MTHM017 Advanced Topics in Statistics
Assignment

Please make sure that the submitted work is your own. This is NOT a group assignment, therefore approaches, solutions shouldn't be discussed with other students, or anyone outside of this organisation. Plagiarism and collusion with other students are examples of academic misconduct and will be reported. More information on academic honesty can be found *here*.

This assessment is AI-supported and permits ethical and responsible use of GenAI tools. You may use GenAI tools to improve the structure of your work, debug your code, or correct your grammar and spelling. You MUST NOT use GenAI tools to help with your modelling, statistical analysis, or write your code from scratch. If markers suspect that you have used AI tools not in a permitted way then you will be required to attend a viva (oral exam) in order to demonstrate your understanding.

Your submission should include not just your answers to the questions, but also all the code and all the relevant output that your code produces. the relevant marks may not be awarded if the code or output is missing from the submission. In your submission you should declare all uses of GenAI tools and reference these appropriately, as well as document the prompts and outputs from these tool. Please refer to the Faculty guidelines on the use of GenAI: see here.

The assignment has two main parts. Part A involves fitting a mixture model to assess reaction times in schizophrenic patients. Part B involves using different methods for classification of data into two groups.

A. Bayesian Inference [65 marks]

In Part A, you must use the functions and syntax covered in the module material. If you use methods outside the scope of the module in addition to these, the source should be clearly cited (here the source cannot be GenAI), and the underlying theory briefly explained. You should also ensure that you read the instructions carefully, as failure to follow them could result in zero marks being awarded for certain parts of the questions.

In Part A we will fit a finite mixture model using the `rtimes` dataset, which contains the reaction times of 17 people (11 non-schizophrenics and 6 schizophrenics, stored in this order) in a psychological experiment. Each person's reaction time was measured 30 times.

1. *[6 marks]* Read in the data, then for each person produce a histogram of that given person's reaction times. The range of the `x` axis should be the same on each histogram. Visually compare the reaction time distributions of schizophrenic and non-schizophrenic individuals. What differences/similarities can you observe? Reference the histograms of specific individuals to support your conclusions.

It is suggested that schizophrenics suffer from attention deficit on some trials, as well as a general motor reflex retardation. Motor reflex retardation affects the response time of all trials, while attention deficit only affects some of the responses. To address this theory we will fit a model, where the response times of non-schizophrenics are described by a normal random-effects model, and the response times of schizophrenic individuals are modeled as a two-component mixture model.

To reflect the attention deficit, let y_{ij} denote the **logarithm** of the jth measured reaction time of person i. Then:

- For the responses of the ith *non-schizophrenic* person ($i = 1, 2, \dots, 11$) we have

  $$y_{ij} \sim N(\alpha_i, \sigma_y^2), \quad i = 1, 2, \dots, 11, \quad j = 1, 2, \dots, 30.$$

  That is, the responses are normally distributed with person-specific mean α_i and some common variance σ_y^2.

- For the responses of the ith *schizophrenic* individual ($i = 12, 13, \dots, 17$), with probability $(1 - \lambda)$ there is no delay, and the response is normally distributed with mean α_i and variance σ_y^2; and with probability λ the response is delayed so that the observations have mean $\alpha_i + \tau$ and variance σ_y^2. That is

  $$\begin{aligned} &y_{ij} \sim N(\alpha_i + \tau z_{ij}, \sigma_y^2), \\ &z_{ij} \sim \text{Bernoulli}(\lambda), \qquad i = 12, 13, \dots, 17; \quad j = 1, 2, \dots, 30. \end{aligned}$$

  Note that in the above model z_{ij} is an indicator function that takes the value 1 whenever the response is delayed, and the value 0 otherwise. Furthermore, τ is the amount of time by which the response is delayed. To ensure that the model is identifiable we will **restrict τ to be positive**.
  The two cases (schizophrenic and non-schizophrenic) could be brought to the same form by adding indicator variables z_{ij} to the non-schizophrenic part of the model. However in the non-schizophrenic case these variables will always take the value 0!

The magnitude of the schizophrenics' motor retardation is captured by the distribution of the α_i parameters. In particular,

- For *non-schizophrenic* individuals we assume that α_i follows a normal distribution with mean μ and variance σ_α^2, that is

  $$\alpha_i \sim N(\mu, \sigma_\alpha^2), \quad i = 1, 2, \dots, 11.$$

- For the *schizophrenics* we assume that the mean of α_i is $\mu + \beta$, while the variance remains σ_α^2. That is

  $$\alpha_i \sim N(\mu + \beta, \sigma_\alpha^2), \quad i = 12, 13, \dots, 17.$$

2. *[5 marks]* The above model uses the logarithm of measured reaction times. Explain why taking the logarithm is necessary here (referencing the relevant output), then perform the transformation yourself. For each person compute the standard deviation of the log transformed reaction times of that individual.

