On the Credible Interval under the Zero-Inflated Mixture Prior in High Dimensional Inference

October 6, 2011

Abstract: In this paper, we consider the construction of the credible set under the canonical Bayes model when assuming the parameters of interest θ_i's follow a prior distribution which is a mixture of *zero* with probability π_0 and another non-trivial distribution with probability $\pi_1 = 1 - \pi_0$. In modern application with high dimension, π_0 is usually very large, implying that the posterior probability $P(\theta_i = 0|\mathbf{X})$ is also large, saying greater than 5%. The traditional approaches constructing 95% posterior credible intervals, such as highest posterior density (HPD) region or the equal-tail credible interval, will enclose *zero* very often and thus lead to non-informative inference procedures.

In this paper, we use the decision Bayes approach to guide us constructing a mixture credible interval. When there is overwhelming evidence that $\theta_i \neq 0$, we scrutinize the distribution $\pi(\theta_i|\mathbf{X}, \theta_i \neq 0)$ and consider the HPD region. Otherwise, the interval is the union of such a HPD region and *zero*. The paper provides a systematic way in deciding when to include *zero*, guaranteeing the average posterior coverage probability. By separating the zero mass and the non-zero component of the posterior, the resulting intervals have less shrinkage effect especially for the non-zero θ_i's, leading to intervals enclose zero less frequently than its alternatives.

We apply this general approach to a normal mean problem with unknown and unequal variances and model selection. It is demonstrated that the new approach is way more powerful than the traditional and other alternatives.

Keyword: Decision Bayes, Loss Function, Mixture Prior, Two Groups Model.

1 Introduction

High dimensional statistics has been the most popular research area in both theoretical studies and real applications. How to do statistical inference for high dimensional data is an important, urgent, and difficult problem. Recently, there are many literature regarding to the point estimation and hypothesis testing. However, there are relatively much less research in the interval estimation. This doesn't mean that the confidence intervals are not important. Indeed, not only does intervals provide an answer to a yes/no question as hypothesis testing tries to answer, they also provide an effect size estimation for the parameters. This additional information drives the problem more difficult and involves more techniques than a hypothesis testings. Despite its difficulty, there are several attempts, such as [6], [1], [18], [15], [**?**], and many others.

In this paper, we consider the construction of the credible interval under a canonical Bayes model with a zero-inflated prior. We especially focus on the post-posterior inference, namely, the inference after we obtain the posterior draw of $\boldsymbol{\theta}$ given the observation, by using various technique, such as the Gibbs sampling, Metropolis Hasting algorithms and others. This general setting can be applied in various settings, such as the normal mean problem as we have done in Section 4. The credible interval for β_i's under the regression setting is shown in Section 5.

To many statisticians and scientists, the credible interval seems like a naive problem if the posterior distribution of $\boldsymbol{\theta}$ given the observation $\boldsymbol{X}$ is available. Many statisticians will simply stop once they derived the posterior distribution, which is actually far away from the final inference. The high dimensionality and the zero-inflated prior drive such inferences to be difficult, requiring a qualitative change in statistical theory. In [20], they put much effort in explaining ways to summarize the posterior distribution in hypothesis testing. The construction of the credible interval is even more challenging.

For instance in microarray experiments where the dimension p is between several thousands to thirty thousands, most of genes are non-differentially expressed. Consequently, it is reasonable to assume a mixture prior for the parameter θ_i, the true differential expression. With probability π_0, θ_i is identically zero and it follows another non-trivial distribution with probability $\pi_1 = 1 - \pi_0$. Usually, π_0 is very large, e.g., being greater than 90% as stated in [7]. The posterior probability of θ_i being 0 could be also large. In a simulated toy example in Section 2, all these posterior probabilities are greater than 10%, implying that the traditional equal tail credible intervals with 90% coverage probabilities will always enclose zero. The HPD region also includes zero because of the existence of the point mass at zero. Consequently, the usual posterior credible intervals don't provide sound statistical procedure. The same issue arises under the regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ where $\boldsymbol{X}$ is a $n \times p$ matrix with $p >> n$. It is generally assumed that most of β_i's are zero as we stated in Section 5. The posterior probability of β_i being zero given $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be high. We cannot rely on the traditional approach in constructing the posterior credible interval which leads to intervals that enclose zero too often. In Section 3, we adopt the Bayes approach to construct satisfying intervals.

Application of decision approach to interval/set estimation has a long history (see [11], [2], and [14]). Recently, [15] have constructed the double shrinkage *empirical* Bayesian confidence interval when using a non-zero inflated prior for θ_i's. However, the loss functions all the people have used so far need adjustments for the zero-inflated prior we consider here. Otherwise, the decision rule will always enclose zero and the interval is statistical non-informative. We provide detailed argument in Section 3 and a new loss function is thus proposed to deal with the zero-inflated prior for θ_i's. The decision Bayes confidence interval is later derived which is forced to include *zero* if there is overwhelming evidence that θ_i is 0, or equivalently, the local fdr score (see [5], [10]) is large enough. To be precisely, we compare the local fdr score with some constant k_2, a tuning parameter of the loss function. The zero component is forced to be included if the local fdr score is greater than k_2.

The choice of k_2 is very critical and difficult. Many existing literature uses some ad-hoc choice of such parameter, such as 0.2 as suggested by [5], [8], and [**?**]. In our paper, we introduce a systematic way in getting the parameter k_2 which guarantees the controlling of the average posterior coverage probability. The proposed k_2 is often larger than α, implying that the proposed interval doesn't necessarily enclose zero even if $P(\theta_i = 0|\boldsymbol{X}) > \alpha$. Therefore, this decision Bayes intervals are different from the traditional equal tail credible intervals which encloses *zero* whenever $P(\theta_i = 0|\boldsymbol{X}) > \alpha$.

Since we mix zero with another connected set, the derived interval could be disconnected, which is in agreement with a comment from [9], stating that "this kind of disconnected description is natural to the two groups models". To be precise, the estimator we construct for each θ_i is a confidence set. However, we shall still adopt the word "confidence interval" for convenience in the later discussion.

We construct the credible intervals under a canonical model which can be applied under various settings. For instance, we generalize the model considered by [15]. We introduce a zero-inflated prior for θ_i's, derive the Gibbs sampling for this special model, and construct credible intervals for all the parameters, guaranteeing the average posterior coverage probability. Under the regression setting, [17] and [16] proposed the hierarchical Bayes LASSO model. They argued that the advantage of such a hierarchical model to the popular variable selection method, such as LASSO ([21]) is that they can do statistical inference. However, they did not put a zero-inflated prior for the vector $\boldsymbol{\beta}$, which does not reflect the its sparsity structure, as commonly known in application. We therefore introduce the hierarchical representation of the Bayesian LASSO with zero-inflated prior in Section 5. We derive Gibbs sampler for calculating the posterior $\beta_i|\boldsymbol{X}, \boldsymbol{Y}$ and then apply the general interval construction to such a problem. By consider the zero-inflated in both the normal mean and regression problem, the derived intervals have less shrinkage effect especially for those non-zero parameters, and thus don't enclose zero as often as the other commonly used approach. The failure of the traditional equal tail credible intervals has been demonstrated in numerical studies.

