THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT6011/7611/6111/3317 COMPUTATIONAL STATISTICS (2016 Fall)

Assignment 4, due on November 28

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

1. In the hidden Markov model, consider tossing a fair die (with probabilities 1/6 for all the six side numbers $\{1, \ldots, 6\}$) and a loaded die (with probabilities $\{1/12, 1/12, 1/12, 1/12, 1/3, 1/3\}$ for all the six numbers $\{1, \ldots, 6\}$). Suppose that we observe a sequence of tosses with numbers

   $\{251326344212463366565535614566523665561326345621443235213263461435421\}$

   determine the underlining coin status: fair or loaded, under each of the four transition matrices.

   (a) If the transition matrix between the fair and loaded dice is

	Fair	Loaded
Fair	0.8	0.3
Loaded	0.2	0.7

   determine the underlining coin status: fair or loaded.

   (b) If the transition matrix between the fair and loaded dice is

	Fair	Loaded
Fair	0.9	0.15
Loaded	0.1	0.85

   determine the underlining coin status: fair or loaded.

   (c) If the transition matrix between the fair and loaded dice is

	Fair	Loaded
Fair	0.95	0.6
Loaded	0.05	0.4

   determine the underlining coin status: fair or loaded.

   (d) If the transition matrix between the fair and loaded dice is

	Fair	Loaded
Fair	0.5	0.5
Loaded	0.5	0.5

   determine the underlining coin status: fair or loaded.

2. Consider the standard normal distribution $X \sim N(0,1)$ with density

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

   (a) Show $\int_{-\infty}^{\infty} f(x)dx = 1$ using polar-coordinates transformation.

   (b) Use Gaussian Hermite quadrature to approximate the integral $\int_{-\infty}^{\infty} f(x)dx$, based on 5, 10, 20, 30 nodes and weights. Present the nodes, weights, and the approximation results.

3. Consider an integral

$$\int_{-2}^{1} \frac{x^9}{\sqrt{x^2+1}} dx.$$

   Present the nodes, weights, and the approximation results.

   (a) Use Gaussian Legendre, Chebyshev 1, Chebyshev 2, and Jacobi quadrature to approximate the integral with 5 nodes and weights.

   (b) Use Gaussian Legendre, Chebyshev 1, Chebyshev 2, and Jacobi quadrature to approximate the integral with 10 nodes and weights.

4. Suppose in a class of 100 students we observe an equal number of individuals with each of the four phenotypes A, B, O, AB, i.e., $n_A = n_B = n_O = n_{AB} = 25$.

   (a) Use the EM notation to derive the step-by-step algorithm to compute the MLEs of the underlying allele frequencies p_A, p_B, and p_O.

   (b) Implement the EM algorithm using Python to estimate p_A, p_B, and p_O.

5. 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. For each dataset, simulate x_{ij} from the standard normal distribution, simulate y_{ij} with $\beta_0 = -1$, $\beta_1 = 1$, $\sigma_u = 0.5$, $\sigma_\epsilon = 1$, $I = 100$, $J = 2$, and use the EM algorithm to obtain the parameter estimates for each simulated dataset. Repeat the simulation process 500 times and present the bias (averaged over 500 simulations) and standard deviation. Comment on your findings.