Bayesian Statistics and Modelling

Team: PizzaPizzaPizza

Team Members:

1. Akshett Jindal(2019114001)
2. Keshav Bansal (2019101019)
3. Shivang Gupta (2019101117)
4. Zishan Kazi (2019111031)

Objective

The main goal of this project is to gain understanding about Bayesian Statistics and Modelling by applying the knowledge on a couple of datasets modelling some parameters and understanding the distribution obtained.

Problem Statement:

In Bayesian statistics, we try to analyse the data and estimate parameters using the Bayes Theorem. In this method, we assume some joint distribution of the parameters (called **Prior Distribution**). It is generally chosen at random (from distributions like Normal, Poisson, etc.) and does not include any information from the data. Then we compute the likelihoods based on the data and apply Bayes Theorem to get the **Apriori Distribution.**

According to Renyl, if A and B are two events, then the conditional probability is given by:

$$P(B|A)=P(B\cap A)P(A)\text{But,}P(B|A)\neq P(A|B)\text{ and }P(A\cap B)=P(B\cap A)$$

These principles can be extended to the situation of data and model parameters. With data set y and model parameters θ and can be written as:
$P(\theta|y)=P(y|\theta)P(\theta)P(y)$

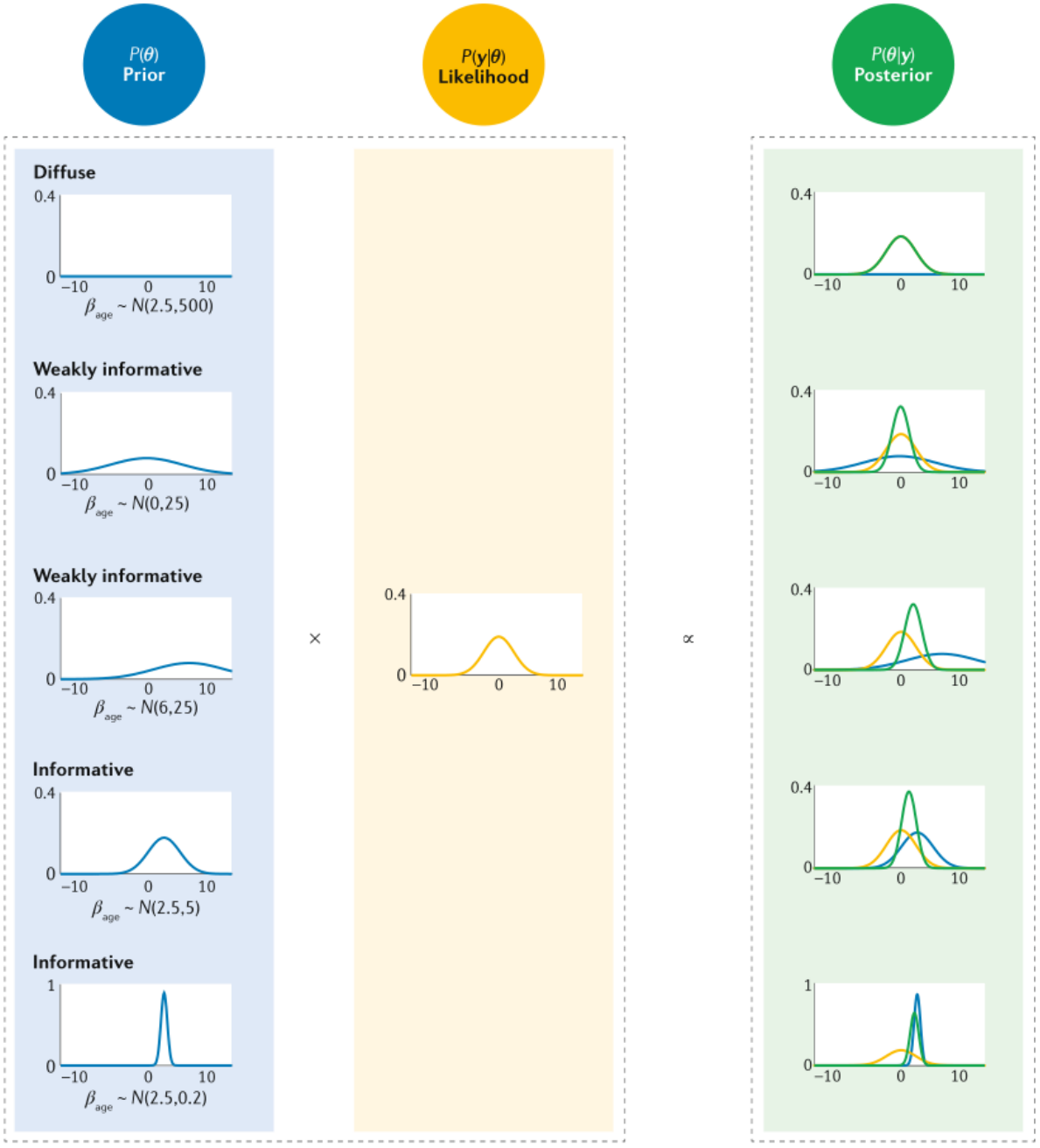
Prior: Probability distribution representing knowledge or uncertainty of a data object prior or before observing it