Here is how the article is organized. In Section 2, we introduce the model setting and provide a toy example showing how the traditional intervals fail. In Section 3, we introduce a newly proposed loss function and derive the decision Bayes intervals for the mixture prior model. We have also provided the choice of the tuning parameter k_2, guaranteeing the coverage probability. In Section 4, we apply this general approach to a normal mean problem with mixture prior for the parameter assuming the variances are unknown and unequal. In Section 4.2, a spike-in data set has been used to assess the performance of our interval approach. In Section 5, we introduce a zero-inflated prior for the hierarchical Bayesian LASSO model and construct the credible intervals for all the regression coefficients. We compare the proposed approach with the equal tail credible intervals and the interval based on the original Bayesian LASSO model of [17], without assuming the zero-inflated prior for $\boldsymbol{\beta}$.

2 Model Assumptions

We will start from the following canonical Bayes model where the observation $\boldsymbol{X}$'s follows

$$\boldsymbol{X}|\boldsymbol{\theta} \sim f(\boldsymbol{x}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ and θ_i follows a zero-inflated prior as

$$\theta_i \begin{cases} = 0 & \text{with probability } \pi_0, \\ \sim \psi(\theta_i) & \text{with probability } \pi_1 = 1 - \pi_0, \end{cases} \tag{2}$$

where π_0 is the prior probability of θ_i being zero. This prior distribution is appropriate in the modern high dimensional data assuming the sparsity. For instance, most of genes in a microarray experiment are non-differentially expressed, and it is generally assumed that π_0 can be as large as 90%. This canonical model can be used in many application, such as the normal mean model with unknown and unequal variances as described in Section 4. One can also incorporate the Dirichlet process when defining the prior distribution $\psi(\theta_i)$ depending on the application. Note that we don't put any independence assumption of either $\boldsymbol{X}$ and $\boldsymbol{\theta}$.

Due to the existence of the mixture of zero component in the prior distribution, the posterior distribution of θ_i is also a mixture of zero and another distribution. Direct calculation shows that

$$\psi(\theta_i|\boldsymbol{X}) \begin{cases} = 0 & \text{with probability } fdr_i(\boldsymbol{X}), \\ \sim \psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0) & \text{with probability } 1 - fdr_i(\boldsymbol{X}), \end{cases} \tag{3}$$

where

$$fdr_i(\boldsymbol{X}) = P(\theta_i = 0|\boldsymbol{X}).$$

Here the $fdr_i(\boldsymbol{X})$ is the local fdr as given in [5, 8, 9, 10]. Concisely speaking, the posterior distribution of θ can be written as

$$\psi(\theta_i|\boldsymbol{X}) = fdr_i(\boldsymbol{X})\delta_0 + (1 - fdr_i(\boldsymbol{X}))\psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0).$$

In this article, we construct the $(1-\alpha)$ Bayes credible interval for θ_i based on the posterior distribution of θ_i. This seems to be a naive question. For instance, statisticians might simply construct such an interval by throwing away certain portions on each tail of the posterior with the total probability of $100\alpha\%$. Such constructions include the equal-tail posterior interval which guarantees that $P(\theta_i \in CI_i|\boldsymbol{X}) \geq 1 - \alpha$, $\forall i = 1, 2, \cdots, p$. However, such a construction is problematic especially in the high dimensional data.

Theorem 2.1 *Let CI_i be a posterior interval for θ_i such that $P(\theta_i \notin CI_i|\boldsymbol{X}) \leq \alpha$. If $fdr_i(\boldsymbol{X}) > \alpha$, then $0 \in CI_i$.*

Consequently, this posterior credible intervals appear statistically non-informative whenever $fdr_i(\boldsymbol{X}) > \alpha$. Furthermore, when π_0 is close to 1, it is quite often to obtain a local fdr score, exceeding the α level. To better understand this, consider the following toy example.

Toy Example:

Assume that $X_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$ and $\theta_i = 0$ with probability $\pi_0 = 0.8$ and $\theta_i \sim N(0, 1)$ with probability $\pi_1 = 0.2$. The dimension $p = 10,000$. We randomly generate the i.i.d. random vectors (X_i, θ_i)'s, $i = 1, 2, \cdots, p$ and calculate $fdr_i(\boldsymbol{X})$ for each θ_i. In Figure 1, we have plotted the histogram of $fdr_i(\boldsymbol{X})$. It is seen clearly that almost all the $fdr_i(\boldsymbol{X})$'s exceed the 10% level. In other words, for those observations, any posterior credible intervals include zero and appear to be statistically useless. We shall calculate the percentage of X_i's which have a local fdr scores greater than 10%. Under this setting, it is easily known that

$$fdr_i(\boldsymbol{X}) = P(\theta_i = 0|\boldsymbol{X}) = \frac{\pi_0\phi(x_i)}{\pi_1\phi(x_i) + \pi_1\phi(x/\sqrt{2})/\sqrt{2}}.$$

Therefore, $P(fdr_i(\boldsymbol{X}) > 10\%) = P(|X_i| < 3.96) = 0.1\%$. In other words, more than 99.9% of the posterior intervals include zero and the construction is thus very non-informative.

Another commonly used method in constructing the Bayesian confidence intervals is using the highest posterior density (HPD) region based on the posterior $\psi(\theta_i|\boldsymbol{X})$. Due to the existence of point mass of the posterior at *zero*, the density function around *zero* becomes infinity. Consequently, the HPD region for all parameters θ_i's will cover *zero* as well.

Such a phenomenon has been well known in the hypothesis testing. It appears to be powerless if we reject a null hypothesis if and only if the local fdr is less than α due to the relatively large local fdr scores in high dimension. In [9], he suggested the threshold of the local fdr for rejecting a hypothesis as 0.2, an ad-hoc chosen constant. In [20], they have argued that it is important to seperate the zero component and $P(\theta_i|\boldsymbol{X}, \theta_i \neq 0)$. They further applied the decision theory to derive the testing procedure with a loss function involving a tuning parameter c which is chosen ad-hoc. [19] considered the cumulative local fdr. However, none of these papers provide an answer to the construction of confidence intervals. Simply inverting the acceptance region doesn't offer an interval with a guaranteed coverage probability. It thus requires much more work and thinking in constructing valid statistical sound intervals, in terms of good coverage probability and large power in identifying nonzero parameters. In the next section, we will introduce the decision Bayes approach for the interval construction.

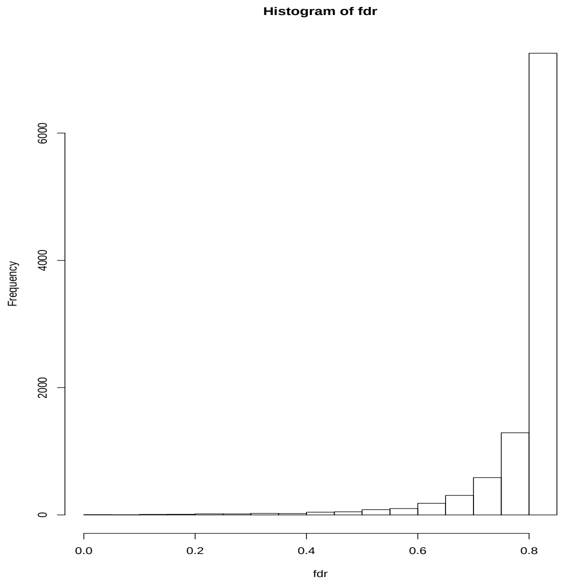

Figure 1: This plots the histogram of $\pi_i^0 = p(\theta_i = 0|\boldsymbol{X})$ when assuming $x_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$ and $\theta_i = 0$ with probability 80% and $\theta_i \sim N(0, 1)$ with the rest of the probability. We generate the random sample pairs (θ_i, x_i) 10,000 times and calculate the posterior probability of θ_i being zero. Here is the output of the histogram.

