University of Edinburgh

School of Mathematics

Bayesian Data Analysis, 2023/2024, Semester 2

Assignment 2

```
knitr::opts_chunk$set(echo = TRUE)
rm(list = ls(all = TRUE))
```

Figure 1: **The dataset is about the houses found in a given California district and some summary stats about them based on the 1990 census data.**

```
library(INLA)
```

```
## Loading required package: Matrix
```

```
## Loading required package: sp
```

```
## This is INLA_23.09.09 built 2023-10-16 17:29:11 UTC.
##  - See www.r-inla.org/contact-us for how to get help.
```

```
housing<-read.csv("housing.csv")
#removing rows with NA's, there are only a few of these
housing=housing[complete.cases(housing), ]
#creating a new covariate
housing$average_bed_rooms=housing$total_bedrooms/housing$households

head(housing)
```

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1   -122.23    37.88                 41         880            129        322
## 2   -122.22    37.86                 21        7099           1106       2401
## 3   -122.24    37.85                 52        1467            190        496
## 4   -122.25    37.85                 52        1274            235        558
## 5   -122.25    37.85                 52        1627            280        565
## 6   -122.25    37.85                 52         919            213        413
##   households median_income median_house_value ocean_proximity average_bed_rooms
## 1        126        8.3252             452600        NEAR BAY         1.0238095
## 2       1138        8.3014             358500        NEAR BAY         0.9718805
## 3        177        7.2574             352100        NEAR BAY         1.0734463
## 4        219        5.6431             341300        NEAR BAY         1.0730594
## 5        259        3.8462             342200        NEAR BAY         1.0810811
## 6        193        4.0368             269700        NEAR BAY         1.1036269
```

The covariates in the dataset are as follows:

longitude, latitude, housing_median_age (median age of houses in district), total_rooms (total rooms in all houses in district), total_bedrooms (total bedrooms in all houses in district), population (population of district), households (number of households in district), median_income (median income in district), median_house_value (median house value in district), ocean_proximity (categorical covariate about proximity of district to ocean), average_bed_rooms (average number of bedrooms of houses in district).

```
#We split the original dataset into two parts, training and test
housing.training<-housing[seq(from=1,to=nrow(housing),by=2), ]
housing.test<-housing[seq(from=2,to=nrow(housing),by=2), ]
```

Q1)[10 marks]

Fit a Bayesian Linear regression model in INLA (with Gaussian likelihood) using the housing.training dataset such that the response variable is the log(median_house_value), and the covariates in the model are as follows:

longitude, latitude, housing_median_age, log(median_income), ocean_proximity, average_bed_rooms.

Use scaled versions of the non-categorical covariates in your model.

Print out the model summary and interpret the posterior means of the regression coefficients.

Compute the DIC, NLSCPO and WAIC scores.

Check the sensitivity of your results to changing the priors.

Explanation: (Write your explanation here)

Q2)[10 marks]

Update your model in Q1 to also include an rw1 random effect model for the housing_median_age, and an ar1 random effect model for log(median_income).

Print out the model summary and interpret the posterior means of the regression coefficients.

Plot the posterior means of the random effects for housing_median_age and log(median_income). The x-axis should be the covariate value (such as housing_median_age), and the y-axis should be the posterior mean of the random effect.

Compute the DIC, NLSCPO and WAIC scores.

Check the sensitivity of your results to changing the priors.

Explanation: (Write your explanation here)

Q3)[10 marks]

In this question, we will use a spatial random effects model for the location.

Create a Bayesian regression model in INLA or inlabru with Gaussian likelihood using the housing.training dataset with log(median_house_value) as the response variable, and the fixed effects in the model are as follows:

longitude, latitude,

housing_median_age, $(housing\_median\_age)^2$,$(housing\_median\_age)^3$,$(housing\_median\_age)^4$

log(median_income), $(\log(median\_income))^2$, $(\log(median\_income))^3$, $(\log(median\_income))^4$ **,**

housing_median_age*log(median_income),

ocean_proximity, average_bed_rooms.

Use scaled versions of the non-categorical covariates in your model.

Include a spatial (spde2) random effect for the location (longitude, latitude), with Matern covariance. [Hint: You must create a mesh first; see the code for Lecture 7 and the solutions of Workshop 5.]

Print out the model summary and interpret the posterior means of the regression coefficients.

Plot the posterior mean of the spatial random effect in terms of the location.

Compute the DIC, NLSCPO and WAIC scores.

Compare the models in Q1) - Q3) in terms of DIC, NLSCPO and WAIC scores.

Check the sensitivity of your results to changing the priors and using a finer mesh.

Explanation: (Write your explanation here)

Q4)[10 marks]

In this question, we will evaluate the predictive performance of these models.

Do the following two tests for all 3 models.

First, compute the posterior mean of the log(median_house_value) for the districts in the training dataset housing.training. Compute the median absolute difference between the posterior means of the log(median_house_value) and its true values on the training dataset. This can be done by including the posterior means in an array v **, the true values in an array** t**, and computing** $\mathrm{median}(|v-t|)$.

Second, evaluate the log(median_house_value) 's posterior predictive means on the test dataset housing.test. Compute the median absolute difference between the log(median_house_value) 's posterior predictive mean and its true value on the test dataset.

Discuss the results.

Explanation: (Write your explanation here)

Q5)[10 marks] Perform posterior predictive checks (using replicates) on all 3 models Q1-Q3 fitted on the housing.training dataset. Choose your test functions to provide insight into the model. Discuss the results.

Explanation: (Write your explanation here)