Posterior: Conditional probability distribution representing what parameters are likely after observing the data object

Likelihood: The probability of falling under a specific category or class.

Prior Uncertainity: An informative prior is one that reflects a high degree of certainity. e.g. an informative normal prior

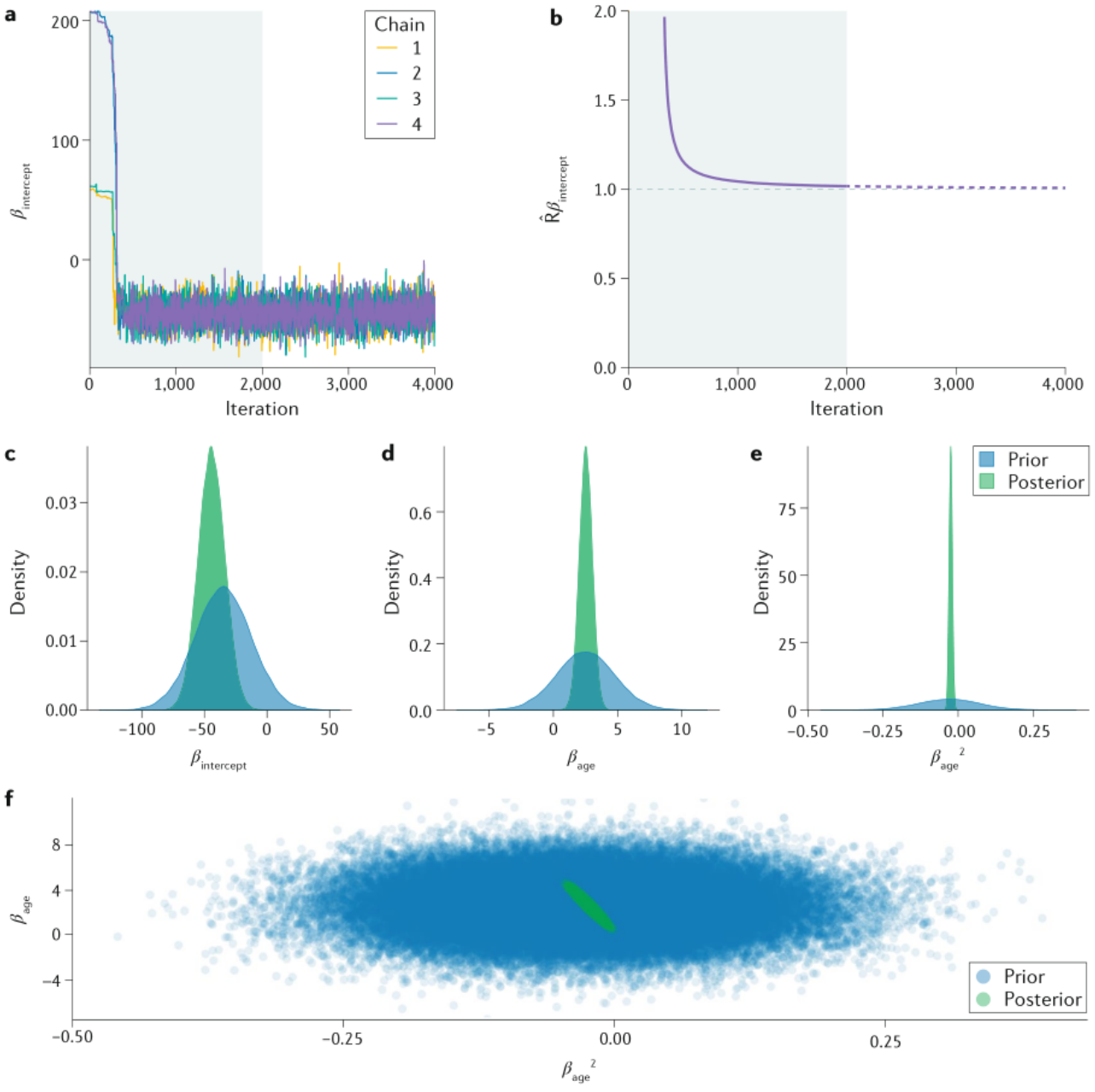
A weakly informative prior has a middling amount of certainty, being neither too diffuse nor too restrictive. A weakly informative normal prior would have a larger variance hyperparameter than an informative prior. Such priors will have a relatively smaller impact on the posterior compared with an informative prior, depending on the scale of the variables, and the posterior results are weighted more by the data observations as expressed in the likelihood.