3 Decision Bayes Confidence Interval

Historically, there have been many attempts to apply the decision Bayes approach to construct confidence sets/intervals. [11] considered a linear loss function for confidence set CS of the vector $\boldsymbol{\theta}$ as $L(\boldsymbol{\theta}, CS) = kVolume(CS) - I_{CS}(\boldsymbol{\theta})$. Also [2] uses the same loss where the tuning parameter k was determined so that the usual $1-\alpha$ confidence set is minimax. [14] used $L(\theta_i, CI_i) = kLen(CI_i) - I_{CI_i}(\theta_i)$ as the loss function for the interval estimator CI_i of the parameter θ_i. [15] modified the loss function above as $L(\theta_i, CI_i) = \frac{k}{\sigma_i}Len(CI_i) - I_{CI_i}(\theta_i)$ when assuming unknown and unequal variances and constructed the confidence interval that shrinks both the means and variances. However, all these loss functions are not appropriate for the zero-inflated prior model (1-2) with θ_i following a mixture of zero and another distribution $\psi(\theta_i)$.

If assuming any of the loss function listed above, for any given confidence interval CI_i^{old}, one can construct a new interval $CI_i^{new} = CI_i^{old} \cup \{0\}$. Because of the positive mass at zero, this new approach boosts the coverage probability and causes no change in the length. Consequently, the risk of CI_i^{new} is always less than or equal to that of the original approach CI_i^{old}, indicating that the inclusion of zero is always beneficial. Consequently CI_i^{new} dominates CI_i^{old} and the decision rule suggests that *zero* should always be included, leading to a powerless interval.

In order to avoid this paradox, we should penalize the inclusion of *zero* when θ_i is indeed non-zero. We thus define the loss function as,

$$L(\boldsymbol{\theta}, \boldsymbol{CI}) = \sum_i \{k_1^i(Len(CI_i) - I_{CI_i}(\theta_i))1(\theta_i \neq 0) + 1(0 \in CI_i)(k_2 - 1(\theta_i = 0))\}, \tag{4}$$

where $0 \leq k_2^i \leq 1$. Inside the summation, the first term balances the length and the true coverage. The second term $1(0 \in CI_i)(k_2 - 1(\theta_i = 0))$ affects the loss function only when the corresponding interval does include *zero*. If $0 \in CI_i$ and θ_i is indeed *zero*, then $k_2 - 1(\theta_i = 0) = k_2 - 1 \leq 0$, and including *zero* reduces the loss and is beneficial. On the other hand, if θ_i is non-zero, the second term $k_2 - 1(\theta_i = 0) = k_2$ is positive and becomes a penalty term. Unlike the loss functions we have mentioned earlier in this section, the inclusion of zero is not always beneficial. Ideally, we only want the interval to include zero if $\theta_i = 0$, as suggested by using such a loss function. When θ_i is nonzero, the tuning parameter k_2 decides the amount of penalty on the inclusion of zero. When assuming a non-zero-inflated prior, the loss function (4) reduces to the loss function to the usual loss by setting $k_2 = 0$. In the above loss function, we allow the tuning parameter k_1^i depend on the observation later in the discussion. Otherwise, it will lead to some paradox, as demonstrated in [3].

Now, we have all the pieces to construct the decision Bayes rule, i.e. we want to construct Bayes interval CI_i^B's, minimizing $E(L(\boldsymbol{\theta}, \boldsymbol{CI}|\boldsymbol{X}))$ for any observation $\boldsymbol{X}$ when assuming the mixture model (1-2) and the loss function (4).

Theorem 3.1 *Assume the model (1-2) and the loss function (4), then conditioning on $\boldsymbol{X}$, we have*

$$EL(\boldsymbol{\theta}, \boldsymbol{CI}|\boldsymbol{X}) = \sum_i \{fdr_i(\boldsymbol{X}) \int_{CI} (k_1^i - \psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0)) d\theta_i + 1(0 \in CI_i|\boldsymbol{X})(k_2 - fdr_i(\boldsymbol{X}))\}. \tag{5}$$

Consequently, the decision Bayes interval for θ_i is

$$CI_i^{DB} = \begin{cases} \{\theta : k_1^i < \psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0)\} \setminus \{0\} & \text{if } fdr_i(\boldsymbol{X}) < k_2, \\ \{\theta : k_1^i < \psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0)\} \cup \{0\} & \text{if } fdr_i(\boldsymbol{X}) \geq k_2. \end{cases} \tag{6}$$

The single point zero is included in the interval if there is overwhelming evidence that $\theta_i = 0$, measured by the local fdr score. If the score is smaller than the tuning parameter k_2, implying that there is relatively overwhelming evidence that the corresponding parameter θ_i is non-zero, the *zero* is excluded from the interval. If $P(\theta_i = 0|\boldsymbol{X}) = fdr_i(\boldsymbol{X}) > k_2$, implying that the parameter is indeed zero, we force the interval to include *zero*.

In [18], they consider the mixture interval

$$CI_i^{QH} = \begin{cases} MX_i \pm \sqrt{M}\sigma_i z_{\alpha/2} & |X_i| > c \\ MX_i \pm \sqrt{M}\sigma_i z_{\alpha/2} \cup \{0\}, & |X_i| \leq c \end{cases}$$

when assuming the following normal mixture model where

$$X_i|\theta_i \sim N(\theta_i, \sigma^2), \theta_i \sim \pi_0 1(\theta_i = 0) + \pi_1 N(0, \tau^2).$$

with known and equal variance σ^2. Under such an assumption, $fdr_i(\boldsymbol{X})$ is monotonic decreasing with respect to $|x|$. Appropriate choice of k_2 and c results in the same intervals CI_i^{HQ} and CI_i^{DB}. However, when $\mu = E(\theta_i|\theta_i \neq 0)$ is nonzero or the variance σ_i^2 for each X_i is unequal and unknown, one should not use the absolute value of x alone to decide whether zero should be included. Additionally, such a construction requires a very strong parametric setting and independence of X_i's and thus becomes less practical. We generalize their intervals to a much broader setting and provide a theoretical justification for the mixture type intervals.

