UNIVERSITY OF TECHNOLOGY SYDNEY

School of Mathematical and Physical Sciences

37458 Advanced Bayesian Methods

ASSIGNMENT 8

Due time and date: 10:05am, Friday 7th June, 2019.

Submission method and location: Hand to Professor Wand at start of class in Room CB5C.01.011.

NOTE: For the benefit of participants requiring assistance with this assignment, a help session will be held at 3pm-4pm on Wednesday 5th June 2019 in Room CB07.06.006.

1. This question involves model selection for regression analysis of a data set concerning production of cheddar cheese. According to a source: "As cheese ages, various chemical processes take place that determine the taste of the final product. This data set contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample. The variables "acetic" and "H2S" are the natural logarithm of the concentration of acetic acid and hydrogen sulphide respectively. The variable "lactic" has not been transformed."

   (a) Download the file `cheese.txt` from the subject web-site.

   (`http://matt-wand.utsacademics.info/37458.html`).

   (b) Issue the following commands to visualise the cheese data:

   ```
   cheese <- read.table("cheese.txt",header=TRUE)
   pairs(cheese)
   ```

   (c) The remainder of this question is concerned with selection of regression models that predict or explain taste in terms of the three chemical variables. This will done using the methodology in the recent research article: *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.* by A. Vehtari, A. Gelman and J. Gabry (2017). This requires installation of the corresponding R package named loo. Ensuring that you have an Internet connection issue the command:

   ```
   install.packages("loo")
   ```

   (d) Download the script `cheeseLOOana.Rs` from the subject web-site. This script uses the rstan and loo packages to choose among Bayesian linear regression involving all subsets of the predictors. To avoid scaling issues, it works with the following standardised variables:

$$\begin{aligned} y &= \text{standardised taste}, \\ x_1 &= \text{standardised acetic}, \\ x_2 &= \text{standardised H2S}, \\ x_3 &= \text{standardised lactic}, \end{aligned}$$

There are 8 possible linear regression models containing predictors as follows:

$$
\begin{aligned}
&1\\
&1,\ x_1,\\
&1,\ x_2,\\
&1,\ x_3,\\
&1,\ x_1,\ x_2\\
&1,\ x_1,\ x_3\\
&1,\ x_2,\ x_3\\
&1,\ x_1,\ x_2,\ x_3
\end{aligned}
$$

For example, the last (and biggest) model is

$$y_i|\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \beta_3\, x_{3i}, \sigma^2)$$

Issue the command:

`source("cheeseLOOana.Rs")` to loop through each of these 8 models and compute the leave-one-out deviance criterion for each model (defined such a way that "lower" means "better"). Which model is selected according to this criterion?

(e) Open the file `cheeseLOOana.Rs` in an editor and look at the Stan code inside the `generated quantities` section to see how the log-likelihood of the model must be specified so that the loo package can obtain the leave-one-out deviance criterion values.

(f) A possible limitation of the models considered in by `cheeseLOOana.Rs` is that interactions between the predictors are not considered. Now consider the following additional 10 models in which pairwise interactions that accompany main effects already in the model are considered:

$$
\begin{aligned}
&1,\ x_1,\ x_2,\ x_1x_2\\
&1,\ x_1,\ x_3,\ x_1x_3\\
&1,\ x_2,\ x_3,\ x_2x_3\\
&1,\ x_1,\ x_2,\ x_3,\ x_1x_2\\
&1,\ x_1,\ x_2,\ x_3,\ x_1x_3\\
&1,\ x_1,\ x_2,\ x_3,\ x_2x_3\\
&1,\ x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3\\
&1,\ x_1,\ x_2,\ x_3,\ x_1x_2,\ x_2x_3\\
&1,\ x_1,\ x_2,\ x_3,\ x_1x_3,\ x_2x_3\\
&1,\ x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3,\ x_2x_3
\end{aligned}
$$

For example, the last (and biggest) interaction model is

$$y_i|\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0+\beta_1\, x_{1i}+\beta_2\, x_{2i}+\beta_3\, x_{3i}+\beta_{12}\, x_{1i}x_{2i}+\beta_{13}\, x_{1i}x_{3i}+\beta_{23}\, x_{2i}x_{3i}, \sigma^2)$$

and allows for all pairwise interactions of the three predictors.

(g) Copy `cheeseLOOana.Rs` to a new file named `cheeseLOOanaWithInteracs.Rs`.

(h) Open `cheeseLOOanaWithInteracs.Rs` and change the line

`numModels <- 8` to `numModels <- 18`

to reflect the fact that the model space is being expanded to include interactions.

(i) Inside the loop that starts with `for (iModel in 1:numModels)` add new lines for the additional models. The first such line is

`if (iModel==9) X <- cbind(1,x1,x2,x1*x2)`. Do any of the new interaction models improve upon the first set of models in terms of lowering the leave-one-out deviance criterion?

