Assignment 4

Yang Can, HKUST

2020/4/19

Problem 1

Consider $M = 20,000$ hypotheses and their corresponding p-values $\{P_1, \ldots, P_M\}$. Suppose the p-values are from the following mixture distribution:

$$P_i = \begin{cases} \text{Uniform}[0,1], & \text{if } y_i = 0, \\ Beta(0.5, 1), & \text{if } y_i = 1, \end{cases}$$

where y_i is a variable indicate whether the i-th hypothesis is null ($y_i = 0$) or nonnull ($y_i = 1$). Suppose we also know $\Pr(y_i = 0) = 0.9$ and $\Pr(y_i = 1) = 0.1$.

(a) What is analytic expression of the local false discovery rate, i.e., $\text{fdr}(P) = \Pr(\text{null}|P)$, where P can be any values in (0,1]. Plot function $\text{fdr}(P)$ using R or Python based on your analytic calculation.

(b) Implement Bonferroni correction (BC) and Benjamini-Hochberg Procedure (BHP). Use computer program to check, on average, how many null hypothesis would be rejected by BC (family-wise error rate≤ 0.1) and BHP (global false discovery rate ≤ 0.1). (Hint: you need to generate p-values $\{P_1, \ldots, P_M\}$ from the mixture distribution and apply BC and BHP to this p-value set. Then repeat this process L times, e.g., $L = 1,000$, to check the average numbers.)

(c) Do you know how to check where BHP controls global false discovery rate at a desired level, e.g., ≤ 0.1? (Hint: Compare the hypothesis rejected by BHP to the ground truth given by the indicator variable $\{y_1, \ldots, y_M\}$.)

Problem 2

Consider $M = 20,000$ hypotheses and their z-statistics $z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$,

$$\mu_i = \begin{cases} \mathcal{N}(0.0, \sigma_0^2), & \text{if } y_i = 0, \\ \mathcal{N}(2.5, \sigma_1^2), & \text{if } y_i = 1, \end{cases}$$

where $\sigma_0^2 = \sigma_1^2 = 0.5$ and y_i is a variable indicate whether the i-th hypothesis is null ($y_i = 0$) or nonnull ($y_i = 1$). Suppose we also know $\Pr(y_i = 0) = 0.95$ and $\Pr(y_i = 1) = 0.05$.

(a) Derive the analytic solution for the marginal density function $f(z)$ of z.

(b) Write down the expression for local false discovery rate $\text{fdr}(z) = \Pr(\text{null}|z)$ and plot $\text{fdr}(z)$ using R or Python.

(c) Obtain $\mathbb{E}(\mu_i|z_i)$.

Problem 3

Consider the problem motivating the James-Stein Estimator: $z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, N.$, how to obtain a good estimate of $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^T$ from $\mathbf{z} = [z_1, \ldots, z_N]^T$?

(a) Assign prior distribution $g(\mu)$: $\mu_i \sim \mathcal{N}(0, \alpha^{-1}), i = 1, \ldots, N$. Now consider $\boldsymbol{\mu}$ as the latent variable and α^{-1} as the model parameter. Derive an EM algorithm for parameter estimation and then use the estimated parameter $\hat{\alpha}$ to obtain an estimate of $\boldsymbol{\mu}$.

(b) Implement the above EM algorithm and compare the estimate with the result obtained from the James-Stein estimator.

Problem 4

Let $\mathbf{y} = [y_1, \ldots, y_K]^T$ be a vector of random variables from the multinomial distribution

$$\mathrm{Mult}(\mathbf{y}|\boldsymbol{\mu}, N) = \binom{N}{y_1\, y_2 \ldots y_K} \prod_{k=1}^{K} \mu_k^{y_k},$$

where$0 \leq \mu_k \leq 1$, $\sum_{k=1}^{K} \mu_k = 1$ and $\sum_{k=1}^{K} y_k = N$. Assume that the piror distribution of $\boldsymbol{\mu}$ is the Dirichlet distribution

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

where $\alpha_k > 0$.

(a) Suppose we have collected a data set $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, where

$$\mathbf{y}_i \sim \mathrm{Mult}(\mathbf{y}_i|\boldsymbol{\mu}_i, N), \quad \boldsymbol{\mu}_i \sim \mathrm{Dir}(\boldsymbol{\mu}_i|\boldsymbol{\alpha})$$

Derive an EM algorithm to estimate $\boldsymbol{\alpha}\mathbb{R}^K$, where $\boldsymbol{\mu}$ is viewed as the latent variable. Let $\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \log p(\mathcal{D}|\boldsymbol{\alpha}) = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log p(\mathbf{y}_i|\boldsymbol{\alpha})$ be the MLE, where

$$p(\mathbf{y}_i|\boldsymbol{\alpha}) = \int p(\mathbf{y}_i|\boldsymbol{\mu}_i) p(\boldsymbol{\mu}_i|\boldsymbol{\alpha}) d\boldsymbol{\mu}_i.$$

Can you guarantee that your EM algorithm converges to MLE $\hat{\boldsymbol{\alpha}}$? Explain your reason.

(b) Consider the generative model is as follows:

$$\mathbf{y}_i \sim \mathrm{Mult}(\mathbf{y}_i|\boldsymbol{\mu}_i, N), \quad \boldsymbol{\mu}_i \sim \mathrm{Dir}(\boldsymbol{\mu}_i|\boldsymbol{\alpha}_i), \quad \boldsymbol{\alpha}_i = \exp(\boldsymbol{\beta}^T \mathbf{x}_i).$$

where the parameter $\boldsymbol{\alpha}_i$ in the Dirichlet prior is modulated by side information encoded in $\mathbf{x} \in \mathbb{R}^p$ and $\boldsymbol{\beta}$ is a $p \times K$ matrix. Please derive an EM algorithm to estimate $\boldsymbol{\beta}$.

(c) Suppose the data set $\mathcal{D} = \{\mathbf{y}_i, \mathbf{x}_i\}_{i=1,\ldots,n}$ is generated via the probabilistic model in (b), where $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,4}]^T$ is a vector of $K = 4$ random variables and $\sum_{k=1}^{K} y_{i,k} = 50$, $\mathbf{x}_i$ is a vector of $p = 5$ random variables. Apply your algorithm derived in (b) to the given data set $\mathcal{D}$ to estimate $\boldsymbol{\beta} \in \mathbb{R}^{5\times 4}$. The data set is given in *data.txt*.

Requirement

- You need to submit a report, in which you should clearly describe your method and explain your idea. The code should also be included.
- You can use R or Python for coding.
- Your report should be in the **pdf** or **html** format, which is automatically generated by either R markdown or Jupyter notebook.
- The report is due to May, 7, 23:59 pm, 2020 (HK time).