In the mixture interval (6), the major component $\{\theta_i : k_1^i < \psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0)\}$ relies on the tuning parameters k_1^i's, and the posterior density of θ_i given X_i and $\theta_i \neq 0$. One can thus choose the k_1^i such that the resulting interval is a $(1-\alpha)$ HPD interval $CI_i(\alpha)$ based on $\psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0)$. The decision Bayes interval can be simplified as

$$CI_i^M = \begin{cases} CI_i(\alpha) \setminus \{0\}, & \text{if } fdr_i(\boldsymbol{X}) < k_2, \\ CI_i(\alpha) \cup \{0\} & \text{if } fdr_i(\boldsymbol{X}) \geq k_2. \end{cases} \tag{7}$$

It is worthy to mention that the interval construction (7) can be disconnected when $0 \notin CI_i(\alpha)$ and $fdr_i(\boldsymbol{X}) \geq k_2$. This happens when the posterior mass of zero is high and $\psi(\theta_i|\boldsymbol{X}, \theta_i \neq 0)$ centers around a value which is far away from zero. It is natural and necessary to consider such a mixture interval as stated in [9]. On the other hand, one should not rely on $fdr_i(\boldsymbol{X})$ solely to decide whether the interval CI_i^M include zero. One can declare a parameter to be non-zero if $fdr_i(\boldsymbol{X})$ is small enough and $0 \notin CI_i$. Consequently, we modify the decision Bayes rule (7) as

$$CI_i^M = \begin{cases} CI_i, & \text{if } fdr_i(\boldsymbol{X}) < k_2, \\ CI_i \cup \{0\} & \text{if } fdr_i(\boldsymbol{X}) \geq k_2. \end{cases} \tag{8}$$

Theorem 3.1 shows that the tuning parameters k_2 is the key threshold in deciding when to include *zero*. A natural question is how to choose such parameters k_2's. It is tempting to use α as k_2 and we hence force the interval to enclose *zero* if $fdr_i(\boldsymbol{X}) = P(\theta_i = 0|\boldsymbol{X}) \geq \alpha$. However, as illustrated in Section 2, the value of $fdr_i(\boldsymbol{X})$ tends to be large especially for large dimension and large π_0 and such a choice of k_2 leads to statistically useless intervals. In [9], he suggested a value 0.2 as a threshold in rejecting a null hypothesis. There is no guarantee that such an ad-hoc chosen this choices can lead to a valid confidence intervals with a good coverage probability. In what follows, we will introduce a systematic way in choosing such a tuning parameter.

Theorem 3.2 *Assume the model (1-2), and the interval is constructed according to (8) where k_1^i's are chosen such that $P(\theta_i \in CI_i|\boldsymbol{X}, \theta_i \neq 0) = 1-\alpha$, then the average posterior coverage probability can be controlled as*

$$\frac{1}{p}\sum_i P(\theta_i \notin CI_i^M|\boldsymbol{X}) \leq \alpha + \frac{1}{p}\sum_i fdr_i(\boldsymbol{X})(I(fdr_i(\boldsymbol{X}) < k_2) - \alpha)$$

Let

$$k_2 = argmax_{k_2}\{k_2 : \frac{1}{p}\sum_i fdr_i(\boldsymbol{X})(I(fdr_i(\boldsymbol{X}) < k_2) - \alpha) \leq 0\}. \tag{9}$$

According to Theorem 3.2, the choice of k_2 in (9) guarantees that the average posterior coverage probability is no less than $1 - \alpha$. This criteria is different than the controlling of the posterior coverage probability $P(\theta_i \in CI_i|\boldsymbol{X})$ for every $i = 1, 2, \cdots, p$. This criteria is of little interest if there is only one parameter θ. However, it is appropriate when the dimension p is very large and we are constructing multiple credible intervals. For those parameters with $fdr_i(\boldsymbol{X})$'s being large, the intervals enclose zero and thus have large posterior coverage probability. Consequently, we can adjust the confidence coefficient $P(\theta_i \in CI_i|\boldsymbol{X})$ to be smaller for those parameters with small local fdr scores. Namely, the posterior coverage probability levels are not universal across all $i = 1, 2, \cdots, p$. It various according to the evidence of the parameters being zero and non-zero. When we reject a hypothesis, we then reduce the corresponding confidence level to be at least $(1 - fdr_i(\boldsymbol{X}))(1 - \alpha)$. If $fdr_i(\boldsymbol{X}) < k_2$ and θ_i is indeed non-zero, it is guaranteed that the posterior probability $P(\theta_i \in CI_i|\boldsymbol{X}, \theta_i \neq 0) \geq 1 - \alpha$.

Furthermore, when assuming that (X_i, θ_i)'s, $i = 1, 2, \cdots, p$ are identically distributed and we construct the confidence interval for all the θ_i's according to (8), then it is guaranteed that at least $100(1 - \alpha)\%$ of the intervals covers the true parameter.

The choice of k_2 according to (9) is much larger than α in general. Consequently, the proposed confidence interval (8) is much more informative than the traditional approach. In the toy example we have mentioned in Section 2, the numerical calculation shows that k_2 is 0.739. As argued in Section 2, less than 0.1% percent of the intervals do not enclose zero. However, according to the mixture construction, it can be show that $P(fdr_i(\boldsymbol{X}) < k_2) = 12.7\%$, which is way larger than the traditional posterior based approach.

In summary, we proposed the mixture confidence interval (8), and introduced a systematic way in choosing the tuning parameter k_2 in this section. This interval is constructed under a canonical Bayes model and can thus be applied in various application settings, such as the normal mean problem in the next section and zero-inflated Bayesian LASSO in Section 5.

4 Application

4.1 Normal mean model with unknown and unequal variances

The intervals we have constructed in Section 3 have broad applications. In this section, we will revisit the confidence interval construction for the normal mean problem with unknown and unequal variances. Let θ_i, $i =$

$1, 2, \cdots, p$ be the parameters that we are interested. For each i, the observation (X_i, S_i^2) is available, satisfying

$$X_i|\theta_i, \sigma_i^2 \overset{ind}{\sim} N(\theta_i, \sigma_i^2), S_i^2|\sigma_i^2 \overset{ind}{\sim} \sigma_i^2 \frac{\chi_d^2}{d}.$$

Here σ_i^2 is the nuisance parameter which is unknown and unequal across all the observations but estimable. In the ANOVA setting, X_i is the ANOVA estimator of θ_i and S_i^2 is the mean-squared error. When the dimension p is large, it seems reasonable and necessary to put a prior distribution for θ_i's as well as for σ_i^2's. It is known that most of the parameters θ_i's are *zero* or close to *zero*. Consequently, it is natural to assume a mixture prior for θ_i in a way that $\theta_i = 0$ with probability π_0 and $\theta_i \sim N(\mu, \tau^2)$ with probability $\pi_1 = 1 - \pi_0$. We put an inverse gamma prior with shape and scale parameter a and b for σ^2. By setting $\pi(\tau^2) \propto \frac{1}{\tau^2}$ and $\mu \sim Unif(-3, 3)$, we have the following hierarchical model:

$$\begin{cases} X_i|\theta_i, \sigma_i^2 \overset{ind}{\sim} N(\theta_i, \sigma_i^2), \\ \theta_i|\tau^2 \overset{iid}{\sim} \pi_0\delta_0 + \pi_1 N(\mu, \tau^2), \\ \pi_0 \overset{ind}{\sim} Beta(k\eta, k(1-\eta)), \\ \frac{S_i^2}{\sigma_i^2}|\sigma_i^2 \overset{ind}{\sim} \frac{\chi_d^2}{d}, \\ \sigma_i^2 \overset{ind}{\sim} INGamma(a, b) \\ \tau^2 \propto \frac{1}{\tau^2}, \mu \sim Unif(-3, 3). \end{cases} \tag{10}$$

In [15], they have approximated the $\log S_i^2$ by a normal distribution and put a lognormal prior for σ^2 and thus defined the so-called Log-normal model. They have further constructed the *empirical* Bayes confidence interval for each parameter θ_i. In that model, they assume that θ_i follows a normal prior $N(\mu, \tau^2)$ without taking the mixture component. [18] has constructed the *empirical* Bayes confidence interval for a mixture prior of θ_i, assuming that the variances σ_i^2's are equal and known. The interval construction we proposed here deals with a model which can be viewed as a generalization of the models in both papers. Dealing with hierarchical Bayes model, we will use the Gibbs sampler to calculate the posterior distribution $\theta|\boldsymbol{X}, \boldsymbol{S}$, which is a mixture of zero and another random distribution. We then summarize this posterior and construct confidence intervals according to (8), guaranteeing the controlling of the coverage probability.

