

Bayesian learning and Monte Carlo Simulations: final project 2024

F. Bassetti

May 20, 2024

1 General Instructions

You are expected to write a short report (about 8 pages long - additional details may be 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 did you select the specific set of values for the hyperparameters in the prior (if hyperparameters are present)? Be clear on the parametrization and the distribution you use, both in the likelihood and in the prior.
- Posterior analysis and interpretation of the results: Some plots of the posterior distributions as well as summary statistics of the posterior distributions are expected (e.g., posterior means and variances, confidence intervals). Any sensitivity analysis should be included here, for example, trying different values for prior variance in the case of LM.
- If needed: Model selection or comparison with other models. You can compare different models; e.g., in regression problems, you can compare your results with a linear model if the data are not normally distributed. Alternatively, you can perform model selection first 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). Similar exercises can be conducted in the case of time series model.
- 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 Final examination

- **Presentation Format:** The project presentation should be conducted using slides (20-25 minutes). It is required that each member of the group actively participates and discusses a specific portion of the project. This will allow for a comprehensive and well-rounded presentation, showcasing the collective effort of the entire group.
- **Questioning by the Examining Committee:** During the project discussion, the examining committee reserves the right to ask questions related to the project or topics covered in the course. These questions may be directed at one or more members of the group. We encourage all group members to be well-prepared and knowledgeable about the project and related concepts, including details and explanation on the codes.

3 Datasets

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

In order to decide the dataset, please have a look to the Rdm file on webeep for more information and visualization of the data.

1. Bike sharing data
2. Industrial production index
3. Airline Customer satisfaction
4. Wine data
5. JFK Passengers
6. US GDP & Inflation
7. Brazil forest fires
8. CO2 data
9. Acidity
10. Covid Data

3.1 Bike sharing data

Source. <https://www.kaggle.com/code/juniorbueno/rental-bikes/notebook>

name of the file. [bike.csv](#)

Short description. The data contains the number of casual/registered users in bike sharing systems and various additional covariates (related to the weather) as well information on days/month/year.

1. instant: record index
2. dteday : date
3. season : season (1:winter, 2:spring, 3:summer, 4:fall)

4. yr : year (0: 2011, 1:2012)
5. mnth : month (1 to 12)
6. hr : hour (0 to 23)
7. holiday : weather day is holiday or not
8. weekday : day of the week
9. workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
10. weathersit : 1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
11. temp : Normalized temperature in Celsius. The values are derived via $(t-tmin)/(tmax-tmin)$, $tmin=-8$, $tmax=+39$ (only in hourly scale)
12. atemp: Normalized feeling temperature in Celsius. The values are derived via $(t-tmin)/(tmax-tmin)$, $tmin=-16$, $tmax=+50$ (only in hourly scale)
13. hum: Normalized humidity. The values are divided to 100 (max)
14. windspeed: Normalized wind speed. The values are divided to 67 (max)
15. casual: count of casual users
16. registered: count of registered users
17. cnt: count of total rental bikes including both casual and registered

Task. Fit a regression model for predicting the count of total rental bikes (or casual/registered). Since the counts are large numbers, you can scale them (e.g., by dividing by 100 and subtracting 10). Discuss the impact of the various variables. Pay attention to the fact that there are many categorical variables. Perform a model selection and prediction exercise.

3.2 Industrial production index

Source.

<https://fred.stlouisfed.org/series/GDP>

name of the file. [indprod.csv](#)

Short description. The dataset consists in 12 economic indexes for the US economy. For each index reported is the Percent Change from Year Ago, Seasonally Adjusted, and the data are monthly. A detailed description is given in the additional explanation file. The indexes are:

1. Industrial Production: Total Index (INDPRO)
2. Wilshire 5000 Price Index (WILL5000PR)
3. New One Family Houses Sold: United States (HSN1F)

4. Crude Oil Prices: Brent - Europe (DCOILBRENTU)
5. Total Vehicle Sales (TOTALSA)
6. Consumer Price Index for All Urban Consumers: Food in U.S.
7. City Average (CPIUFDSL)
8. Japanese Yen to U.S. Dollar Spot Exchange Rate (DEXJPUS)
9. University of Michigan: Inflation Expectation (MICH)
10. CBOE Volatility Index: VIX (VIXCLS)
11. All Employees, Total Nonfarm (PAYEMS)
12. Producer Price Index by Commodity: All Commodities (PPIACO)
13. Sticky Price Consumer Price Index less Food and Energy (CORESTICKM159SFRBATL)

Task. Perform a linear regression on the dataset using the Industrial Production Index as the response variable and all the other variables as predictors. Discuss the importance of the various predictors and develop a parsimonious model. Discuss model selection, prediction and out-of-sample validation.