2. In this question the code from the previous question is used to perform model selection for a different data set. As with the cheddar cheese data it has a continuous response variable and three potential predictors. The data set corresponds to 21 days of operation of a plant for the oxidation of ammonia to nitric oxide, and is part of base R (so does not have to be loaded in).

(a) Enter the following R commands to quickly visualise the data and then find out more about the data:

```
pairs(stackloss)
help(stackloss)
```

(b) Copy the file `cheeseLOOanaWithInteracs.Rs` to a new file named

`stacklossLOOanaWithInteracs.Rs`. Then modify the code so that the same 18 models from the previous question are fitted to the stackloss data with

$$\begin{aligned} y &= \text{standardised stackloss}, \\ x_1 &= \text{standardised air flow}, \\ x_2 &= \text{standardised water temperature}, \\ x_3 &= \text{standardised acid concentration} \end{aligned}$$

and compared using the leave-one-out deviance criterion provided by the loo package in R. Note that the four variables have can be accessed from the `stackloss` as follows: `stackloss$stack.loss`, `stackloss$Air.Flow`, `stackloss$Water.Temp` and `stackloss$Acid.Conc`

Write down the selected model.

(c) What is the interpretation of the selected model with regards to how the predictors impact stack loss?

3. In Assignments 5 and 6 three Bayesian mixed models, labelled 1, 2 and 3, were fit to data from an experiment involving 30 young rats. The residual plots suggested that model 3 was the best fit to the data, but this was subjective. In this question the leave-one-out deviance criterion provided by the loo package in R is used to choose among the three models. Since the Stan implementation of the models differs quite a bit between Model 1 and Models 2 and 3, separate scripts that obtain leave-one-out deviance criterion values have been prepared.

(a) Download the following three files from the subject web-site:

- `ratsModel1LOOana.Rs`

- `ratsModel2LOOana.Rs`
- `ratsModel3LOOana.Rs`

(b) Open `ratsModel1LOOana.Rs` in an editor. In the Stan code for model specification, study the `generated quantities` component to see how creation of the `log_lik` quantity, for integration with the loo package, is coded.

(c) Source each of `ratsModel1LOOana.Rs`, `ratsModel2LOOana.Rs` and `ratsModel3LOOana.Rs` and make a table that compares the leave-one-out deviance criterion values.

(d) Is the embellishment of third model with random intercepts and slopes and a quadratic fixed effect term justified?

4. This question involves visualisation of a three-dimensional surface fit to the Wolfcamp Aquifer data from Assignment 7, and assumes that the R package rgl is installed in your R environment.

(a) Download the script `wolfcampPlaneFit.Rs` from the subject web-site. The files `wolfcamp.txt`, `rglSetup.Rs` and `pointsInPoly.r` (from a previous assignment) also are needed for this part. In an R session, issue the command:

`source("wolfcampPlaneFit.Rs")`

The script fits a plane to a transformed version of the data. The planar fit can be visualised by spinning the graphical object using the mouse or touchpad on your computer. The fitting is done using ordinary (non-Bayesian) least squares but a fit obtained using a Bayesian approach and Stan would be very similar.

(b) The Bayesian plane model has the form:

$$y_i|\beta_0,\beta_1,\beta_2,\sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0+\beta_1\,x_{1i}+\beta_2\,x_{2i},\sigma^2), \quad 1\le i\le 85.$$

where

$$\begin{aligned} y &= \text{transformed piezometric head},\\ x_1 &= \text{transformed x-coordinate},\\ x_2 &= \text{transformed y-coordinate}. \end{aligned}$$

This can be re-expressed in matrix notation as:

$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}) \tag{1}$$

where, for example, $\boldsymbol{X}=[1\ x_{1i}\ x_{2i}]_{1\le i\le 85}$ is the 85×3 matrix containing a column of ones and the x_1 and x_2 data .

(c) Download the script wolfcampBivPenSplLOOana.Rs from the subject web-site and issue the command:

`source("wolfcampBivPenSplLOOana.Rs")` and record the leave-one-out deviance criterion value.

(d) The script from the previous part of this question fits a bivariate penalized spline model with matrix algebraic representation:

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2\boldsymbol{I}) \tag{2}$$

where $\boldsymbol{Z}$ contains bivariate spline functions of the (x_1, x_2) data. However, the simple planar fit seems reasonable and there is the question of whether the bivariate penalized spline fit overs significant prediction improvement. Copy `wolfcampBivPenSplLOOana.Rs` to a new file named `wolfcampPlaneLOOana.Rs`. Then edit `wolfcampPlaneLOOana.Rs` so that it fits model (1) rather than model (2) and obtain the leave-one-out deviance criterion for the reduced model.

Hint: Conversion of `wolfcampBivPenSplLOOana.Rs` to one that fits a plane essentially involves removal of the parts of the code connected with the $\boldsymbol{Zu}$ term.

(e) Does the bivariate Bayesian penalized spline model offer a better fit compared to the Bayesian planar model in terms of the leave-one-out deviance criterion produced by the loo package?