Team 3 : SMAI Project Report

Bayesian Statistics and Modelling

Introduction:

Bayesian Statistics is an approach to data analysis and parameter estimation based on Bayes theorem. We discuss the importance of prior and posterior predictive checking, selecting a proper technique for sampling from a posterior distribution,variational inference and variable selection.

Typically, Bayesian workflow consists of three steps:

- Capturing prior knowledge about a given parameter in a statistical model via the prior distribution (before data)
- Determining the likelihood function using the information about the parameters available in the observed data
- Combining both the prior distribution and the likelihood function using Bayes' theorem in the form of the posterior distribution.

Prior distribution:

Prior distributions play a defining role in Bayesian statistics. Priors can come in many different distributional forms, such as a normal, uniform or Poisson distribution, among others. Priors can have different levels of informativeness. The information reflected in a prior distribution can be anywhere on a continuum from complete uncertainty to relative certainty. Although priors can fall anywhere along this continuum, there are three main classifications of priors that are used.

Informative prior:

An informative prior is one that reflects a high degree of certainty about the model parameters being estimated. For example, an informative normal prior would be expected to have a very small variance.

Weakly Informative 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.

Diffuse priors:

A diffuse prior reflects a great deal of uncertainty about the model parameter. This prior form represents a relatively flat density and does not include specific knowledge of the parameter. A researcher may want to use a diffuse prior when there is a complete lack of certainty surrounding the parameter. In this case,the data will largely determine the posterior.

Likelihood function:

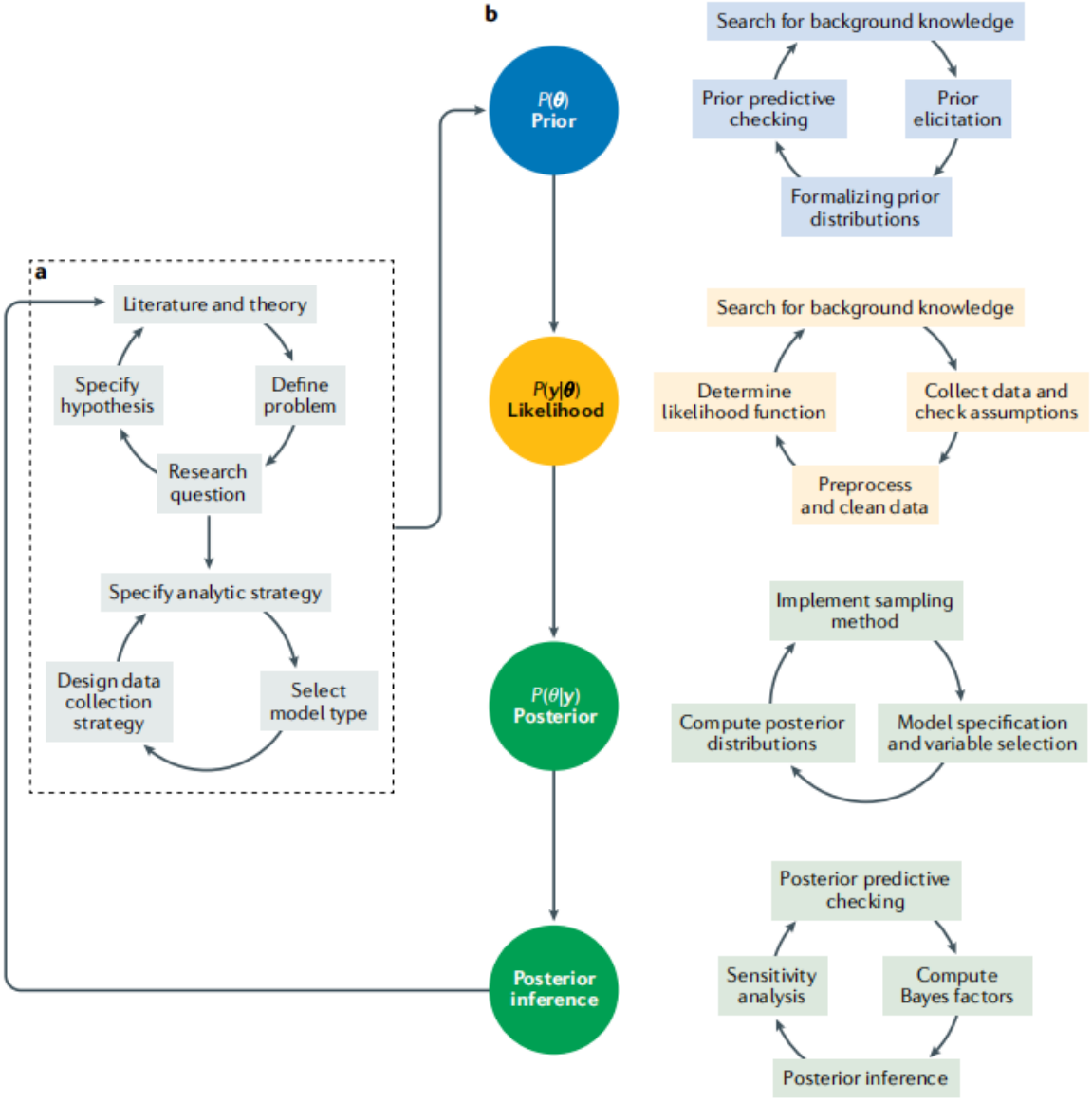
The likelihood is the conditional probability distribution $p(y|\theta)$ of the data (y), given fixed parameters (θ). In Bayesian inference, unknown parameters are referred to as random variables in order to make probability statements about them. The (observed) data are treated as fixed, whereas the parameter values are varied; the likelihood is a function of θ for the fixed data y. Therefore, the likelihood function summarises the following elements: a statistical model that stochastically generates all of the data, a range of possible values for θ and the observed data y.

