# April 11, 2019

## 1 Introduction

Problem involving a large number of parameters, inference is often made on parameters that are selected based on the data.

### 1.1 Winner's Curse

- The bias introduced by the selection
- Cause the usual confidence interval to have an extremely low coverage porbability

### 1.2 Toy Example

$$Y_i|\beta_i \overset{ind}{\sim} N(\beta_i, 1)$$
$$\beta_i \sim \begin{cases} 0 & \text{with probability } 0.8 \\ N(0,1) & \text{with probability } 0.2 \end{cases}$$

Let $Y_{(1)} = \max_{1\leq i\leq p} Y_i$ and $\beta_{(i)}$ be the coreesponding parameter. Constructing usual 95% confidence interval. $CI_{(1)} = Y_{(1)} \pm 1.95$. As reapeated 10,000 times to simulate the coverage probability. It was 42.4% and it is easy to see that $E[Y_{(1)}] \geq \beta_{(1)}$.

- Developing a confidence interval for a selected parameter that is statistically sound is important
- Constructing confidence intervals for multiple selected parameters subject to controlling an overall measure of false coverage

### 1.3 Zero Inflated Mixture Prior(ZIMP)

$$\boldsymbol{Y}|\boldsymbol{\beta} \sim f(\boldsymbol{y}|\boldsymbol{\beta})$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and

$$\beta_i \sim \pi(\beta_i) = \pi_0 1(\beta_i = 0) + (1-\pi_0)\psi(\beta_i)$$

$\pi_0$ is the prior of $\beta_i$ being zero, and $\psi(\beta_i)$ is the distribution of $\beta_i$ given $\beta_i \neq 0$

- This model is useful in genetic experiments where many of the genes are believed to be non-differentially expressed, and in regressions with sparsity structure.
- If $\pi_0$ is large, the posterior probability of $\beta_i$ being 0 can also large
- this is problematic for equal-tail credible interval since it can obtain zero a high proportions of times
- high-posterior-density(HPD) regions : HPD regions always include zero due to the existence of point mass at zero

### 1.4 Credible interval

We can say that our 100(1–$\alpha$)% credible interval $C$ defines the subset on the parameter space, which we'll call $\theta$, such that the integral

$$\int_C \pi(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta} = 1 - \alpha$$

$\pi$ here is the posterior probability distribution. So, for instance, if you needed a 95 percent credible interval, you would be working to find the interval over which the integral of the posterior probability distribution sums to 0.95

- Choosing the narrowest interval, which for a unimodal distribution will involve choosing those values of highest probability density including the mode. This is sometimes called the highest posterior density interval.
- Choosing the interval where the probability of being below the interval is as likely as being above it. This interval will include the median. This is sometimes called the equal-tailed interval.

### 1.5 Loss Function

Consider a loss function that penalizes the inclusion of zero when $\beta_i$ is indeed non-zero.

- The Bayesian decision interval is then forced to include zero if there is overwhelming ovidence that $\beta_i$ is 0
- equivalently the local fdr score$P(\beta_i = 0|\boldsymbol{Y})$ is large.
- Local fdr score is compared to a tuning parameter $k_2$ used in loss function

- the zero component is included in the interval if the local fdr score is greater than $k_2$

We determine the $k_2$ so that the posterior false coverage rate(PFCR) or the Bayes false coverage rate(BFCR) is controlled at a desired level. $k_2$ is often larger than $\alpha$, the proposed interval doesn't have to include zero even $P(\beta_i = 0|\boldsymbol{Y}) > \alpha$

## 2 Traditional Bayes Credible Intervals

There is two measures of false coverage. First one is Posterior False Coverage Rate(PFCR) and second one is Bayeian False Coverage Rate(BFCR)

Let

- $\mathcal{R}(\boldsymbol{Y})$ be the set of indices of the parameters selected based on the observation $\boldsymbol{Y}$
- $R$ is the total count of $\mathcal{R}(\boldsymbol{Y})$
- Given the credible intervals $CI_i$ for $\beta_i,\ i \in \ \mathcal{R}(\boldsymbol{Y})$
- $\mathcal{V}$ consist of $i \in \ \mathcal{R}(\boldsymbol{Y})$ such that $\beta_i \notin CI_i$
- $V$ is the total count of $\mathcal{V}$
- $Q$ is the proportion of the selected parameters that are not covered by their respective intervals $Q = V/R$ if $\ R > 0$, and $Q = 0$ if $\ R = 0$

### 2.1 False Coverage Rate

$$FCR = E[Q|\beta]$$

Measure of false coverage among the selected parameters in a frequentist sense.

### 2.2 PFCR

$$PFCR(\boldsymbol{Y}) = E[Q|\boldsymbol{Y}] = \begin{cases} \frac{1}{R}\sum_{i\in\mathcal{R}} P(\beta_i \notin CI_i|\boldsymbol{Y}) & \text{if } R > 0 \\ 0 & \text{if R} = 0 \end{cases}$$

### 2.3 BFCR

$$BFCR = \int PFCR(\boldsymbol{Y})m(\boldsymbol{Y})\boldsymbol{dY}$$

where $m(\boldsymbol{Y})$ is the marginal density of $\boldsymbol{Y}$

if $P(\beta_i \notin CI_i|\boldsymbol{Y}) \leq \alpha$ for all $i = 1, 2, \dots, p$ then both PFCR and BFCR are less than or equal to $\alpha$ for any selection rule $\mathcal{R}(\boldsymbol{Y})$. Thus $100(1-\alpha)\%$ credible intervals obtained from posterior can avoid adjusting for selection rule, but do not have good inferential properties.

let $\psi(\beta_i|\boldsymbol{Y}, \beta_i \neq 0)$ be the posterior distribution of $\beta_i$ given $\beta_i \neq 0$

$$\psi(\beta_i|\boldsymbol{Y}) = fdr_i(\boldsymbol{Y})1(\beta_i = 0) + (1 - fdr_i(\boldsymbol{Y}))\psi(\beta_i|\boldsymbol{Y}, \beta_i \neq 0)$$

$fdr_i(\boldsymbol{Y})$ and $\psi(\beta_i|\boldsymbol{Y}$ need to MCMC.

**Theroem 1.** Let $CI_i$ be a posterior interval for $\beta_i$ such that $P(\beta_i \notin CI_i|\boldsymbol{Y}) \leq \alpha$ . If $fdr_i(\boldsymbol{Y}) > \alpha$, then $0 \in CI_i$