# Bayesian learning and Monte Carlo Simulations: final project

F. Bassetti

May 27, 2022

## 1 General instructions

You are expected to write a short report (about 8 pages long - additional details may added in an appendix) with the following sections

- Description of the problem and the data.
- Model specification: what is the model for the data (likelihood)? What is the prior? Why you selected the specific set of values for the hyperparameters in the prior (if hyperparameters are present)?
- Posterior analysis and interpretation of the results. Some plots of the posterior distributions as well summary statistics of the posterior distributions are expected. Any sensitivity analysis should be included here, for example try different value for prior variance in case of LM.
- If needed: model selection or comparison with other models. You can compare different model, e.g. regression problems you can compare your results with a linear model if the data are not normally distributed. Alternatively, you can perform first model selection and then analyze the posterior results of the selected model(s).
- Final comments and conclusions.

Additional suggestions

- Any prediction exercise will be appreciated. If covariates are present, you can split the data in two parts, fit the model with the first part of the data and use the second part to do a prediction exercise (out of sample).
- You can add an Appendix with any additional analysis (such as auto-correlation plots and trace plots or other diagnostic tests).
- The code should be in different R file(s) (ready to be run by me if needed) with some short comments to be able to understand what you have done.

**Mandatory: the R file(s) and the the Project (pdf format) need to be submitted at least 3 day before the examination.**

## 2 Datasets

All the datasets (and some additional file of comments) are available on the Webeep pages.

### 2.1 Forest Fires data

This dataset is public available for research. The details are described in Cortez and Morais(2007). The dataset contains the following variables

1. X x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month: month of the year: "jan" to "dec"
4. day of the week: "mon" to "sun"
5. FFMC index from the FWI system: 18.7 to 96.20
6. DMC index from the FWI system: 1.1 to 291.3
7. DC index from the FWI system: 7.9 to 860.6
8. ISI index from the FWI system: 0.0 to 56.10
9. temp temperature in Celsius degrees: 2.2 to 33.30
10. RH relative humidity in %: 15.0 to 100
11. wind speed in km/h: 0.40 to 9.40
12. rain outside rain in mm/m2: 0.0 to 6.4
13. area the burned area of the forest (in ha): 0.00 to 1090.84.

In this dataset we are interested to model the burned area of the fo rest as a function of the rest of the variables.

Data are taken from: Cortez P. and Morais A. "A Data Mining Approach to Predict Forest Fires using Meteorological Data." In J. Neves, M. F. Santos and J. Machado Eds., "New Trends in Artificial Intelligence", Proceedings of the 13th EPIA 2007 Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

### 2.2 Ames House Price data

The data are taken from the database of the Ames City Assessor's Office. It includes a large number of variables and observations within the data set and they refer to 2930 property sales that had occurred in Ames, Iowa between 2006 and 2010. A mix of 76 nominal, ordinal, continuous, and discrete variables were used in the calculation of the assessed values. The dataset included physical property measurements in addition to computation variables used in the city's assessment process. All variables focus on the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information

that a typical home buyer would want to know about a potential property (e.g. When was it built? How big is the lot? How many feet of living space are in the dwelling? Is the basement finished? How many bathrooms are there? – for more details see at De Cock (2011)). The Ames Iowa dataset is referring at property sales that had occurred in US between 2006 and 2010.

The main response variable is the value of the house (SalePrice in $) to be explained using the other covariates.

Before starting the analysis is better to use only complete cases.

Additional info here:

http://jse.amstat.org/v19n3/decock.pdf

### 2.3 Ford Car Price Prediction

The dataset collects information about prices and features of Ford cars, in particular:

1. model: the Ford Car Brand
2. year: Production Year
3. price: Price of car in $
4. transmission: Automatic, Manual or Semi-Auto mileage: Number of miles traveled
5. fuel-type: Petrol, Diesel, Hybrid, Electric, Other
6. tax: Annual Tax
7. mpg: Miles per Gallon
8. engineSize: Car's Engine Size

Here the task is fit a suitable regression model which tries to explain the car prices in light of the other variables. Then, one can consider to predict the price of a new car model, given their features.

### 2.4 CO2 data

Human emissions of carbon dioxide and other greenhouse gases – are a primary driver of climate change – and present one of the world's most pressing challenges.

https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

Some data possibly related to C02 emissions, have been extracted from

https://ourworldindata.org.

Data have been selected for various nations and various years.

1. Country: name of the country.
2. y: year.
3. EnergyUse: Energy use (kg of oil equivalent per capita).
4. GDP: Gross Domestic Product per capita, PPP (constant 2017 international $).

5. pop: Population (historical estimates).

