Advanced Topics in Statisics Assignment

The assingment has three main parts. Part A involves fitting an auto-regressive process to time series data model using the BUGS language and assessing the effect of using different model structures on the estimation of missing data. Part B involves using different methods for classification of data into two groups. Part C involves producing a narrated power point presentation based on question 3 of part B.

Part A gives 50% of your final marks, Part B gives 30% of your final marks and Part C gives 20% of your final marks. [Assignment: 160 marks in total]

A. Bayesian Inference [80 marks]

The dataset contains measurements of particulate matter (PM10) air pollution in London (measured at the Heathrow and Haringey sites) for 2000 to 2004. The data can be found in London_Pollution.csv.

1. *[4 marks]* Summarise the two sets of data and calculate the number of missing data points for each monitoring location, by year. Comment on whether the patterns of missingness have changed over time.
2. *[3 marks]* Plot the PM10 measurements against time for the two sites, highlighting (showing clearly) the periods of missing data.
3. *[5 marks]* The locations in Eastings and Northings of the two locations are Heathrow: (508399, 176744); and Haringey: (533890, 190638). Plot these two monitor locations on a map of London and comment on any difference you found in the summaries of the data in the context of the geographical location of the monitoring sites. The necessary shapefiles are on the ELE page of the course.

Considering the Heathrow data, there is missing data. We are going to fit a model that allows us to estimate these missing data by treating them as model parameters that will be estimated (and we find posterior distributions for them). As we have time series data, we are going to use the fact that day-to-day measurements will be correlated, i.e. today's measurement will correlate with yesterday's.

A random walk process of order 1, RW(1), is defined at time t as

$$\begin{aligned} Y_t - Y_{t-1} &= w_t \\ Y_t &= Y_{t-1} + w_t \end{aligned}$$

Where w_t are a set of realisations of random (or white) noise, e.g. $w_t \sim N(0, \sigma_w^2)$. Note the first line refers to the differences in the values at consecutive time points being white noise.

We are interested in fitting a random walk model to the Heathrow data (*Heathrow*). The model will be of the following form:

$$\begin{aligned} Heathrow_t &\sim N(Y_t, \sigma_v^2) \\ Y_t &\sim N(Y_{t-1}, \sigma_w^2) \end{aligned}$$

Where σ_w^2 is the variance of the white noise process associated to the random walk. We then make noisy measurements of this random walk process, thus $Heathrow_t$, the measurement we have at time t, equals to the true value of the underlying process Y_t plus some measurement error. In the formula above, σ_v^2 is the variance of this measurement error.

4. *[16 marks]* Code this model in JAGS to analyse the Heathrow data from 1st January 2000 to 31st December 2003 (NOTE the end year). Hint: due to the nature of the model you will have to explicitly specify a value for Y_1 in the model (i.e. for the first time point as Y_0 doesn't exist). One suggestion might be $mu \sim dnorm(0, 0.001)$. Run the model for 10,000 iterations, with 2 chains, discarding the first 5,000 as 'burn-in'. Produce trace plots for the chains and summaries for the fitted parameters (including the missing data). Hint: You should initialise both chains. One suggestion might be using the mean and median to initialise the missing values of *Heathrow*, and using random uniforms (with a narrow interval centred around say 20) to initialise Y.

5. *[3 marks]* Comment on whether the chains for all the parameters have converged.

6. *[5 marks]* Extract the posterior means and 95% credible intervals for $\widehat{Y}_t$, and plot them against time, together with the original data (the measurements). Comment on the width of the credible interval during the periods of missing data.

An alternative model is a random walk process of order 2, RW(2). This assumes that the 'differences between differences' is white noise and is defined at time t as

$$\begin{aligned}(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) &= w_t \\ Y_t &= 2Y_{t-1} - Y_{t-2} + w_t\end{aligned}$$

Where again w_t are a set of realisations of random (or white) noise, e.g. $w_t \sim N(0, \sigma_w^2)$.

That is now we are interested in fitting a random walk model of order 2 to the Heathrow data. The model will be of the following form:

$$\begin{aligned}Heathrow_t &\sim N(Y_t, \sigma_v^2) \\ Y_t &\sim N(2Y_{t-1} - Y_{t-2}, \sigma_w^2)\end{aligned}$$

Again, σ_w^2 is the variance of the white noise process, and σ_v^2 is the variance of the measurement error.

7. *[12 marks]* Code this RW(2) model in JAGS to analyse the Heathrow data from 1st January 2000 to 31st December 2003 (NOTE the end year). Run the model for 10,000 iterations, discarding the first 5,000 as 'burn-in'. Produce trace plots for the chains and summaries for the fitted parameters (including the missing data). Comment on the differences between the smoothing effects of the two models. For this you might find it useful to plot the outcome for the first quarter of 2000 separately (for both models). Note that getting this model to converge might be quite tricky. Instead of spending much time trying to get it to converge, you should instead try to explain why we might see lack of convergence here.

8. *[8 marks]* Use both of your models to predict the measurements of PM10 at Heathrow for the first week of 2004.

9. *[6 marks]* For both models, plot the predicted values of PM10 for the first week of 2004, along with the actual measurements, against time. By calculating appropriate measures of comparison, comment on how good you think the models are at forecasting. Hint: you may want to re-run the model with an extra line to calculate the root mean squared prediction error $\sqrt{\sum_{t=1}^{n} \frac{(\hat{Y}_t - Y_t)^2}{n}}$, noting that this value will also have a posterior distribution as it is a function of the predicted values (that are treated as unknown paramaters that need to be estimated).