3. *[5 marks]* List the parameters of the model and assign non-informative uniform prior distributions to each parameter, paying attention to the values these parameters are allowed take.

4. *[13 marks]* Code up the above model in JAGS using functions covered in the module. Fit the model with 10000 iterations, discarding the first 5000 as burn-in. Make sure you set the model up in a way that demonstrates your full understanding of the Bayesian model fitting process as taught in the module.
   Note that part of the JAGS model was written for you. To write your JAGS model fill in the gaps of the draft model definition below by adding i) the likelihood component that describes the reaction time of schizophrenics, and ii) the prior distributions of all the model parameters. You shouldn't modify the part that's already given.

   ```
   jags.model <- function(){
     # reaction time of non-schizophrenics
     for(i in 1:11){
       for(j in 1:30){
         y.ns[i,j] ~ dnorm(alpha.ns[i], p.y)
       }
       alpha.ns[i] ~ dnorm(mu,p.alpha)
     }
     # reaction time of schizophrenics
     ...
     # priors
     ...
   }
   ```

5. *[8 marks]* Using only methods covered in the module, investigate whether the MCMC chains have converged (convergence should be checked for all the nodes where this is appropriate). Include all the relevant evidence that supports your conclusions.

6. *[10 marks]* The primary interest to psychologists lies in the parameters β, λ and τ. Plot the posterior distributions of these three parameters, then produce numerical summaries of the distributions. Check if you have enough samples for posterior inference.
   Remembering that the response time was modeled on the log scale (and therefore both τ and β are on the log scale), give the median and a 95% posterior interval for each of these parameters on their original scale. Based on these estimates what conclusions can you make about the reaction times of schizophrenics compared to non-schizophrenics?

7. *[15 marks]* Next we will use prediction to check the fit of the model. Follow the steps below to assess how well the model can explain the variability in the data.

   (a) Edit your previous model definition so that it predicts 30 additional (log) response time measurements $\tilde{y}_{ij}$, $j = 1, 2, \ldots, 30$, for each schizophrenic individual $i = 12, 13, \ldots, 17$ in the study. Note that the prediction of $\tilde{y}_{ij}$ should use the posterior of α_i.

   (b) Then add futher nodes to your model to i) find the standard deviation of the 30 predicted measurements for each individual, and to ii) get the minimum and maximum values of these 6 standard deviation values.
   That is, if for individual i the simulated response times are

   $$\tilde{\boldsymbol{y}}_i = (\tilde{y}_{i1}, \tilde{y}_{i2}, \ldots, \tilde{y}_{i30}), \quad i = 1, 2, \ldots, 6,$$

   then the model should first compute the standard deviations

   $$sd_i = sd(\tilde{\boldsymbol{y}}_i), \quad i = 1, 2, \ldots, 6,$$

   then find the smallest and largest of these six values,

   $$\begin{aligned} S_{min} &= min(sd_1, sd_2, \ldots, sd_6), \\ S_{max} &= max(sd_1, sd_2, \ldots, sd_6). \end{aligned}$$

   (c) Fit your edited model with 6000 iterations, discarding 5000 as burnin. In the model fitting step you should make sure that you demonstrate your understanding of all the steps of the JAGS model fitting; but you may skip convergence checking.

   (d) Extract the minimum and maximum standard deviation values from the fitted model, and produce a scatterplot of the (S_{min}, S_{max}) pairs. (Note that each iteration of the model fit will produce a minimum-maximum pair). Find the minimum and maximum of the raw standard deviation estimates obtained in Question 2, and add an additional point to your scatterplot showing this raw minimum-maximum pair.

   Based on the scatterplot, would you say that the model can accurately explain the variation in the within-person response time variance?