It is worthy to emphasize that the main focus of our work is the post-posterior inference, i.e. constructing the confidence interval based on the posterior distribution of $\boldsymbol{\theta}$ or a random draw from this distribution. How to derive the posterior is not our primary interest. People can use other tools to calculate it depending on their own application.

Gibbs Sampler:

1.

$$\widehat{fdr_i(\boldsymbol{X},\boldsymbol{S})} = p(\theta_i = 0|Rest) = \frac{\pi_0 \frac{1}{\sigma_i}\phi(\frac{x}{\sigma_i})}{\pi_0 \frac{1}{\sigma_i}\phi(\frac{x}{\sigma_i}) + \pi_1 \frac{1}{\sqrt{\tau^2+\sigma_i^2}}\phi(\frac{x-\mu}{\sqrt{\tau^2+\sigma_i^2}})}.$$

2.

$$\theta_i|Rest = \widehat{fdr_i(\boldsymbol{X},\boldsymbol{S})}\delta_0 + (1 - \widehat{fdr_i(\boldsymbol{X},\boldsymbol{S})})N(M_iX_i + (1-M_i)\mu, M_i\sigma_i^2).$$

3. Let $Z_i = 1(\theta_i = 0)$, then

$$\pi_0|Rest \sim Beta(\alpha + \sum_i Z_i, \beta + p - \sum_i Z_i);$$

4.

$$\tau^2|Rest \sim INGamma(\frac{p - \sum_i Z_i}{2}, \frac{\sum_{\theta_i \neq 0}(\theta_i - \mu)^2}{2});$$

5.

$$\sigma_i^2|Rest \sim INGamma(\frac{d+1}{2} + a, \frac{(x_i - \theta_i)^2}{2} + \frac{dS_i^2}{2} + b).$$

6.

$$\mu \sim N(\frac{\sum_{\theta_i \neq 0} \frac{x_i}{\tau^2+\sigma_i^2}}{\sum_{\theta_i \neq 0} \frac{1}{\tau^2+\sigma_i^2}}, \frac{1}{\sum_{\theta_i \neq 0} \frac{1}{\tau^2+\sigma_i^2}}).$$

4.2 Data Analysis

In this section, we apply the procedure derived based on model (10) to an Affymetrix control data set: the golden spike in data set of [4]. All the true parameters of this data set are prechosen and known and it can be used to assess various statistical procedures (see [15], [**?**], [22] and etc.) [4] has a detailed description of this data set.

We download the data from the web site `http://www.elwood9/net/spike` which has already been processed and normalized at the probe set level. The data is of the size 13997×6. For each gene, there are three replicates from both the treatment and control group, denoted as Y_{1i} and Y_{2i} where $i = 1, 2, 3$. Let S_{1i}^2 and S_{2i}^2 be the sample variances of the i-th gene from both the treatment and control group. Define the observations (X_i, S_i^2) for the i-th gene as

$$X_i = \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}, S_i^2 = (\frac{S_{1i}^2}{3} + \frac{S_{2i}^2}{3}).$$

The degrees of freedom d's, calculated by using the Satterthwaite approximation, vary from 2 to 4.

In order to calculate the credible interval, one wants to specify the prior for σ^2. In model (10), the σ^2 is assumed to follow the inverse-gamma prior with the shape parameter a and scale parameter b. To avoid any subjective priors, we set $a = b = 0$, $k = 2$, and $\eta = 1/2$. Consequently, $\pi(\sigma_i^2) \propto \frac{1}{\sigma_i^2}$ and $\pi \sim U(0, 1)$ and both of the priors are non-informative. We are aiming at controlling the coverage probability to be 95%.

We run the Gibbs sampler 101,000 times, with the first 1000 burn-in iteration. We then choose every 10th value as the realization of the posterior density to avoid the dependence. Consequently, there are 10,000 random samples generated from each parameter. In Figure 2, we plot the histogram of the local fdr $fdr_i(\boldsymbol{X}, \boldsymbol{S})$ for all genes. Among all these genes, the minimum value of $fdr_i(\boldsymbol{X}, \boldsymbol{S})$'s is 0.0358. We first calculate the traditional equal-tail confidence intervals, summarized in the third column of Table 1. Such a construction has a valid coverage probability 95.49%. However, all the intervals constructed include zero. Therefore, such a credible interval construction is statistically useless.

The threshold k_2 calculated according to (9) is $k_2 = 0.514$. Among all the 13,997 genes, 1653 of them have the local fdr smaller than k_2, and the intervals for the rest of genes all include *zero*. For each genes, the equal-tail 95% credible interval CI_i based only on the non-zero draws $\theta_i | \boldsymbol{X}, \boldsymbol{S}, \theta_i \neq 0$ is constructed and summarized in the second column of Table 1. The average coverage probability is 97.28%, which is greater than 95%, the nominal level. The average length is 1.91. It turns out that this average length is much larger than 1.16, the average length of the equal-tail intervals. This is not surprising, because the latter one is built upon the mixture distribution $fdr_i(\boldsymbol{X}, \boldsymbol{S})\delta_0 + (1 - fdr_i(\boldsymbol{X}, \boldsymbol{S}))\psi(\theta | \boldsymbol{X}, \boldsymbol{S}, \theta_i \neq 0)$. Because of the existence of the large mass at zero, the variation is very small and there is a large shrinkage effect. On the other hand, we scrutinize the non-zero component and construct the credible interval based on $\theta_i | \boldsymbol{X}, \boldsymbol{S}, \theta_i \neq 0$, leading to intervals with less shrinkage. Consequently, our interval appears to be longer than the naive approach. However, such an approach is necessary for a sound statistical inference. By mixing zero components, the traditional equal-tail intervals appears to be powerless completely. On the other hand, 319 of

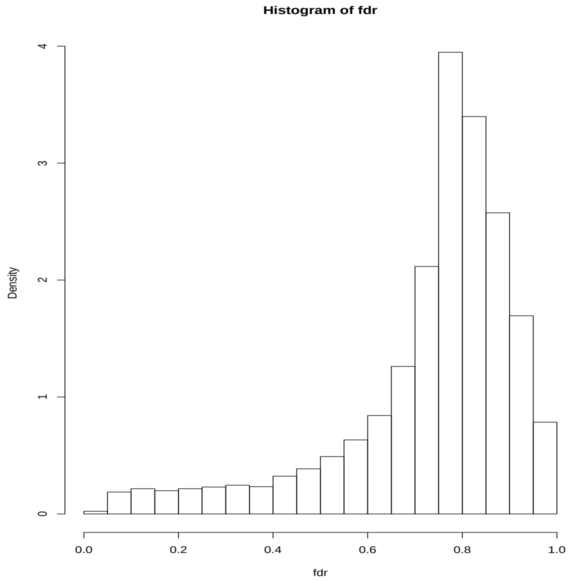

Figure 2: This plots the histogram of $fdr_i(\boldsymbol{X}, \boldsymbol{S})$ for the golden spike in data set with $p = 13,997$. Among all these observations, the minimum value of $fdr_i(\boldsymbol{X}, \boldsymbol{S}) = 0.0358$.

proposed intervals doesn't enclose zero and can be declared as significant. This pattern can be clearly seen in Figure 3, where the two intervals are plotted for the 3-th and 4000-th genes.