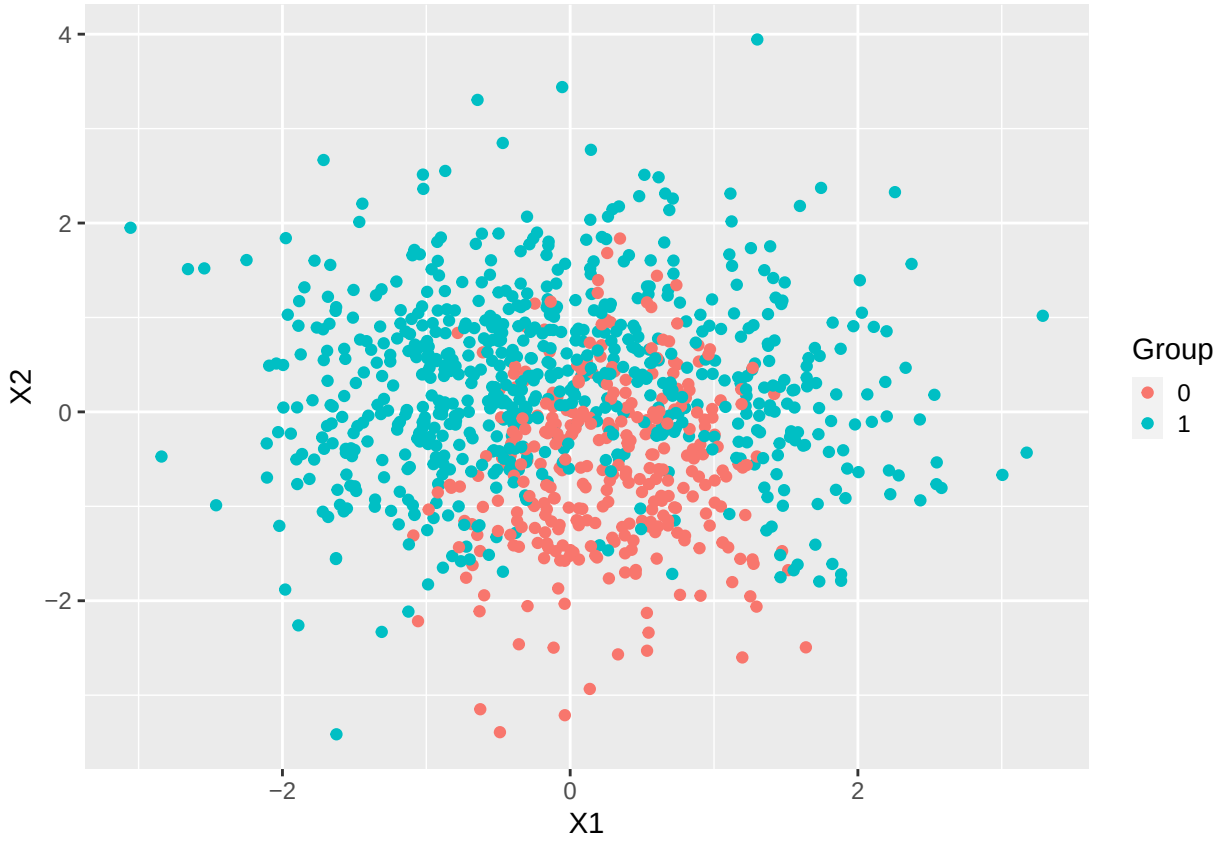
We are now going to repeat this analysis for the Haringey site.

10. *[8 marks]* Fit the RW(1) and RW(2) models in JAGS to the Haringey data for 2000 to 2003. Use non-informative priors. Comment on how well the chains have converged and how well both models fit the data.

11. *[10 marks]* Now re-run these analyses using informative priors, using what you have learnt from fitting the Heathrow model. By comparing the results (e.g. summmaries of the posterior distributions, convergence etc), comment on any effect that using different priors has (or has not had) on the results.

B. Classification [48 marks]

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. *[3 marks]* Summarise the two groups in terms of the variables X_1 and X_2. Describe your findings.
2. *[2 marks]* Select 80% of the data to act as a training set, with the remaining 20% for testing/evaluation.
3. Perform classification using the following methods. In each case, briefly describe how the method works, present the results of an evaluation of the method and describe your findings. Where appropriate optimise the parameters of the method (e.g. by using ROC curve, cross validation). In each case describe carefully how the optimisation method works.
   (a) *[5 marks]* Linear discriminant analysis.
   (b) *[5 marks]* Quadratic discriminant analysis.
   (c) *[8 marks]* Logistic regression.
   (d) *[10 marks]* Support vector machines.

(e) *[8 marks]* K-nearest neighbour regression.

4. *[3 marks]* Compare the results from these five approaches and select what you think is the best method for classification in this case, explaining your reasoning.

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

C. Presentation [32 marks]

The presentation is based on PartB/Q3 only. You should submit a narrated power-point presentation that should be 5 minutes long, and you should aim for 5 slides in total (this could mean 1 slide on each method).

In this you should explain what the problem is, how you approached it, and what your findings are.

You should pay attention to the clarity/pace/coherency of the delivery, the style/information-balance on the slides, clear description of methodology and time management.

The deadline for submission is Noon (12pm), 22nd February. Note that late submissions will be penalised.

You should submit the narrated power point presentations and a pdf that will contain your answers to the questions via ELE. Your `R` code should be included in your report as an Appendix.