PHS 43010: Final Project: Bayesian inference clinical trials and nonparametric models

Presentation May 25; reports due May 27.

1 Overview

We are going to spend the next two weeks on two mini-research projects, one for each team. The stat team will take on the nonparametric Bayesian project and non-stat team will take on a project for dose-finding trial designs.

Teams The stat team vs non-stat team

Stat Team: Captain: Ma, Liangzong

Non-Stat Team: Captain: Zhang, Xinyi

Objective To learn Bayesian modeling and inference using hierarchical models for real-life problems.

2 Non-stat Team Project: CRM

Review Read the review paper on CRM (`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6339349/pdf/12874_2018_Article_638.pdf`) and the main references therein. The relevant references are among the first 20 cited papers.

Tasks We will fulfill these tasks.

Review Write a review of the CRM design for phase I dose-finding trials.

Repeat Reproduce results in Figures 6 & 7 by fitting a logistic regression model with two parameters (Table 1 in Wheeler et al., 2019). That is,

$$log\left(\frac{p_d}{1-p_d}\right) = \beta_1 + \beta_2 \cdot d.$$

Research Now consider running simulation studies using the CRM with logistic regression (two parameters) for the ssHHT trial in Figure 6. Use the dose levels in the figure for fitting the logistic regression. Assume the sample size is 36. Assume there are two scenarios, one with true toxicity probability of (0.25, 0.3, 0.5, 0.6, 0.7, 0.75) and the other with true toxicity probability of (0.01, 0.05, 0.2, 0.3, 0.5, 0.6). Conduct 1,000 simulated trials. Report the percentage of trials each dose is selected as the MTD and the average number of patients assigned to each dose over simulated trials.

I am assuming
we follow
their
methodology?

Question 1 Is there a difference in the performance of CRM for the two scenarios?

Question 2 What is the reason for the difference?

Question 3 (Bonus) How can the difference be mitigated?

Report Write up a report and attach the R program for your report. If you can, try to use RMarkdown (`https://rmarkdown.rstudio.com/`) for the report writing. But using RMarkdown is not required.

3 Stat Team Project: HDP

Objective Learn and implement a slice sampler (SS) for the HDP model introduced in the lecture given by Dehua Bi (Thursday). The project will extend the slice sampler for DP, also discussed in the same lecture. Below is the HDP model in the stick-breaking representation.

HDP model: Denote y_{ij} the observation of subject i in group j, $F(\cdot)$ some kernel function, and H the prior distribution for the location ϕ_k. The stick-breaking representation of HDP is:

$$y_{ij}|\theta_{ij} \sim F(\theta_{ij}),\ \theta_{ij}|G_j \sim G_j,$$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk}\delta_{\phi_k},\ \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl}),$$

$$\pi'_{jk} \sim Beta\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right),\ \beta_k = \beta'_k\prod_{l=1}^{k-1}(1-\beta'_l),$$

$$\beta'_k \sim Beta(1,\gamma),\ \phi_k \sim H$$

For simplicity, please assume the observation y_{ij} is univariate. Also assume a normal kernel for $F(\cdot)$, i.e., $F(\cdot) = N(\cdot,\cdot)$, $\theta_{ij} = \{\mu_{ij}, \sigma^2_{ij}\}$, and $\phi_k = \{\mu_k, \sigma^2_k\}$.

Readings Below are the readings and reading assignments.

SS for HDP Divide your team into two subteams. The first team should read the papers on slice sampler for mixture models Walker, 2007; Kalli et al., 2011), and the second team should read the HDP paper (Teh et al., 2004). Each subteam should prepare slides on the review of these papers, especially on the theory of the derivation of the stick-breaking representation of the HDP, the Chinese Restaurant Franchise (CRF) process, and the slice sampler for DP. In addition, it is important to understand the idea of using auxiliary variables to reduce the infinite sum in HDP to a finite sum (since we will be using it in the slice sampler).

Tasks We will fulfill these tasks.

Review Write a review of the readings.

Repeat Implement the Chinese Restaurant Franchise sampling method presented in Teh et al. (2004)'s paper. Choose a prior H (any prior that makes sense in this case, including the conjugate one). You do not need to sample the concentration parameters, i.e., you can set $\alpha_0 = \gamma = 1$.

Research First, write the HDP model using latent indicator variables z_{ij} for each subject i in group j. Hint: G_j is a.s. discrete, so θ_{ij} must be equal to one of the ϕ_ks. How do you write $y_{ij}|\theta_{ij} \sim F(\theta_{ij})$ in terms of the latent indicator z_{ij} and the locations ϕ_k?

Next, implement a slice sampler for the HDP model with z_{ij}, and use it to infer the cluster membership of patients with warts in the immunotherapy (treatment = 1) and cryotherapy (treatment = 0) groups based on a univariate observation of the log area of warts.

Finally, compare inference results obtained using the slice sampler with those obtained using the Chinese Restaurant Franchise sampling method.

Report Write up a report and attach the R program for your report. If you can, try to use RMarkdown (`https://rmarkdown.rstudio.com/`) for the report writing. But using RMarkdown is not required.