Firstly, we consider the 3-th gene. Based on the $\psi(\theta|\boldsymbol{X}, \boldsymbol{S}, \theta_i \neq 0)$, we use the kernel method to estimate the posterior density. The black solid line represents the local fdr $fdr_i(\boldsymbol{X}, \boldsymbol{S})$. The red solid bars represent the proposed interval while the green dashed line is the equal-tail intervals, which encloses zero. The proposed interval is essentially built upon $\psi(\theta_i|\boldsymbol{X}, \boldsymbol{S}, \theta_i \neq 0)$ and is thus longer. However, the interval doesn't enclose zero.

For the 400-th gene, all the intervals cover zero. Since the local fdr score here is large, the equal-tail interval can be much shorter than the proposed approach.

In [6], Efron has shown that the low volume by itself does not always guarantee good inferential properties for the interval construction. In our approach, we sacrifice the width of an interval by separating the *zero* components of the parameter θ_i's. But the interval construction becomes much more statistical informative. We do better because we scrutinize carefully

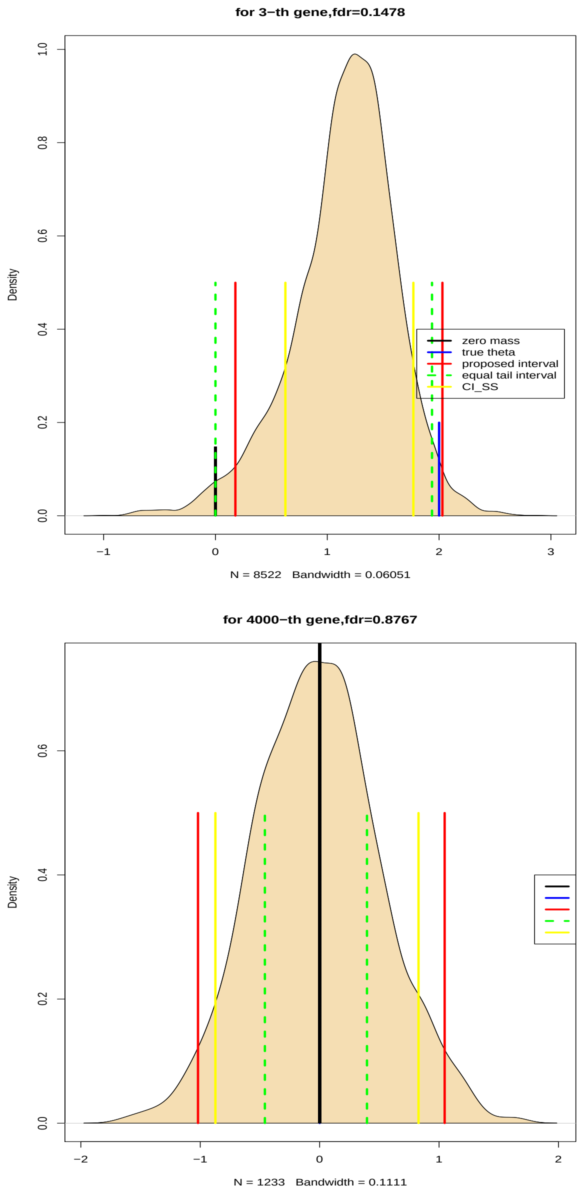

Figure 3: The various confidence intervals constructed for the 3-rd gene which is differentially expressed and the 4000-th gene which is non-differentially expressed.

the non-zero component. Such a consideration is necessary especially in the high dimensional data.

What is also included in Table 1 is the traditional t-interval. Aiming at controlling the coverage probability at 95%, the t-interval only covers the true parameters 88.66% of times. This is very likely due to the Satterthwaite approximation. Nevertheless, our proposed approach still works well. The average length of the t-interval is 3.12, which is even longer than the proposed interval. One of the reasons for the poor performance of the t-interval is that it centers around the unbiased estimator $\boldsymbol{X}$, while the proposed interval centers around a shrinkage estimation of the location parameter. Another observation regarding to this t-intervals is its ability in finding positive genes,. Unfortunately, most of those positives are false positive. Among 1558 genes with the corresponding intervals don't enclose zero, 634 of them are falsely discovered, resulting in a false discovery proportion as 40.69%. Based on the proposed method, that proportion is only 8.15%.

We have also constructed the double shrinkage confidence intervals proposed by [15]. In that construction, they did not use the zero-inflated prior and assume that $\theta_i \sim N(\mu, \tau^2)$. After estimating the hyper-parameter from the data, they proposed *empirical* Bayes double shrinkage confidence intervals. When applied to this data, the coverage probability is 93.5%. However, when focusing on the intervals for those 1331 genes which are differentially expressed, 57.48% of them fails to cover the true parameter. A typical pattern is shown in Figure 3 where CI_{SS} fails to enclose the true parameter from its right side. When assuming such a non-zero-inflated prior for θ_i's, they tends to over shrink the center and the length of the intervals especially for those nonzero θ_i's. It thus suggests less shrinkage effect especially for those differentially expressed genes and it is important to scrutinize the non-zero component of the posterior distribution $\psi(\theta_i|\boldsymbol{X}, \boldsymbol{S}, \theta_i \neq 0)$ as the proposed procedure has done.

In summary, it is easily seen that the proposed interval is the best choice when dealing with the credible intervals when assuming the zero-inflated prior.

5 Application to Variable Selection

In this section, we will apply the general approach (8) to a statistical inference problem in model selection, which gains much attention in the last two decades. Various procedures, such as LASSO ([21]), SCAD ([12]), Screening method, ([13]), and many of their extensions. However, all these approaches

	Proposed	Equal-tail interval	T-interval	CI_{SS}
Coverage Probability(%)	97.28	95.49	88.66	93.5
Average Length	1.91	1.16	3.12	1.68
Number of Significance	319	0	1558	857
Number of False Significance	26	0	634	145
False Discovery Proportion(%)	8.15	0	40.69	16.92
Proportion of Non-Coverage among Differentially Expressed Genes(%)	26.6	47.4	71.6	57.5

Table 1: The coverage probability, average length, number of (false) rejection, and false discovery proportion for various interval construction.

focus on the model selection and its estimation, it is difficult to do any further inference, such as the construction of confidence intervals. Recently, [17] considers the Bayesian version of the LASSO. They derive the Gibbs sampler and construct the credible interval based on the posterior draw of each parameter. They argue that one can use such credible intervals as a criteria in choosing the model. [16] further extend this work to other method, such as Group Lasso, Fused Lasso, Elastic Net.

In these two papers, they assume a non-zero-inflated prior for the parameter $\boldsymbol{\beta}$, which is impractical. In model selection, it is quite common to assume a sparsity structure for the $\boldsymbol{\beta}$, in other words, the proportion of non-zero in the parameter vector $\boldsymbol{\beta}$ is close to 0. Based on this assumption, we introduce the following hierarchical model with a zero-inflated prior for $\boldsymbol{\beta}$.