3.3 Airline Customer satisfaction

Source.

<https://www.kaggle.com/datasets/raminhuseyn/airline-customer-satisfaction>

name of the file. `airline_sub.csv`, full dataset (very big): `Airline_customer_satisfaction.csv`

Short description.

The dataset provides insights into customer satisfaction levels within an undisclosed airline company. While the specific airline name is withheld, the dataset is rich in information, containing 22 columns.

1. Satisfaction. Indicates the satisfaction level of the customer.
2. Customer Type. Type of customer: 'Loyal Customer' or 'Disloyal Customer'.
3. Age: Age of the customer.
4. Type of Travel. Purpose of the travel: 'Business travel' or 'Personal Travel'.
5. Class: Class of travel. 'Business', 'Eco', or 'Eco Plus'.
6. Flight Distance. The distance of the flight in kilometres
7. Seat comfort. Rating of seat comfort provided during the flight (1 to 5).
8. Departure/Arrival time convenient. Rating of the convenience of departure/arrival time (1 to 5).
9. Food and drink. Rating of food and drink quality provided during the flight (1 to 5).

10. Gate location. Rating of gate location convenience (1 to 5).
11. Inflight wifi service. Rating of inflight wifi service satisfaction (1 to 5).
12. Inflight entertainment. Rating of inflight entertainment satisfaction (1 to 5).
13. Online support. Rating of online customer support satisfaction (1 to 5).
14. Ease of Online booking. Rating of ease of online booking satisfaction (1 to 5).
15. On-board service. Rating of on-board service satisfaction (1 to 5).
16. Leg room service. Rating of leg room service satisfaction (1 to 5).
17. Baggage handling. Rating of baggage handling satisfaction (1 to 5).
18. Checkin service. Rating of check-in service satisfaction (1 to 5).
19. Cleanliness. Rating of cleanliness satisfaction (1 to 5).
20. Online boarding Rating of online boarding satisfaction (1 to 5).
21. Departure Delay in Minutes. Total departure delay in minutes.
22. Arrival Delay in Minutes. Total arrival delay in minutes.

We select from the original dataset 1000 customers.

Task. Perform a suitable regression for the categorical response variable "Satisfaction." Identify the variables that are correlated with "Satisfaction." A prediction exercise is required. You may discuss variable selection if desired.

Use only some of the covariates, starting with: Age, Seat Comfort, Flight Distance, Class, Departure Delay, and Arrival Delay.

3.4 Wine data

Source.

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.

name of the file. [wine.csv](#)

Short description. 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 physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

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)

Task. The Dataset can be viewed as classification or regression tasks. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods. You can try different models, but you need to try a Binomial regression. You can also aggregate votes, resulting in a different range of the Binomial. Discuss the differences if you estimate more than one model.

3.5 JFK Passengers

Source.

https://github.com/alan-turing-institute/TCPD/tree/master/datasets/jfk_passengers

name of the file. [JFK.csv](#)

Short description. The Port Authority collects monthly data for domestic and international, cargo, flights, passengers and aircraft equipment type from each carrier at PANYNJ-operated airports.

Task. This dataset can be analyzed using a change point model with two different means (before and after the change point), possibly adding a linear drift in the mean after the change point. More complex models with a change point can also be fitted and discussed. Time series models are suggested, two different trends and two different variances can be considered (before and after the change point).

3.6 US GDP & Inflation

Source.

<https://fred.stlouisfed.org/series/HSN1F>

name of the file. [gdp_inflation.csv](#)

Short description. The data consists in two time series:

1. Gross Domestic Product (GDP)
2. Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPI-AUCSL)

More info in the additional file.

