THE UNIVERSITY OF HONG KONG

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT6011/STAT4610 COMPUTATIONAL STATISTICS

Please ensure to take note of the following instructions for the take-home midterm project:

- **The project is due at 6 pm on November 13.**
- **All numerical computations must be performed using Python.**
- **Two files should be submitted. The first file should be an executable .ipynb file containing the code and analysis. The second file should be a merged PDF file that includes the .ipynb file as well as any handwritten (or LaTeX-typed) procedures.**
- **Do not submit a .zip file.**
- **Both files should be readable and easily understandable.**
- **When submitting the files, it is essential to include not only the final answer but also the intermediate steps.**
- **Any submissions that do not meet the above requirements will receive a score of 0 without further notice.**

Let p_j denote the toxicity probability of dose level $j, j = 1, \ldots, J$, and the toxicity probability is assumed to increase monotonically with the dose level, i.e., $0 < p_1 < \cdots < p_J < 1$. In a clinical trial with five dose levels, i.e., $J = 5$, the number of observed toxicity y_j and the number of treated patients n_j are

Dose Level j	1	2	3	4	5
y_j	0	1	4	3	0
n_j	3	6	12	6	0

We consider three models (model M_1 is the true model as it satisfies the monotonic constraint):

$$M_1: \ p_1 < p_2 < p_3 < p_4 < p_5$$

$$M_2: \ p_1 > p_2 > p_3 > p_4 > p_5$$

$$M_3: \ p_1 < p_2 < p_3 > p_4 > p_5$$

Given the observed data D, the marginal likelihood under model $M_k (k = 1, 2, 3)$ is

$$P(D|M_k) \propto \int f(p_1, \ldots, p_J|M_k) \prod_{j=1}^{J} \{p_j^{y_j}(1 - p_j)^{n_j - y_j}\} \mathrm{d}p_1 \cdots \mathrm{d}p_J$$

where $f(p_1, \ldots, p_J|M_k)$ is the joint prior distribution of $p_1, \ldots, p_J$ under M_k. The posterior probability that model M_k is true is given by

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{j=1}^{K} P(D|M_j)P(M_j)}.$$

where $P(M_k)$ is the prior probability of M_k. We specify a discrete uniform distribution for the prior model probability; that is, $P(M_k) = 1/3, k = 1, \ldots, 3$.

We assign the joint prior distribution $f(p_1, \ldots, p_5|M_k)$ to be from Uniform(0, 1), but $(p_1, \ldots, p_5)$ must satisfy the order constraint under each model M_k.

Write a scentific report to address the following points and compute the posterior model probability $P(M_k|D), k = 1, 2, 3$.

1. There are three possible approaches to generate the joint uniform prior of $(p_1, \ldots, p_5)$ under the model constraint:

- Generate $p_1, \dots, p_5$ from the joint uniform prior in a sequential order, e.g., under M_3, simulate $p_1, \dots, p_5$ as follows

$$\begin{aligned} p_1 &\sim \text{Unif}(0, 1), \\ p_2 &\sim \text{Unif}(p_1, 1), \\ p_3 &\sim \text{Unif}(p_2, 1), \\ p_4 &\sim \text{Unif}(0, p_3), \\ p_5 &\sim \text{Unif}(0, p_4). \end{aligned}$$

- Generate $p_1, \dots, p_5$ from the joint uniform prior by first simulating five uniform(0, 1) variates, and then reordering them according to each model constraint. Note that model M_3 does not fully specify the order.
- Derive the prior full conditional distribution, and use the Gibbs sampler to sample from the joint uniform prior distribution $f(p_1, \dots, p_5|M_k)$.

2. Plot the prior density functions of $p_1, \dots, p_5$ under each model (one plot for each model) based on 100,000 prior sample.

3. Derive the posterior full conditional distribution, and use the Gibbs sampler to sample from the joint posterior distribution $f(p_1, \dots, p_5|D, M_k)$. Note that $f(p_1, \dots, p_5|M_k) \propto 1$. How to use the posterior samples to compute $P(D|M_k)$ is a very important question, so that you need to check the literature or google. For example, see the JRSS-B (2008, 70, pp. 589–607) paper "Marginal likelihood estimation via power posteriors" by Friel and Pettitt and the references therein.

4. Use the numerical method to solve the integration in $P(D|M_k)$ directly, yet under the model constraint. You can use *sympy.Symbol* to create the symbol representation of the integrand and calculate the integration using *sympy.integrate*.

5. Compare all results in terms of the posterior model probability $P(M_k|D)(k = 1, 2, 3)$, and comment which method(s) are correct.

********** **END OF PAPER** **********