Posterior distribution:

Once the statistical model has been defined and the associated likelihood function derived, the next step is to fit the model to the observed data to estimate the unknown parameters of the model. In Bayesian statistics, the focus is on estimating the entire posterior distribution of the model parameters. This posterior distribution is often summarised with associated point estimates, such as the posterior mean or median, and a credible interval. Direct inference on the posterior distribution is typically not possible, as the mathematical equation describing the posterior distribution is usually both very complicated and high-dimensional, with the number of dimensions equal to the number of parameters.

Markov chain Monte Carlo (MCMC):

MCMC is able to indirectly obtain inference on the posterior distribution using computer simulations. MCMC permits a set of sampled parameter values of arbitrary size to be obtained from the posterior distribution, despite the posterior distribution being high-dimensional and only known up to a constant of proportionality. MCMC combines two concepts: obtaining a set of parameter values from the posterior distribution using the Markov chain; and obtaining a distributional estimate of the posterior and associated statistics with sampled parameters using Monte Carlo integration.



Prior elicitation:

Prior elicitation is the process by which a suitable prior distribution is constructed. Strategies for prior elicitation include asking an expert or a panel of experts to provide values for the hyperparameters of the prior distribution. MATCH is a generic expert elicitation tool, but many methods that can be used to elicit information from experts require custom elicitation procedures and tools. Prior elicitation can also involve implementing data based priors. Then, the hyperparameters for the prior are derived from the sample data using methods such as maximum likelihood or sample statistics.

Prior Predictive checking:

The prior predictive distribution is a distribution of all possible samples that could occur if the model is true. In theory, a ‘correct’ prior provides a prior predictive distribution similar to the true data-generating distribution. Prior predictive checking compares the observed data, or statistics of the observed data, with the prior predictive distribution, or statistics of the predictive distribution, and checks their compatibility. There are multiple popular methods to accomplish this, prominent ones being:

- Comparing the obtained Kernel Density Estimation and predicted curves to measure similarity
- **Bayes factor** - a metric useful to compare two different models. It is defined as the ratio between posterior and prior odds of both models. This translates to the ratios of marginal likelihoods when the hypotheses in the models are contradictory. The Bayes factor would favour the more precise prior.