6. co2: Annual CO2 emissions (per capita)

7. lowcarbon: Low-carbon energy (% sub energy). Low-carbon energy is defined as the sum of nuclear and renewable sources. Renewable sources include hydropower, solar, wind, geothermal, wave and tidal and bioenergy. Traditional biofuels are not included.

8. urb: urban population (%) .

9. internet: number of internet users (OWID based on WB & UN).

Data taken from:
https://ourworldindata.org/grapher/energy-use-per-capita-vs-gdp-per-capita
https://ourworldindata.org/grapher/co2-emissions-vs-gdp
https://ourworldindata.org/grapher/low-carbon-energy-consumption?country=OMN Africa IDN
https://ourworldindata.org/grapher/urbanization-vs-gdp
https://ourworldindata.org/grapher/number-of-internet-users-by-country

The basic task is to consider a regression model to explain C02 emission with the other variables. You can transform some of the variables.

Additional questions: C02 and GDP are dependent? Historically, CO2 emissions have been strongly correlated with how much money we have. This is particularly true at low-to-middle incomes. The richer we are, the more CO2 we emit. This is because we use more energy – which often comes from burning fossil fuels. This relationship is stil true at higher incomes?

In addition you can: consider and compare various years. Consider the time as a covariate. Add more covariates (taking them from the web). Consider time series models.

### 2.5 Bank Marketing data

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

### 2.6 Salary data

Salary Data from 1994 Census database. The following dataset collets data about salaries taken from the 1994 Census database. Here the task is to determine whether a person makes over 50K a year or not.

Along with the salary variable (a binary variable), we have the following variables at out disposal:

1. age: continuous.

2. workclass: dummy - Private, Self-emp, gov, other (Without-pay, Never-worked).

3. fnlwgt: continuous, the final weight, i.e. the number of people the census believes the entry represents.
4. education-num: continuous, it is the number representation of the education level
5. is.married: marital-status: yes/no =1/0
6. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
7. is.white: White=1, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black=0
8. is.marel: Female=0, Male=1.
9. capital-gain: continuous.
10. capital-loss: continuous.
11. hours-per-week: continuous.
12. salary: $\leq 50K$ or $> 50K$

### 2.7 Wine data

There are two datasets that are related to red and white variants of the Portuguese "Vinho Verde" wine. We consider the white one. Due to privacy and logistic issues, only physic-ochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods. Input variables (based on physicochemical tests):

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH

10. sulphates

11. alcohol

12. Output variable (based on sensory data): quality (score between 0 and 10)

You can perform a classification using the different categories of wine using a categorical distribution (dcat in JAGS).

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

### 2.8 Covid Data

Nowcasting Covid Data. The dataset records some statistics related to COVID pandemia in Lombardia from 2020-12-06 to 2021-7-05.

The data arte taken from

https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita-regioni.csv

https://raw.githubusercontent.com/tsiotas/covid-19-zone/main/covid-19-zone.csv

The variables have been already selected and cleaned.

1. newpos: number of detected COVID positive subjects.

2. intcar: number of patients in intensive care.

3. hosp: number of patients in hospital.

4. newpos-av7D: average number of detected COIVD positive subjects over the previous 7 days.

5. color: color of the region (7 days before)

6. day: day (of the previous 5 statistics). R as.data format.

7. "hospH8": number of patients in hospital (7 day head)

8. "intcarH8":number of patients in intensive care (7 day head)

9. "dayH8": (day +7).

The nowcast exercise is to build and estimate a model to forecast on the basis of the variables at day t (newpos, intcar, hosp, newpos-av7D, color, day of the week) the number of patients in hospital and in intensive care at time t+7 (7 day head nowcasting).

You can start by fitting a model only for one of the variables, e.g. "hospH8". You can add any other covariate you like (just take them from the web). You can do the exercise using different time windows (e.g. after the omicron variant).

## 2.9 Forecasting Daily Coffee Price

Coffee is a brewed drink prepared from roasted coffee beans, the seeds of berries from certain flowering plants in the Coffea genus. From the coffee fruit, the seeds are separated to produce a stable, raw product: unroasted green coffee. The seeds are then roasted, a process which transforms them into a consumable product: roasted coffee, which is ground into fine particles that are typically steeped in hot water before being filtered out, producing a cup of coffee.

The coffee dataset contains daily coffee prices from 2015 to 2017 as they has been sold in the exchange. For every day, we have the following variables:

1. Open: opening price
2. High: highest price for the given day
3. Low: lowest price for the given day
4. Volume: the amount of coffee sold in the day

All prices are in USD. In this case, the task is to predict the coffee price for the next day, fitting a suitable AR model using JAGS. Use only one of the prices (e.g. the opening price). You may use Volume and or year as a covariates.