MATH574 Final Exam, Spring 2018

Note: Total=100

(i) *This exam is a take-home exam. You can use any resources available to you. But you have to do the exam independently without discussing with others.*

(ii) *Please make sure your handwriting is clear and readable. Any solutions that are not clear or recognizable will not be graded!*

(iii) *Please answer all the questions clearly. Important intermediate steps are necessary in order to gain full credits.*

(iv) *For the programming questions, you must answer your questions using words, numbers or figures as asked for. I will not look into your codes to find your answers.*

(v) *Please save all your codes in a separate file and submit it as well. It will be used as proof of your work.*

(vi) *Submit your solutions and codes to Blackboard by 11:59 pm pm on May 3rd, 2018. No late submission is accepted. Emailing me after the due time will be NOT be accepted.*

Problem 0 Course Evaluation Please take 5 minutes of your time to take the course evaluation. Here is how you can do it. Go to Blackboard, then click the "Activity Stream" on the side bar of the left and scroll down to find the Course Evaluation on April 10th. Please do so as soon as possible. The deadline is April 29.

I really appreciate your feedbacks. To express my appreciation, if you paste a screen capture of the confirmation of the submission of your course evaluation of MATH 574, a 5-point extra credits (on top of the 100 points) will be added to your final exam score.

Problem 1 (10 points) Derive a method to sample independent random variables from Exponential distribution ($\exp(\mu)$) based on uniform random number generation, which has a pdf $f(x) = \mu \exp(-\mu x)$, $0 < x < \infty$.

Problem 2 (20 points) Suppose random samples $x_1, x_2, \ldots, x_{100}$ are i.i.d. observations from a mixture of two normal distributions with means μ_1 (which is known to be equal to 0) and μ_2. The two normal distributions have a common variance σ^2. Let α denote the proportion of samples that are from the first normal distribution. Drive an EM approach to find the maximum likelihood estimate of unknown parameter $(\alpha, \mu_2, \sigma^2)$.

Problem 3 (20 points) Straight-line fitting with variation in x and y: suppose we wish to model two variables, x and y, as having an underlying linear relation with added errors. That is, with data $(x,y)_i$, $i = 1, \ldots, n$, we model $\binom{x_i}{y_i} \sim N\left(\binom{u_i}{v_i}, \Sigma\right)$, and $v_i = a + bu_i$. Here the unknown parameters are (a,b), and Σ is known with

$$\Sigma = \left(\begin{array}{cc} \sigma_x^2, & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y, & \sigma_y^2 \end{array} \right),$$

where σ_x^2 and σ_y^2 are the variance of x and y respectively and rho is the correlation between x and y. Assume that the values u_i follow a normal distribution with mean μ and variance τ^2. Write the likelihood of the data given the parameters; you can do this by integrating over $u_1, \ldots, u_n$ or by working with the multivariate normal distribution.

Hints: x_i is not related to (a,b), thus the likelihood $p(x,y|a,b)$ can be factorized into $p(y|a,b)p(x)$. Thus we only need to derive $p(y|a,b)$. Using the conditional normal distribution and hierarchical model you can derive $p(y_i|a,b)$. The conditional normal distribution can be derived from the joint distribution of (x_i, y_i), i.e.,

$$y_i|x_i, v_i, u_i \sim N(v_i + \rho\sigma_y\sigma_x^{-1}(x_i - u_i), \sigma_y^2(1-\rho^2)), \text{ for } i = 1, \ldots, n.$$

The marginal distribution of x_i according to the joint distribution of (x_i, y_i) is $x_i|u_i \sim N(u_i, \sigma_x^2)$.

Problem 4 (20 points) [Normal approximation to posterior under generalized linear models] Suppose we have a random sample $y_i \sim Poisson(\lambda_i)$, $i = 1, 2, \ldots, n$, where $\log(\lambda_i) = (X\beta)_i$ and $X \in \mathbb{R}^{n\times n}$ and $\beta \in \mathbb{R}^p$. Matrix X is known. However, coefficient β is unknown. Suppose n is large. Given a noninformative (i.e. uniform) prior, please derive the approximate multivariate normal distribution to the posterior at the posterior mode. You need to specify the man vector and variance-covariance matrix of this multivariate normal.