Kernel Density Estimation:

It is the process in which we use a kernel to smoothen a probability density function without using parameters. This approximates the probability density distribution without having to use a parametric approach giving us a reference curve.

Posterior Predictive checking:

Once a posterior distribution for a particular model is obtained, it can be used to simulate new data conditional 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. This is similar to how prior predictive checking can be used, but much more stringent in the comparison between the observed and simulated data.

Model fitting:

- The dataset we will be working with in this project is a Ph.D delay data sheet along with some attributes namely: Age, Age^2 , Have Children or not, Sex.
- We will try to model the delay in Ph.D in months in terms of Age, which is expected to be of the order 2, implying that our expected model is of the form:

$y = \beta_{intercept} + \beta_{age}\, Age + \beta_{age^2}\, Age^2 + \varepsilon$, where y is the expected delay, β’s are the coefficients of the parameter Age and ε is the residual.

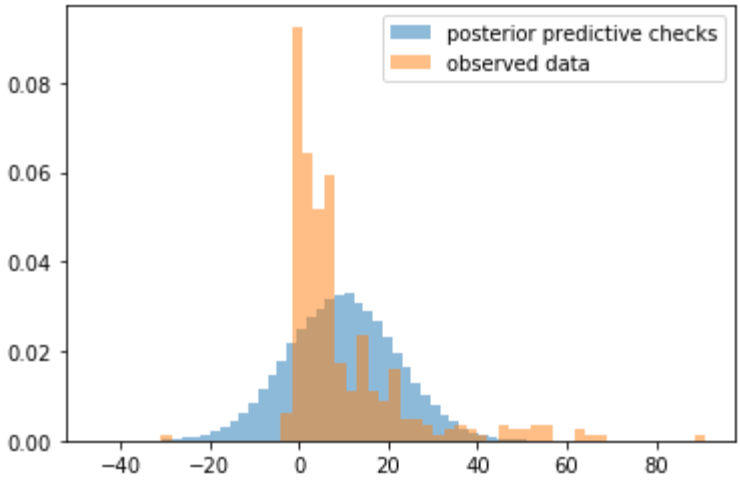
- The following is a visualisation of the dataset provided in this project analysis.

Modelling Procedure:

- We assume a Normal distribution for all of the β's and an Inverse Gamma distribution ε, as priors and try to find the posterior distribution
- The MCMC simulation is done using PyMC3, a Python framework for Bayesian analysis and statistics.
- The likelihood is estimated as a Normal distribution (refer to the code for specifics), from which we sample a large chunk of samples, which are then used to perform inference.
- Inference can be explained as performing a large number of simultaneous random walk simulations starting at a random sample from the obtained chunk, from which we repeatedly try to sample the normal parameters, also known as hyperparameters until a convergence is obtained.

Validating/Assessing the model parameters :

- This can be achieved by performing PPC (Prior and Posterior Predictive Checks)
- We plot the Posterior Predictive distribution and try to compare it with the dataset's histogram or observed data.
- If they are similar, it verifies that the posterior is likely a close resemblance to the observation set given.



- Plotting the correlation plots between parameters across various chains. It is important to avoid correlation between parameters to create a robust model.
- This plot is expected to converge to 0 as the iterations increase, implying a decline in correlation.
- Convergence in the MCMC chains is very important to ensure that our outcome of the model is valid. Divergences usually occur when the prior is too flat or sparse, implying the chain starts to diverge trying to explore a funnel/

encounter a sudden change in slope produced by the samples given (not always consistent)

- $\hat{R}$ statistic is an excellent parameter to ensure this convergence. It is defined as the ratio between inter and intra chain variability. It is expected to converge, hence must tend to 1, as the iterations increase.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
b_intercept	-47.722	9.199	-65.882	-31.399	0.129	0.092	5180.0	6775.0	1.0
b_0	2.541	0.398	1.770	3.263	0.005	0.004	5484.0	7212.0	1.0
b_1	-0.025	0.004	-0.033	-0.017	0.000	0.000	5544.0	7589.0	1.0
eps	3.096	4.875	0.063	10.875	0.073	0.056	7170.0	6032.0	1.0
sg	11.716	0.337	11.060	12.328	0.003	0.002	10644.0	8896.0	1.0

