UNIVERSITY OF TECHNOLOGY SYDNEY

School of Mathematical and Physical Sciences

37457 Advanced Bayesian Methods

ASSIGNMENT 7

Due time and date: 10:05am, Wednesday 25th October, 2023.

Submission method: Hand to Professor Wand at start of Week 11 class.

NOTE: For the benefit of participants requiring assistance with this assignment, a help session will be held at 2pm-3pm on Tuesday 24th October 2023 in Room 006, Level 6, Building 7.

1. This question is concerned with matrix notation typically used in the linear mixed models literature. Suppose that small longitudinal study has $m = 3$ groups and $n_1 = 2$, $n_2 = 3$, $n_3 = 2$ measurements on each group, respectively. The response vector is:

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{bmatrix}.$$

The fixed effects and random effects vectors are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

The fixed effects and random effects design matrices are:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{23} \\ 1 & x_{31} \\ 1 & x_{32} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1}$$

This question is concerned with showing why the design matrices have these particular forms when the Bayesian linear mixed model is expressed using matrix notation. Starting with the matrix algebraic expression:

$$E(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$$

obtain expression for each of the $E(y_{ij}|\boldsymbol{\beta}, \boldsymbol{u})$ ($1 \leq i \leq 3, 1 \leq j \leq n_i$). Be sure to show all steps of the working.

Hint: the answer for the first one is

$$E(y_{11}|\boldsymbol{\beta}, \boldsymbol{u}) = \beta_0 + \beta_1 \, x_{11} + u_1.$$

If correct, your answer verifies (for the special case of $n_1 = 2$, $n_2 = 3$, $n_3 = 2$) that the first part of the Bayesian random intercepts model:

$$y_{ij}|\beta_0, \beta_1, u_i, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\beta_0 + u_i + \beta_1 \, x_{ij}, \sigma_\varepsilon^2), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i$$

can be written succinctly using matrix notation as:

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

with matrices $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ as defined above.

This *general form* of the linear mixed model is useful for various theoretical calculations.

2. This question is concerned with using a *generalized additive mixed model* analysis to compare spinal bone mineral density of adolescent American females among four ethnic groups (Asian, Black, Hispanic and White). Generalized additive mixed models extend both linear mixed models and generalized additive models by accounting for both grouping structure and non-linearity. **NOTE: Most of this question involves following instructions, rather than producing solutions to be handed in. The single hand-in part is highlighted below.**

   - Download the files `femSBMD.txt` and `femSBMDview.R` from the subject website

     (`http://matt-wand.utsacademics.info/37457.html`).
   - Start an R session and issue the command `source("femSBMDview.R")` to visualise the data. Each set of connected points corresponds to repeated measurements spinal bone mineral density measurements of a particular adolescent. It is apparent that age has a non-linear effect on mean spinal bone mineral density.
   - A generalized additive mixed model for these data is:

$$\begin{aligned} y_{ij}|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \boldsymbol{u}_{\text{grp}}, \boldsymbol{u}_{\text{spl}}, \sigma_\varepsilon^2 \\ \overset{\text{ind.}}{\sim} N\Big(\beta_0 + f(x_{1ij}) + \beta_2 \, x_{2i} + \beta_3 \, x_{3i} + \beta_4 \, x_{4i} + u_{\text{grp},i}, \sigma_\varepsilon^2\Big) \end{aligned}$$

     where $f(x) = \beta_1 x + \displaystyle\sum_{k=1}^{K} u_{\text{spl},k} z_k(x)$ where $z_1, \ldots, z_K$ are suitable spline basis functions and the $u_{\text{grp},i} \overset{\text{ind.}}{\sim} N(0, \sigma_{\text{grp}}^2)$ and $u_{\text{spl},k} \overset{\text{ind.}}{\sim} N(0, \sigma_{\text{spl}}^2)$ are random effects that allow for the grouping structure of the responses and the non-linearity of age (which is the x_1 variable). The y variable is spinal both mineral density, the x_2 variable is an indicator that the adolescent has Black ethnicity, the x_3 variable is an indicator that the adolescent has Hispanic ethnicity and the x_4 variable is an indicator that the adolescent has White ethnicity (Asian ethnicity is the reference group, so does not have an indicator variable).

- Using results from Question 1, the first part of the model can be written more succinctly in the form

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}_{\text{grp}}, \boldsymbol{u}_{\text{spl}}, \sigma_\varepsilon^2 \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_{\text{grp}}\boldsymbol{u}_{\text{grp}} + \boldsymbol{Z}_{\text{spl}}\boldsymbol{u}_{\text{spl}}, \sigma_\varepsilon^2 \boldsymbol{I})$$

  where $\boldsymbol{Z}_{\text{grp}}$ is a matrix analogous to the matrix $\boldsymbol{Z}$ in equation (1).
- Issue the following R commands to create the $\boldsymbol{Z}_{\text{grp}}$ matrix for the adolescent female spinal bone mineral density data:

```
femSBMD <- read.table("femSBMD.txt",header=TRUE)
idnum <- femSBMD$idnum
print(idnum)
Zgrp <- 1*(outer(idnum,1:423,"-")==0)
print(Zgrp[1:10,1:5])
```

- The last command shows the top left-hand corner of $\boldsymbol{Z}_{\text{grp}}$ for the current example. In this case $\boldsymbol{Z}_{\text{grp}}$ is quite a large matrix, of dimension 1003×423. Issue the command:

  `image(t(Zgrp[34:1,1:10]))` to visualise a larger portion of top left-hand corner of $\boldsymbol{Z}_{\text{grp}}$. The dark shade indicates positions of the 1s and the lighter shade indicates positions of the 0s.
- Now issue the

