MASc Thesis

andres.castillo

December 2023

Abstract

Acknowledgements

Contents

Chapter 1

Introduction

1.1 bla

1.2 blabla

Chapter 2

Literature review

Chapter 3

An overview of Bayesian inference

This section presents an overview of the basic concepts in Bayesian inference that are used for model estimation in the following chapters. It starts with a brief comparison between the Frequentist and Bayesian approaches, then, it presents the theoretical foundations of Bayesian statistics and how these principles are transformed into a practical framework for parameter inference and modelling in the real world.

This chapter is built on the ideas discussed in *A student's guide to Bayesian statistics* (Lambert, 2018), *Statistical rethinking: A Bayesian course with examples in R and Stan* (McElreath, 2016), and *Bayesian Data Analysis* (Gelman et al., 2013). For further details, these sources are a good starting point.

3.1 The Bayesian paradigm: Frequentist vs. Bayesian

The goal of statistical inference is to draw conclusions about a population by building a model that better represents the process of interest and estimating the parameters that better describe its behaviour. In statistical inference, there are two *Schools of thought*: the Frequentist (or Classical) and the

Bayesian approach.

For frequentists, the data are assumed to be the result of an infinite number of repeated experiments with the same characteristics. Then, the data are randomly sampled from a fixed and defined population, and any source of variation comes from that sampling process. Under this perspective, model parameters are assumed to be fixed but unknown values related to the population of interest, and the objective of inference is to calculate the best point estimate of the true value of the parameters given a data sample.

In contrast, Bayesian statistics assume that data are observed and fixed quantities and the source of variation comes from the uncertainty over the parameters. In this view, parameters are probabilistic values and the objective of inference is to estimate the probability distribution of the model's parameters. Then, we use the data as evidence to update any prior belief about the underlying process.

The debate about which approach is the best is interesting but long and almost philosophical, therefore, it is out of the scope of this document. However, Lambert (2018) argues that the Frequentist approach might make sense if the process of interest can be replicated multiple times as in the case of many natural sciences (i.e. multiple controlled experiments in a laboratory). Conversely, In contexts where the data collection can only be performed once such as in many social sciences, or transportation studies (e.g. democratic elections, population census, travel surveys), a Bayesian approach might be more aligned with the data nature.

3.2 The Bayesian thinking

The Bayesian framework shown in eq. (3.1) starts with some prior beliefs about the process we are interested in modelling. Then, we collect evidence (data) to update our beliefs using the model. The result of this update conforms to what is known as posterior belief.

$$\text{prior} + \text{data} \xrightarrow{model} \text{posterior} \tag{3.1}$$

An example of this update procedure in Bayesian inference can be a simple model of the wage gap by gender in a given labour market. Supported by

many researchers and study results, we can assume there is a gap in salaries between Males and Females and that this difference is normally distributed around a positive value. Then, we collect data from a salary survey and build a linear model $\hat{y} = \theta \cdot \text{gender}$, assuming the salaries are linearly correlated with gender. If we have proper and sufficient data, we obtain the estimated probability distribution of θ which corresponds to the estimated posterior distribution of the salary differences.

In this example, the location and shape of the posterior distribution of θ is determined by several factors: The location and shape of the prior distribution, the data size and quality, and the model specification.

Assuming our model is well specified, if we wrongly believe that there is no gender gap in salaries or that the difference is not significant, our data and model will update this belief and produce a posterior closer to the evidence we have. Conversely, If our prior belief and the evidence that we have supports the existence of a gender gap, our process would be well explained by our model and the uncertainty about the process (parameter variability) would be reduced.

3.3 The basics of Bayesian inference

3.4 Applying the Bayes theorem: Markov Chain Monte Carlo sampling (MCMC)

3.5 Probabilistic programming: A framework to perform MCMC

Chapter 4

Data sources

4.1 Survey of Labour and Income Dynamics - SLID

4.2 Hierarchical structure: Industry and Occupation

4.3 Exploratory Data Analysis

Chapter 5

Salary model estimation

5.1 Data structure: from single to multilevel structure

5.2 Model variables

5.3 Model specification

5.4 Model interpretability

5.5 Sensitivity analysis

5.6 Longitudinal analysis of the estimated parameters

5.7 Model validation

Chapter 6

Model integration

6.1 Integration with the existing implementation

6.2 Model comparison: proposed vs. existent

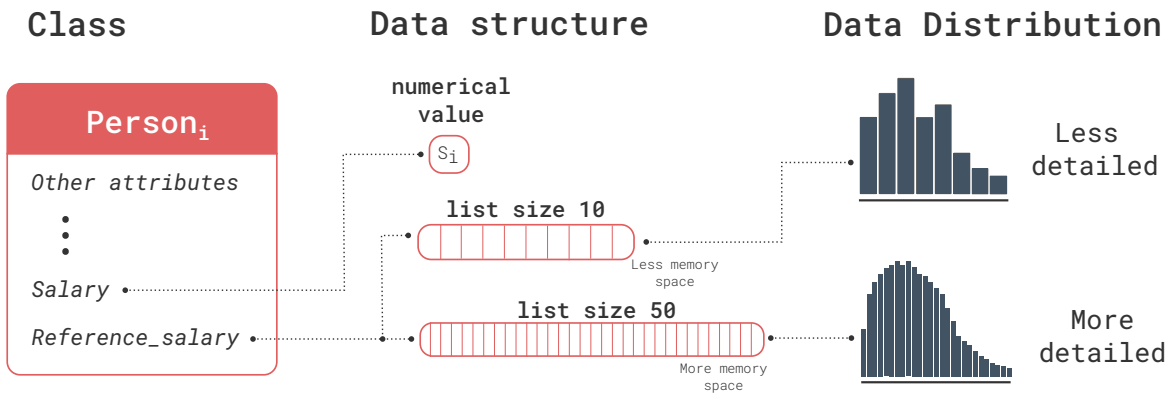

Figure 6.1: Representation of salary within the existent framework.

Chapter 7

Computational cost of Bayesian inference

Chapter 8

Conclusions

Chapter 9

Future work

Bibliography

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Texts in statistical science. CRC Press Boca Raton, Boca Raton, third edition edition.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE Los Angeles, Los Angeles.

McElreath, R. (2016). *Statistical rethinking : a Bayesian course with examples in R and Stan*. Texts in statistical science. CRC Press/Taylor & Francis Group Boca Raton, Boca Raton.