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \\ \beta_i | \tau_1^2, \cdots, \tau_p^2, \pi_0 \sim (1-\pi_0) N_p(0_p, \sigma^2 D_\tau) + \pi_0 \delta_0 \\ \pi_0 \sim Beta(k\eta, k(1-\eta)) \\ \boldsymbol{D_\tau} = diag(\tau_1^2, \tau_2^2, \cdots, \tau_p^2) \\ \sigma^2, \tau_1^2, \cdots, \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod \frac{\lambda^2}{2} \exp(-\lambda^2 \tau_j^2/2) d\tau_j^2 \\ \lambda^2 \sim \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} e^{-\delta\lambda^2}. \end{cases} \tag{11}$$

This model is a generalization of Model (5) in [17]. We call this a zero-inflated Bayesian LASSO model. Next, we will derive the Gibbs sampler to generate a draw from the posterior distribution $\psi(\beta_i | \boldsymbol{X}, \boldsymbol{Y})$.

Define the parenthesis operator (i) as

$$\boldsymbol{\beta}_{(i)} = (\beta_1, \cdots, \beta_{i-1}, \beta_{i+1}, \beta_p),$$

and

$$\boldsymbol{X}_{(i)} = (X_1, \cdots, X_{i-1}, X_{i+1}, X_p)$$

where X_i is the i-th column of the matrix $\boldsymbol{X}$.

Gibbs Sampling

1.

$$\begin{aligned} & \widehat{fdr_i(\boldsymbol{X},\boldsymbol{Y})} = P(\beta_i = 0|rest) \\ = & \frac{\pi_0}{\pi_0 + \pi_1\sqrt{\frac{1}{1+\tau_i^2 X_i^T X_i}}\exp(\frac{[(\boldsymbol{Y}-\boldsymbol{X}_{(i)}\beta_{(i)})^T X_i]^2}{2(X_i^T X_i+\frac{1}{\tau_i^2})\sigma^2})}; \end{aligned}$$

2.

$$\beta_i|Rest \sim \widehat{fdr_i(\boldsymbol{X},\boldsymbol{Y})}\delta_0 + (1-\widehat{fdr_i(\boldsymbol{X},\boldsymbol{Y})})N(\frac{(\boldsymbol{Y}-\boldsymbol{X}_{(i)}\beta_{(i)})^T X_i}{X_i^T X_i + \frac{1}{\tau_i^2}}, \frac{\sigma^2}{X_i^T X_i + \frac{1}{\tau_i^2}});$$

3. Let $Z_i = 1(\beta_i = 0)$, then

$$\pi_0|rest \sim Beta(a + \sum_i Z_i, b + p - \sum_i Z_i);$$

4.

$$\tau_i^2|rest \sim \begin{cases} \frac{\lambda^2}{2}\exp(-\lambda^2\tau_j^2/2)d\tau_j^2, & \text{if} \quad \beta_i = 0 \\ (INGaussian(\sqrt{\frac{\lambda^2\sigma^2}{\beta_i^2}}, \lambda^2))^{-1}, & \text{if} \quad \beta_i \neq 0. \end{cases}$$

5. $\lambda^2|rest \sim INGamma(shape = p + r, rate = \sum_j \tau_j^2/2 + \delta);$

6. $\sigma^2|rest \sim INGamma((n-1)/2 + \#\{\beta_i \neq 0\}/2, \frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\boldsymbol{\beta}^T D_\tau \boldsymbol{\beta}}{2}).$

5.1 simulation

We simulate data from the true model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \epsilon_i \sim N(0,1), for \quad i = 1, 2, \cdots, n.$$

The design matrix $\boldsymbol{X}$ and the $\boldsymbol{\beta}$ are chosen according to the following three examples.
Example 1 Set $p = 8$, $n = 20$, $\Sigma = (\sigma_{ij})_{i,j=1,2,\cdots,p}$ with $\boldsymbol{\Sigma} = (\sigma_{ij})$ with

$\sigma_{ij} = 0.5^{|i-j|}$, and $\sigma = 3$. The parameter vector $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$.

Example 2
Set $p = 40, n = 100$, $cor(x_i, x_j) = 0.5, i \neq j$, and $\sigma = 15$. The vector $\boldsymbol{\beta} = (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{2})$, where $\mathbf{0}$ and $\mathbf{2}$ are 10-dimensional vector.

Example 3
The setting of this simulation is similar to the setting of Example 2 except that we set σ as 7.

Example 4
Set $p = 500$, $n = 100$, and $\sigma = 3$. The p-dimensional vector $\boldsymbol{\beta}$ has only 10 non-zero entries. Five of them equal to 2, located at the 1st, 20th, 40th, 60th, 80th position, and the other five equal to -2, located at the 100th, 200th, 300th, 400th, 450th position. The design matrix $\boldsymbol{X}$ is generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}_p$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1,2,\cdots,p}$ with $\sigma_{ij} = 0.5^{|i-j|}$.

Example 5
Set $p = 100$, $n = 40$, $\boldsymbol{\Sigma}$ is the same as in Example 4. The vector $\boldsymbol{\beta}$ has the element entries $\beta_1 = 2, \beta_{10} = -1.5, \beta_{50} = -2, \beta_{75} = -1.5$, and the rest of the entries are zero. $\sigma = 3$.

We set the hyper parameter $k = 2$, $\eta = 1/2$ for the Beta distribution to avoid any subjective choice. The hyper parameter of the Gamma distribution for λ^2 is chosen as $r = 1$ and $\delta = 10$, resulting in a relative flat prior. We then run the Gibbs sampler derived above 11000 times, with the first 1,000 as the burn-in iterations. For the rest 10,000 iterations, we choose every 10th generated value to avoid sequence dependence. We then construct the regular equal-tail credible interval, and the proposed interval (8). What is also included is the credible interval based on the Bayesian LASSO with assuming the zero-inflated prior. After constructing the 95% confidence intervals for all the parameters, we select the parameter if the corresponding interval does not enclose zero, and count the number of correctly chosen predictors and false selected predictors. We replicate this simulation 100 times, and report the average number of correct and wrong predictors in Table 2. In each cell, the first number is the average number of correctly chosen predictors and the second one is the average number of wrongly chosen predictors. Under all settings, the proposed approach can find the most number of true predictors, with a very reasonable number of false predictors. When the proportion of non-zero component in $\boldsymbol{\beta}$ is large as in Example 1, 2, and 3,

	Proposed	Equal-Tail	No zero-inflated
Example 1	1.35/0.06	0.46/0.01	1.25/0.03
Example 2	2.42/0.49	0.39/0.04	1.53/0.28
Example 3	5.61/0.42	2.17/0.06	4.81/0.28
Example 4	6.85/2.07	2.96/0.00	0.30/0.00
Example 5	1.67/0.56	0.28/0.01	0.93/0.16

Table 2: Simulation result of BLASSO.

the intervals without assuming the zero-inflated prior works slightly worse than the proposed approach. The discrepancy enlarges when the proportion is getting smaller as shown in Example 4 and 5. In Example 4, the equal-tail credible intervals based on the Bayesian LASSO model only choose 0.3 correct predictors on average, which is much smaller than 6.85, the average number of true predictors found by the proposed approach when putting a zero-inflated prior for $\boldsymbol{\beta}$.