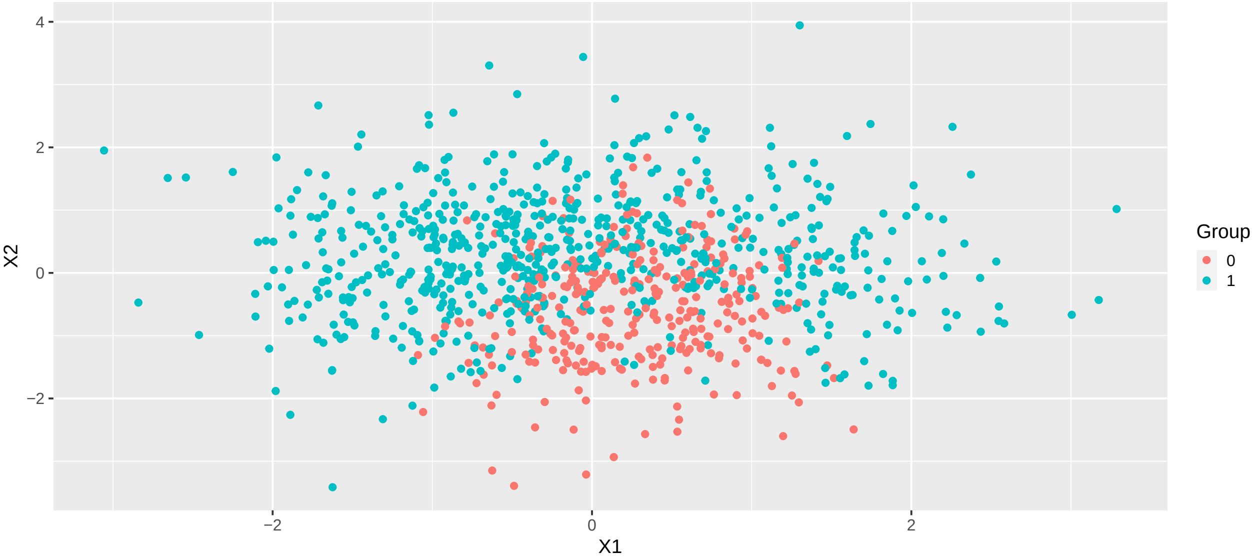
8. *[3 marks]* These marks will be automatically awarded if you used the same functions in the JAGS model fitting as the module's problem sheets. However, if you used a different syntax or different functions, explain what these differences are and cite the (published!) source you used for the model fitting.

B. Classification [35 marks]

In Part B, you should use the theory covered in the module material. If you use methods/explanation outside the scope of the module in addition to these, the source should be clearly cited (here the source cannot be GenAI), and the underlying theory briefly explained.

You should also ensure that you read the instructions carefully, as failure to follow them could result in zero marks being awarded for certain parts of the questions.

The following figure shows the information in the dataset `Classification.csv` - it shows two different groups, plotted against two explanatory variables. This is simulated data - the groupings are determined by a (known, but not to you!) function of X1 and X2 with added noise/random error. The aim is to find a suitable method for classifying the 1000 datapoints into the two groups from a selection of possible approaches.



1. *[5 marks]* Create meaningful summaries of the two groups in terms of the variables X_1 and X_2. Describe your findings. Considering the plot showing the observations and the numerical summaries, which of the following classification methods do you think are suitable for classifying this data and why?

   a. Linear discriminant analysis.

   b. Quadratic discriminant analysis.

   c. K-nearest neighbour classification.

   d. Support vector machines.

   e. Random forests.

2. *[1 marks]* Select 75% of the data to act as a training set, with the remaining 25% for testing/evaluation.

3. *[23 marks]* Choose **four of the methods** listed in Question 1 that are suitable to classify the data. Perform classification using these methods. In each case, briefly describe the theory behind how the classification method classifies the data (using your own words, and referring to the module material), present the results of an evaluation of the method (highlighting different aspects of the model performance) and describe your findings. Make sure that in each case you give a detailed description of the model performance. Where appropriate optimise the (hyper)parameters of the method. Note, if you fit all five models, only the first four will be considered for marking.

4. *[4 marks]* Compare the results from your chosen four approaches and select the best method(s) for classification while considering different modelling objectives. Explain your reasoning.

5. *[2 marks]* The file `ClassificationTrue.csv` contains the true classifications, based on the function of X1 and X2 without the noise. Evaluate how your four chosen methods from Questions 3 compare to the truth (in each case use the previously selected optimal value of the parameters). Do(es) your choice(s) from Question 4 still perform best in this case?

Total for paper = 100 marks

The deadline for submission is Noon (12pm), 14th March. Note that late submissions will be penalised.

You should submit a pdf that contains your answers, code (and all the relevant output/plots!) to the questions via ELE. In Part A you should use the R programming language, but in Part B you can choose to use R or Python (or both).