Stat 8003, Final Exam

Name:

Instructions:

1. This is a take home exam and is due Monday, Dec. 15th, 5:00pm. Please upload your answer(pdf file, tex file and R file) to blackboard;
2. You can consult any references but can not speak with anyone (except for Prof. Zhao) about the exam;
3. Good luck!

Problem 1. Suppose that a single observation X is taken from a uniform density on $[0, \theta]$ and consider testing $H_0 : \theta = 1$ vs $H_1 : \theta = 2$.

(a). For $0 < \alpha < 1$, consider the test that rejects when $X \in [0, \alpha]$. What is the significance level and power?

(b). What is the significance level and power of the test that rejects when $X \in [1 - \alpha, 1]$?

(c). Find another test that has the same significance level and power as the previous one?

(d). What happens if the null and the alternative hypothesis are interchanged $H_0 : \theta = 2$ vs $H_1 : \theta = 1$?

Problem 2. PM2.5. PM2.5 particles are air pollutants with a diameter of 2.5 micrometers or less, small enough to invade even the smallest airways. These particles generally come from activities that burn fossil fuels, such as traffic, smelting, and metal processing. According to Kuenzli, air pollution can cause inflammatory responses both in the body's respiratory tract and in the blood vessels. In the case of the circulatory system, this can eventually lead to thickening of the artery wall and its attendant problems.

The air quality data are obtained at the airdata website, whose primary source is EPA's regulatory air quality system (AQS) repository. The AQS network includes a number of fixed site monitors in each region, each of which measures ambient air pollution levels on a regular basis, either daily, every third day or every sixth day in the case of PM2.5. The following pm2.5 data is the mean of the daily values of different monitoring sites in Boston in 2013.

You can load the data set by using the command:

```
pm2.5 <- read.csv("http://astro.temple.edu/~zhaozhg/Stat8003/data/pm2.5.csv")
```

Answer the following questions:

(a). Can we model the data as a normal distribution? Why or why not?

(b). Model the data according to Gamma distribution $\Gamma(\alpha, \beta)$. Find the MLE of α and β.

(c). Does the Gamma model fit the data well? Test it using the Kolmogorov-Smirnov test at $\alpha = 0.01$ level.

Problem 3. Bulb life. One market monitoring organization would like to compare the life time of two brands of bulbs, Brand A and Brand B. They design the experiment in this way. Let X_i and Y_i be the life time of ith bulb in Brand A and Brand B respectively, which can be approximated by independent random variables with exponential distributions with expectations λ and μ respectively. They pair X_i and Y_i . In the ith experiment, instead of letting both two bulbs burn until they die out, they stop when one of the bulbs burn out, and record the burning time Z_i and indicator W_i of which one burns out. They repeat the experiment n times. Mathematically, Z_i and W_i can be defined as

$$Z_i = min(X_i, Y_i), \text{and } W_i = \begin{cases} 1, & \text{if } Z_i = X_i, \\ 0, & \text{if } Z_i = Y_i, \end{cases}$$

(a). Find closed form expressions for the MLE of λ and μ.

(b). Consider applying the EM algorithm with the complete data take to be $(X_1, Y_1), \cdots, (X_n, Y_n)$. Show that the EM sequence is given by

$$\begin{cases} \lambda^{(k+1)} = \frac{\sum_i w_i z_i + \sum_i (1-w_i)(z_i + \lambda^{(k)})}{n}, \\ \mu^{(k+1)} = \frac{\sum_i (1-w_i) z_i + \sum_i w_i (z_i + \mu^{(k)})}{n}, \end{cases}$$

Hint: in the EM algorithm, the Q function is $El(x, y, \lambda, \mu)|w, z, \lambda^{(k)}, \mu^{(k)}$.

(c). Use the data to find a solution using MLE and EM algorithm with the starting value $\mu^{(0)} = 1$ and $\lambda^{(0)} = 1$.

You can load the data set by using the command:

```
bulb <- read.csv("http://astro.temple.edu/~zhaozhg/Stat8003/data/bulb.csv")
```

Problem 4. Old Faithful Geyser Data. In the faithful data set implemented in R, it contains waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We want to model X_i, the eruption time of the i-th eruption, which can be classified as short eruptions and long eruptions. In class, we model X_i as a mixture of normal distribution with two components and estimate the corresponding parameters using EM algorithm. Next, we will model it according to two-group models.

Let θ_i be an indicator that the i-th eruption is a long eruption. (i.e. $\theta_i = 1$ if the i-th eruption is long and $\theta_i = 0$ otherwise.) Assume the following model,

$$\begin{cases} X_i|\theta_i \overset{\text{ind}}{\sim} (1-\theta_i)N(\mu_s, \sigma_s^2) + \theta_i N(\mu_l, \sigma_l^2), \\ \theta_i \overset{\text{iid}}{\sim} Bernoulli(\pi_1), \\ \pi_1 \sim Unif(0,1), \\ \mu_s, \mu_l \sim N(0, 1000), \\ (\sigma_s^2)^{-1} \sim Gamma(shape = 0.001, rate = 0.001), \\ (\sigma_l^2)^{-1} \sim Gamma(shape = 0.001, rate = 0.001). \end{cases}$$

(a). Derive the Gibbs sampler;

(b). Based on the data, estimate μ_s, μ_l, π_1 based on the posterior median, and construct the corresponding 95% credible intervals;

(c). One researcher believes that there are more short eruptions than the long eruptions. Does your analysis support his statement? Write down your null and alternative hypothesis, calculate an appropriate test statistic and draw your conclusion.

(d). For each eruption, classify it as a long eruption if the corresponding local fdr score $P(\theta_i = 0|X)$ is less than 0.2.

(e). Are the estimations based on the EM algorithm the same as the one calculated above? Discuss the pros and cons of each method.