Similarly as argued in Section 4, by separating the zero mass and non-zero mass in the posterior distribution $\psi(\beta_i|\boldsymbol{X})$, the proposed approach has less shrinkage effect especially for those non-zero parameters than the approach without assuming zero-inflated prior. It thus produces better inference procedure. This traditional equal-tail approach, applied in both [17] and [16], appears to be very non-informative under the zero-inflated model. When assuming the zero-inflated prior, it is necessary to consider the proposed approach because the equal-tail credible intervals is conservative in terms of finding much less number of true predictors than the proposed approach. Consequently, we recommend our approach for applications.

6 Conclusion

In this article, we construct the Bayes credible intervals when assuming the prior of θ_i's follow zero-inflated prior. Since the existence of point mass in the prior distribution with high probability, the traditional method in constructing the credible interval for θ_i's satisfying $P(\theta_i \in CI_i|X) \geq 1-\alpha$ by using $\psi(\theta_i|X)$ solely lacks power due to the large local fdr score. Instead, we are aiming at constructing intervals CI_i such that the average posterior coverage probability is at least $1-\alpha$. We modify the decision Bayes rule and derive a mixture confidence interval which is forced to enclose zero if there is strong evidence that the parameter is zero, determined by the local fdr score. We apply this general approach to both the normal mean model and regression model. The proposed approach appears to be more statistically sound than all its alternatives.

In this article, we have applied this general approach to the normal mean problem with unknown and unequal variances and the zero-inflated Bayesian LASSO model. It is demonstrated that the proposed intervals produce much better statistical inference procedure.

References

[1] 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.

[2] 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.

[3] G. Casella, J. T. Hwang, and C. Robert. A paradox in decision-theoretic interval estimation. *Statist. Sinica*, 3(1):141–155, 1993. ISSN 1017-0405.

[4] 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.

[5] B. Efron. Local false discovery rates. 2005.

[6] B. Efron. Minimum volume confidence regions for a multivariate normal mean vector. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(4):655–670, 2006. ISSN 1369-7412.

[7] B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[8] Bradley Efron. Size, power and false discovery rates. *Ann. Statist.*, 35(4):1351–1377, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001460. URL http://dx.doi.org/10.1214/009053606000001460.

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

[10] Bradley Efron. *Large-scale inference, empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press, 2010.

[11] R. E. Faith. Minimax Bayes point and set estimators of a multivariate normal mean. *Unpublished Ph.D. dissertation, Department of Statistics, University of Michigan*, 1976.

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[13] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

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

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

[16] M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

[17] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. ISSN 0162-1459.

[18] 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.

[19] S.K. Sarkar, T. Zhou, and D. Ghosh. A general decision theoretic formulation of procedures controlling fdr and fnr from a Bayesian perspective. *Statist. Sinica*, 18(3):925–945, 2008.

[20] J.G. Scott and J.O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7): 2144–2162, 2006. ISSN 0378-3758.

[21] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58

(1):pp. 267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

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

Appendix: Sketch of Technical Arguments

Proof of Theorem 2.1: It is easily seen that

$$\begin{aligned} & P(\theta_i \in CI_i|X) \\ = \quad & P(\theta_i \in CI_i|X, \theta_i = 0)P(\theta_i = 0|X) + P(\theta_i \in CI_i|X, \theta_i \neq 0)P(\theta_i \neq 0|X). \end{aligned}$$

If $0 \notin CI_i$, then

$$P(\theta_i \in CI_i|X) \leq P(\theta_i \in CI_i|X, \theta_i \neq 0)(1 - fdr_i(X)) < 1 - \alpha,$$

which leads to a contradiction. Consequently, $0 \in CI_i$. ∎

Proof of Theorem 3.1. Let $L_i(\vec{\theta}, CI)$ be the i-th component of the loss function. Then $EL(\vec{\theta}, CI) = \sum_i EL_i(\vec{\theta}, CI)$. Firstly,

$$\begin{aligned} & EL_i(\vec{\theta}, CI|X) = k_1^i Len(CI_i) P(\theta_i \neq 0|X) \\ - \quad & \int 1(\theta_i \in CI, \theta_i \neq 0)\pi(\theta_i|X)d\theta_i + I(0 \in CI_i|X)(k_2^i - fdr_i(X)). \end{aligned} \tag{12}$$

The integration $\int 1(\theta_i \in CI_i, \theta_i \neq 0)\pi(\theta_i|X)d\theta_i$ can be written as $\int_{CI_i} \pi(\theta_i, \theta_i \neq 0|X)d\theta_i$ where $\pi(\theta_i, \theta_i \neq 0|X) = (1 - fdr_i(X))\pi(\theta_i|X, \theta_i \neq 0)$. Write $Len(CI_i)$ as $\int_{CI_i} 1 d\theta_i$. Then (12) equals to

$$(1 - fdr_i(X)) \int_{CI_i} (k_1^i - \pi(\theta_i|X, \theta_i \neq 0))d\theta_i + I(0 \in CI_i|X)(k_2^i - fdr_i(X)). \tag{13}$$

The minimizer of the first integration is given by $\{\theta_i : k_1^i < \pi(\theta_i|X, \theta_i \neq 0)\}$. Now consider two intervals CI_i^1 and CI_i^2 where $CI_i^1 = \{\theta_i : k_1^i < \pi(\theta_i|X, \theta_i \neq 0)\} \setminus \{0\}$ and $CI_i^2 = \{\theta_i : k_1^i < \pi(\theta_i|X, \theta_i \neq 0)\} \cup \{0\}$. Then both CI_i^1 and CI_i^2 minimize the first integration of (13). Since $0 \in CI_i^2$ and $0 \notin CI_i^1$, then

$$EL_i(CI_i^2|X) = EL_i(CI_i^1|X) + (k_2^i - fdr_i(X)).$$

Consequently, the Bayes interval includes 0 if and only if $k_2^i < fdr_i(X)$, i.e. it is the one that is defined in (8). ∎

Proof of Theorem 3.2

Consider the posterior non-coverage probability $P(\theta_i \notin CI_i^M|X)$,

$$\begin{aligned}
& P(\theta_i \notin CI_i^M|X) \\
= \; & P(\theta_i \notin CI_i^M|X, \theta_i = 0)P(\theta_i = 0|X) + P(\theta_i \notin CI_i^M|X, \theta_i \neq 0)P(\theta_i \neq 0|X) \\
= \; & fdr_i(X)P(\theta_i \notin CI_i^M|X, \theta_i = 0) + (1 - fdr_i(X))P(\theta_i \notin CI_i^M|X, \theta \neq 0) \\
\leq \; & fdr_i(X)I(fdr_i(X) < k_2^i) + \alpha(1 - fdr_i(X)) \\
= \; & \alpha + fdr_i(X)(I(fdr_i(X) < k_2^i) - \alpha).
\end{aligned}$$

Consequently,

$$\begin{aligned}
& \frac{1}{p}\sum_i P(\theta_i \notin CI_i^M|\mathbf{X}) \\
\leq \; & \alpha + \frac{1}{p}\sum_i fdr_i(X)(I(fdr_i(X) < k_2^i) - \alpha).
\end{aligned}$$

∎