# MTH 511A Fall 2019: Numerical Assignment

**Instructions:**

(a) The assignment has two components: (1) a professional quality report with a discussion of the answers, and (2) clean and well commented code.

(b) The assignment is due **November 4th**. The written report is due in class as a hardcopy, and the code is due by 7:00pm on November 4th to be emailed to `dootika.vats@gmail.com`.

(c) Send "roll-no.zip" to the email `dootika.vats@gmail.com`, where "roll-no" is your roll number. So if your roll number is 1234, email the file "1234.zip". The .zip file must contain three files

  (a) A file "em.R" that has the complete code for the first question.

  (b) A file "logistic.R" that has the complete code for the second question.

  (c) A "roll-no.pdf" of your report.

(d) Follow the instructions *very carefully* for how to submit the code for each problem. Make sure your object names in the code is as asked for, otherwise, you will lose points.

(e) Both questions are 50 points each, with 25 points for the code and 25 points for the report.

(f) The code must be well commented (at least as well as the my uploaded codes). Points will be deducted if large junks of code are not well commented.

**Questions:**

1. (50 points) *(Mixture of Gaussians with cross-validation)*

   We have $(X, Y)$ from a bivariate mixture of Gaussians with some unknown number of classes $C$, That is

   $$f(x, y \mid \mu_1, \dots, \mu_C, \mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_C, \pi_1, \dots, \pi_C) = \sum_{c=1}^{C} \pi_c f(x, y \mid \mu_c, \mathbf{\Sigma}_c)\,,$$

   where each $\mu_c \in \mathbb{R}^2$, $\mathbf{\Sigma}_c$ is a $2 \times 2$ covariance matrix, and $\pi_c$ is the mixture probability between 0 and 1.

   (a) Everyone has a different dataset $(X, Y)$, unique to their roll number. Read your dataset

   ```
   dat <- read.table("http://home.iitk.ac.in/~dootika/assets/course/MixG_data/roll-no.txt",
   header = F, col.names = c("X", "Y"))
   ```

   So if your roll number is 12345, your command will be

   ```
   dat <- read.table("http://home.iitk.ac.in/~dootika/assets/course/MixG_data/12345.txt",
   header = F, col.names = c("X", "Y"))
   ```

   Note that, since everyone has a different dataset, no one has the same solution! Your goal is to estimate $\mu = (\mu_1, \dots, \mu_C), \mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_C$, and $\pi = (\pi_1, \dots, \pi_C)$. Assume you know that the number of classes/groups here can only be $C = 2, 3, 4, 5$, or 6.

   (b) Using cross-validation with the negative log-likelihood as a loss function, choose the best model among the 5 models with $C = 2, 3, 4, 5, 6$. Report the negative log-likelihood values from all 5 methods, and the model chosen.

   (c) Using *all* the data, fit 5 models with classes $C = 2, 3, 4, 5, 6$. Calculate the Akaike Information Criterion (AIC) for each model, and choose the model with the smallest AIC. Report AIC values and your model chosen. (This is another model selection criterion, which does not use cross-validation).

   (d) Are the models chosen by these two methods the same? If not, which do you think is a better model, and why?

   (e) After choosing the appropriate number of classes $C$, run the EM algorithm again to find the overall MLE estimates. Report all the MLE estimates. In your code, store the final estimates in a list `mle.est` with three components

   - `mu.est`: a matrix with 2 columns, and your estimated number of rows.
   - `sig.est`: a list of $2 \times 2$ matrices
   - `mix.est`: a vector of mixture probabilities

(f) Plot a scatterplot of data $(X, Y)$, with different point for each group (as estimated by you). Put this in the report.

2. (50 points) *(Bayesian logistic regression with MCMC)*

Consider a Bayesian logistic regression model. For $i = 1, \dots, n$, let $x_i = (1, x_{i2}, \dots, x_{i5})^T$ be the vector of covariates for the $i$th observation and $\beta \in \mathbb{R}^5$ be the corresponding vector of regression coefficients. Suppose response $y_i$ is a realization of $Y_i$ with

$$Y_i \sim \text{Bern}\,(p_i) \quad \text{where} \quad p_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\,.$$

Since this is a Bayesian model, we also assume that $\beta$ has the following prior distribution

$$\beta \sim N_5(0, 100\, I_5)\,.$$

Our goal is to find the posterior distribution and report the posterior mean of $\beta$: $\text{E}(\beta \mid Y)$. Load your unique dataset using

```
dat <- read.table("http://home.iitk.ac.in/~dootika/assets/course/Log_data/roll-no.txt",
header = F)
```

(a) Write down the posterior distribution of $\beta$ in your report (make sure this is correct, since otherwise your whole solution will be incorrect)!

(b) What is a good starting value for the MCMC that samples from the posterior distribution of $\beta$? Report this starting value and your reasons.

(c) Using this starting value, draw samples from the posterior distribution of $\beta$ using a Metropolis-Hastings algorithm. Tune the algorithm so that the acceptance probability is between .23 and .30. Use the following object names in your code:

- The overall acceptance probability is stored in a variable `acc.prob`
- The MCMC output is stored in the object `chain`
- The posterior mean estimates are stored in `beta.est`

(d) Report the autocorrelation plots for *all five* components of $\beta$ and report the estimated marginal density plot of the first and the fifth components of $\beta$.

(e) Finally, report the posterior mean estimates of $\beta$.

At the bottom of your code, add the following lines

```
print(acc.prop)
print(beta.est)
```