Model Fitting

1.) So once the statistical prior model and likelihood function derived so now the task is to fit the model to the observed data.
2.) We use Markov chain Monte Carlo (MCMC) for posterior inference.
3.) MCMC is able to indirectly obtain inference on the posterior distribution using computer simulations.
4.) It combines two steps:
1. Obtaining a set of parameter values from the posterior distribution using the Markov chain.
2. Obtaining a distributional estimate of posterior with sampled parameters using monte carlo integration.

5.) The transition kernel determines the MCMC algorithm, describing how the parameter values and any other additional auxiliary variables are updated at each iteration of the Markov chain.
6.) If the proposed values are accepted, the Markov chain moves to this new state; whereas if the values are rejected, the Markov chain remains in the same state at the next iteration.



Testing

Posterior Predictive Checking

1. Once a posterior distribution for a particular model is obtained, it is used to simulate new data on this distribution that might be

helpful to assess whether the model provides valid predictions so that these can be used for extrapolating to future events.

Bayes factor

1. A Bayes factor is the ratio of the likelihood of one particular hypothesis to the likelihood of another. It tells us what the weight of the evidence is in favor of a given hypothesis. So it can be used to check the correctness of final model obtained.

Datasets

For this project we will use the following datasets:

1. Google Hangouts Chat Data -> In this we will model the `response_time` i.e. the time which a person takes to reply to a message.
2. Covid Testing - In this we will model the number of psoitive cases in a state.

We are also trying to gather some data regarding food wastage in the mess in the college for each day and try to model the food wastage which can be used by the messes/parliament for preventing wastage.

Timeline:

Task	Date
Using Frequentist technique for Estimating Model Parameters	Nov8-Nov 11
Using Bayesian technique for Estimating Model Parameters	Nov 12-Nov 14
Posterior Predictive Checking	Nov15-Nov16
Calculating Bayes Factor	Nov 16-Nov 17
Mid Evaluation	Nov 17-20
Hierarchical Modelling and bayesian Regression	Nov 21-Nov 25
Checking our implementations on various datasets like Google Hangouts Chat and Spam Filter	Nov 26-Dec 3
Final Evaluation and presentation	Dec 4

Work Distribution:

Team Member	Work
Akshett Jindal	Frequentist Technique and Bayesian Regression
Keshav Bansal	Posterior Predictive Checking and Bayes Factor
Shivang Gupta	Bayesian Technique and Hierarchical Modelling
Zishan Kazi	Working with Various Datasets and Markov Chain Monte Carlo