Problem 5 (30 points) [Logistic Regression] For a given professional athelet, his or her performance level will tend to increase until midcareer and then deteriorate until retirement. Let y_i denote the number of home runs hit by the professional baseball player Mike Schmidt in n_i at-bats (opportunities) during the ith season. Table 1 gives Schmidt's age, y_i and n_i for all 18 years of his baseball career. The datafile is named schmidt in the LearnBayes package. The home run rates $\{y_i/n_i\}$ are graphed against Schmidt's year in Figure 1. If y_i is assumed to be binomial (n_i, p), where p_i denotes the probability of hitting a home run during the ith

Table 1: Home run hitting data for baseball player Mike Schmidt.

Age	Home Runs	At-Bats	Age	Home Runs	At-Bats
23	1	34	32	31	354
24	18	367	33	35	514
25	36	568	34	40	534
26	38	562	35	36	528
27	38	584	36	33	549
28	38	544	37	37	552
29	21	513	38	35	522
30	45	541	39	12	390
31	48	548	40	6	148

season, then a reasonable model for the $\{p_i\}$ is the logit quadratic model of the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 AGE_i + \beta_2 (AGE_i - mean(AGE))^2,$$

where AGE is Schmidt's age during the ith season.

(a) (10 points) Assume that the regression vector $\beta = (\beta_0, \beta_1, \beta_2)$ has a uniform noninformative prior. Write a short R function to compute the logarithm of the posterior density of β.

(b) (5 points) Use the function "laplace" to find the posterior mode and associated variance-covariance matrix of β.

(c) (5 points) Based on the output from "laplace", use the function "rwmetrop" to simulate 5000 draws from the posterior distribution of β.

(d) (10 points) One would expect the fitted parabola to have a concave down shape where $\beta_2 < 0$. Use the simulation output from part (c) to find the posterior probability that the fitted curve is concave down.

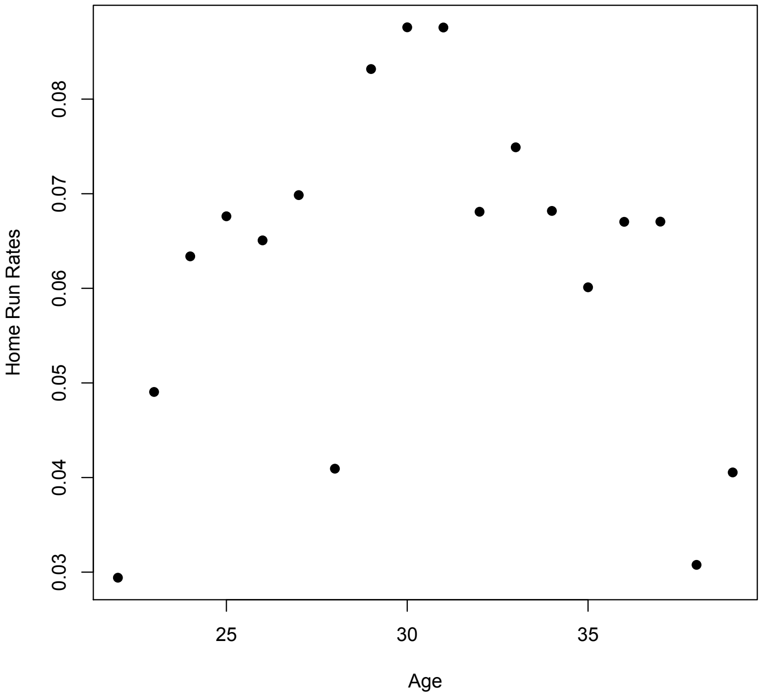

Figure 1: Scatterplot of home run rates HR/AB against age for Mike Schmidt.