  `image(t(Zgrp[dim(Zgrp)[1]:1,]))` to visualise the full matrix.
- Issue the following commands the obtain the dimension and sum of all entries in $\boldsymbol{Z}_{\text{grp}}$:

  `dim(Zgrp) ; sum(Zgrp)`

  THE ONLY HAND-IN PART OF QUESTION 2:

  Based on these results, what percentage of the entries of $\boldsymbol{Z}_{\text{grp}}$ are zero?
- Download the file `femSBMDviaStanBruteForce.R` from the subject web-site. In this file the model specification in the Stan language is done using the code:

  `y ~ normal(X*beta + Zgrp*uGrp + Zspl*uSpl,sigmaEps)`

  where the matrices $\boldsymbol{Z}_{\text{grp}}$ and $\boldsymbol{Z}_{\text{spl}}$ are set up and passed into the Stan inference engine. Issue the command

  `source("femSBMDviaStanBruteForce.R")` This fits the above Bayesian generalized additive mixed model to the data and produces summary results. Key features of the output are:
  - the Black ethnicity group has a statistically significant elevation above the Asian ethnicity group in terms of mean spinal bone mineral density, after accounting for age and within-subject correlation,
  - age has a pronounced non-linear effect on mean spinal bone mineral density.
- Issue the commands:

```
idnumSmall <- c(1,1,2,2,2,2,2,3,3,3,4,4,4,4)
uA <- c("elephant","frog","goose","hyena")
uA[idnumSmall]
uB <- c(8.3,3.4,-1.9,5.6)
uB[idnumSmall]
```

These examples show how, according to an array indexing convention in the R language, vectors `idnumSmall`, `uA` and `uA` can be used to slickly generate vectors of length 14 that contain contain replications of the entries of `uA` and `uA` using the replication information in `idnumSmall`. The Stan language also has this array indexing convention.

- Download the file `femSBMDviaStanSlick.R` from the subject web-site. In this file the model specification in the Stan language is done using the code:

  ```
  y ~ normal(X*beta + uGrp[idnum] + Zspl*uSpl,sigmaEps)
  ```

  Issue the command

  ```
  source("femSBMDviaStanSlick.R")
  ```

  The analysis is identical to the previous one, but use of the "`idnum` trick" means that the analysis is much faster, since matrices with high percentages of zero entries (known as *sparse* matrices) are not used.

3. Ordinary penalized splines are concerned with fitting a curve through a scatterplot. This question extends the notion to *bivariate* penalized splines, in which a *surface* is fitted to a *point cloud*.

   - Start an R session and ensure that you have an Internet connection. Issue the command:

     ```
     install.packages("rgl")
     ```

     This results installation of the R package rgl for visualisation via three-dimensional spin graphics.
   - Download the following files from the subject web-site

     (`http://matt-wand.utsacademics.info/37457.html`):

     - `Ztps.r`
     - `rglSetup.R`
     - `pointsInPoly.r`
     - `wolfcamp.txt`
     - `wolfcampKnots.txt`
     - `wolfcampBdry.txt`
     - `wolfcampViaStan.R`

     The first of these files creates bivariate spline functions for a given set of bivariate knot locations and bivariate data set.
   - The file `wolfcamp.txt` contains data on piezometric head (a liquid pressure measurement) at 85 geographical locations within the Wolfcamp Aquifer, Texas, U.S.A.
   - Issue the command `source("wolfcampViaStan.R")`. This fits a Bayesian bivariate nonparametric regression model of the form

     $$y_i | f(x_{1i}, x_{2i}), \sigma_\varepsilon^2 \sim N(f(x_{1i}, x_{2i}), \sigma_\varepsilon^2), \quad 1 \leq i \leq 85.$$

     where f is a smooth bivariate function, (x_{1i}, x_{2i}) correspond to geographical position and y_i corresponds to piezometric head. The data are transformed to the interval $[0, 2]$ for for Bayesian analysis and plotting. The script produces four graphical displays:

  - a plot showing the geographical positions of the data and bivariate knot locations,
  - Markov chain Monte Carlo summary checks,
  - an image plot showing the fitted surface with the R's "terrain" colour scheme, for which shades of green indicate lower bivariate function values and shades of brown and white indicate higher bivariate function values.
  - a three-dimensional spin graphics plot. Use the mouse or touch pad on your computer to spin the point cloud and fitted surface.

  Note that conversion back to the original units is not done in this script, but such an embellishment could be carried out with additional coding (but left out of this assignment).

- Download the following files from the subject web-site:
  - `ozoneMidwest.txt`
  - `ozoneMidwestBdry.txt`
  - `ozoneMidwestKnots.txt`

  The data in `ozoneMidwest.txt` corresponds to ozone level measurements at 147 geographical locations (recorded in longitude and latitude) in the Mid-West region of U.S.A.

- Copy the file wolfcampViaStan.R to a new file named `ozoneMidwestViaStan.R`. Open `ozoneMidwestViaStan.R` and modify the code so that the bivariate spline fitting is performed for the Mid-West U.S.A. ozone data rather than the Wolfcamp Aquifer data.

  Hint: The most obvious change needed is replacement of the character string `wolfcamp` by the character string `ozoneMidwest`.

- Save the image plot produced by the modified `ozoneMidwestViaStan.R`.

THE ONLY HAND-IN PART OF QUESTION 3:

Include the image plot produced by the last task in your Assignment 7 submission.

♦ ♦ ♦