Task. Fit some time series models for the two series (separately). You can try AR, MA, GARCH or ARMA. In case you use more model, compare the models with some Information criteria (BIC, DIC, WAIC). You can also try a bivariate time series models, e.g. a simple VAR(1) model (in this case, ask to the teacher for more information).

3.7 Brazil forest fires

Source.

<https://www.kaggle.com/code/lucasar/analysis-of-forest-fires-in-brazil-1998-2017>

name of the file. [amazon.csv](#)

Short description. This dataset reports the number of forest fires in Brazil divided by state. The series comprises the period of approximately 10 years (1998 to 2017).

Task. Fit a regression model with the number of fires as the response variable. You can start by considering data from only one state. The two basic covariates are the year and the month. You can try using an integer-valued distribution, or, if that fails, a normal model (scaling the data with some transformation if necessary).

If you want to use integer-valued observations, round the data to obtain integer numbers for the number of fires (since some data are not integers). You can also try a more complex model including data from more states.

3.8 CO2 data

Source.

<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>

name of the file. [CO2.csv](#)

Short description. 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. 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).

Task. Consider a regression model to explain CO2 emission with the other variables. You can transform some of the variables. Additional questions: CO2 and GDP are strongly 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 still 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.

3.9 Acidity

Source. Dataset taken from the R package ‘gamlss.data’.

name of the file. [acidity.csv](#)

Short description. The data shows the log acidity index for 155 lakes in the Northeastern United States.

Task. The students are required to create a model to provide an estimate of the density and of the clustering of data. Also, they should analyze how and why changes in the model and/or in the prior impact the obtained estimates.

3.10 Covid Data

Source.

<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>

name of the file. [covidLom2020.21.csv](#)

Short description. The dataset records some statistics related to COVID pandemia in Lombardia from 2020-12-06 to 2021-7-05. 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 COVID 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 wrt to hosp)
8. "intcarH8": number of patients in intensive care (7 day head wrt to intcar)
9. "dayH8": (day +7).

Task. 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).

Start by fitting a model only for one of the variables, e.g. "hospH8".

More challenging: estimate time series model to describe the behaviour of the pandemia (do not use hospH8,intcarH8,dayH8). Since the pandemia exhibit a peak you can consider a simple time series with a change point. You can also try well-known epidemic model (e.g. SIR). Be prepared that they do not work very well for the COVID.

3.11 SPF

Source.

<https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters>

name of the file. `SPF_GDP.csv`

Short description.

The Survey of Professional Forecasters (SPF) are survey of macrovariables. The US-SPF and ECB-SPF ask forecasters to report point forecasts and density forecasts. Density forecasts have the form of histograms with a set of intervals provided in the survey instrument.

More information on the dataset are provided in the Notebook SPF.

1. "H" horizon
2. "period" 0 corresponds to 10 bin, 1 to 11 bin
3. "YEAR" year
4. "QUARTER" quarter
5. "ID" forecaster id.
6. "INDUSTRY" forecaster type
7. "bin1" "bin2" "bin3" "bin4" "bin5" "bin6" "bin7" "bin8" "bin9" "bin10" "bin11" probability given to the bin (to be ignored)
8. "nbin_tot" number of bins used
9. "openL" "openR" 1 the forecaster gives positive probability to open (left/right) bin
10. "n.b.mode" position of the mode (wrt to bin number)
11. "prob.mode" probability assigned to the modal bin
12. "mode" value of the mode (uniform model fitting)
13. "mean" value of the mean (uniform model fitting)
14. "var" value of the variance (uniform model fitting)
15. "median" value of the median (uniform model fitting)

Task. Understand if individual uncertainty appears to be associated with a prominent respondent effect, while the point forecast (e.g. mean/median) is more affected by the period. That is, while there are marked differences across forecasters in the confidence attached to their predictions, forecasters' confidence changes slowly over time.

Start with a ANOVA type model (see the Notebook). Consider only data with a given Horizon (say 1, 2,...).

Compare variables related to point forecasters (mean/median) with variables related to uncertainty (e.g. variance/probability in the mode). In case you may change the model (e.g for probability in the mode).

Two more models need to be tested. The first is a model which take into considerations possible time effect in the variance of the errors. The second model is a mixture model in which forecaster belongs to common block and their response depends only on the block. Ask to the teacher for more explanation if you are interested in this project.

More complex: you can include in the model also the horizon.