THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT6011/7611/8305 COMPUTATIONAL STATISTICS (2021 Fall)

Assignment 4, due on November 30

All numerical computation MUST be conducted in Python, and attach the Python code.

1. Consider an integral
$$\int_{-2}^{3} \frac{2\cos(x)+5}{\sqrt{x^4+3}}dx.$$
   (a) Plot the above integrand function in the range of $(-2, 3)$.

   (b) Use the Gaussian Legendre, Chebyshev 1, Chebyshev 2, and Jacobi quadratures to approximate the integral with 20 nodes and weights, respectively.

2. Use the EM algorithm to estimate the parameters in the random intercept model, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$,
$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_i + \epsilon_{ij}, \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2), \\ u_i &\sim N(0, \sigma_u^2). \end{aligned}$$
   The unknown parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_u^2, \sigma_\epsilon^2)^{\mathrm{T}}$.

   (a) Derive the Q-function and the M-step of the EM algorithm.

   (b) Conduct simulations as follows. Set the parameters $\beta_0 = 1$, $\beta_1 = 1$, $\sigma_u = 0.5$, $\sigma_\epsilon = 0.5$, $I = 100$, and $J = 2$. For each dataset, simulate x_{ij} from Uniform$(0, 1)$, simulate ϵ_{ij} and u_i from the corresponding normal distributions, and then obtain y_{ij}. Use the EM algorithm to obtain the parameter estimates based on each simulated dataset. Repeat the simulation process 100 times and present the bias (averaged over 100 simulations) and standard deviation for $\boldsymbol{\theta}$. Comment on your findings.

3. Use the dataset q3.csv, the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ are from a mixture of normal distributions, i.e., $Y_i \sim \sum_{j=1}^k \omega_j f_j(y)$, $i = 1, \ldots, n = 300$, where each f_j is a normal density function $N(\mu_j, \sigma_j^2)$, and ω_j is the mixing probability and $\sum_{j=1}^k \omega_j = 1$. Consider the complete data (y_i, u_i), where the missing data u_i indicates which distribution y_i is from.

   (a) Write out the complete-data likelihood.

   (b) Derive the marginal distribution of y_i.

(c) Visualize the data (e.g., using histograms or clustering) and guess the value of k and normal density functions $N(\mu_j, \sigma_j^2)$.

(d) Suppose that we know $k = 3$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$ $(j = 1, 2, 3)$, and $\omega_1 = \omega_2 = \omega_3 = 1/3$, but μ_j's are unknown. Derive the $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^{(0)})$ function in the E step. In the M step, derive the estimators $\{\mu_j^{(1)}\}$ given the previous step values $\{\mu_j^{(0)}\}$. Use the EM algorithm to estimate μ_j.

(e) Suppose that we know $k = 3$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$ $(j = 1, 2, 3)$, but μ_j and ω_j are unknown. If we treat the u_i's as missing data, derive the $Q(\boldsymbol{\omega}, \boldsymbol{\mu}|\boldsymbol{\omega}^{(0)}, \boldsymbol{\mu}^{(0)})$ function in the E step. In the M step, derive the estimators in a closed form, i.e., the iterative equation between $\{\omega_j^{(1)}$, $\mu_j^{(1)}\}$ and $\{\omega_j^{(0)}$, $\mu_j^{(0)}\}$. Use the EM algorithm to estimate μ_j and ω_j.