and J. T. Hwang. Empirical Bayes FCR controlling confidence interval. Submitted, 2010.

Zhigen Zhao. Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Statistics and Its Interface*, 3:533–541, 2010.