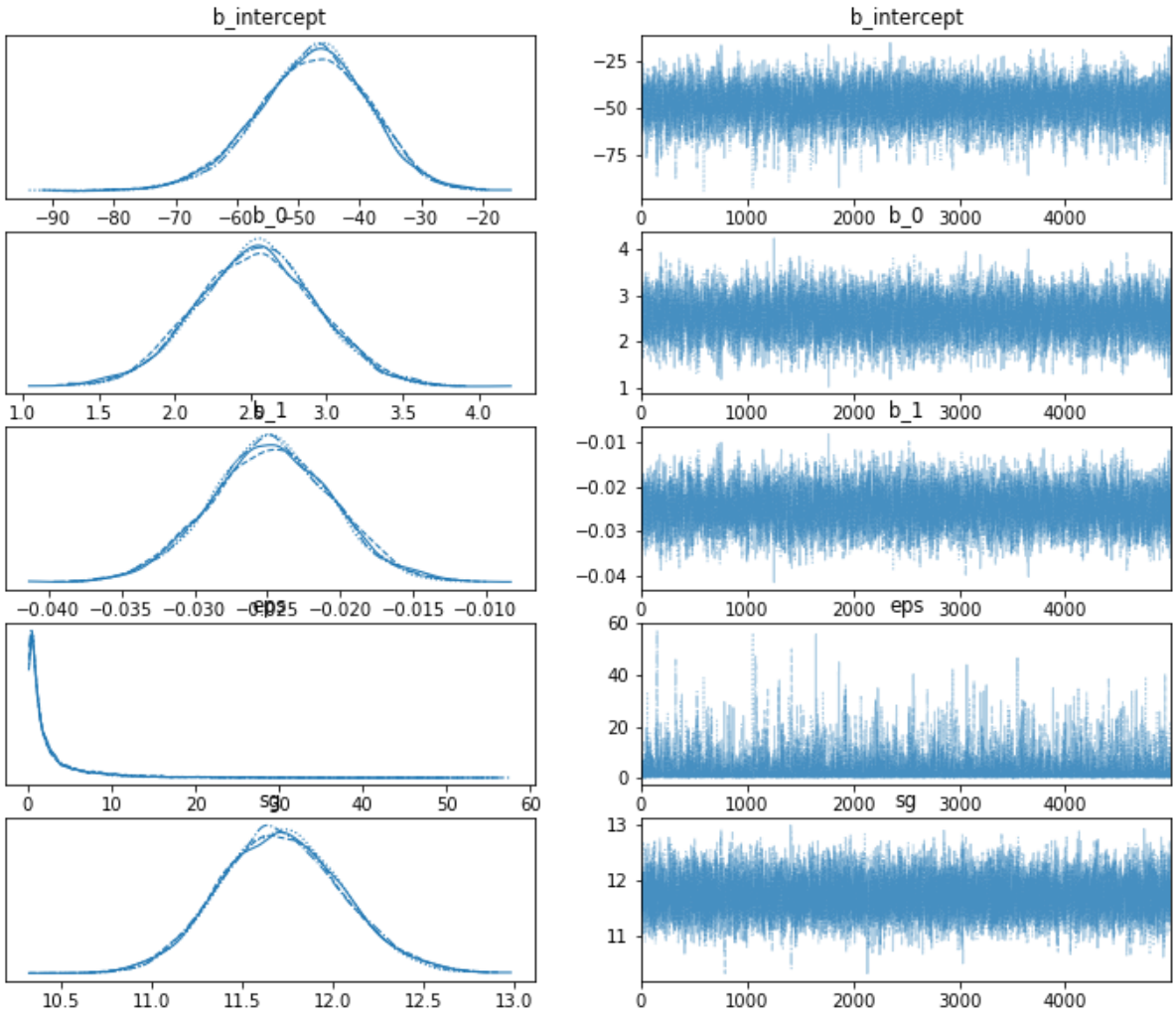
- The above are the obtained statistics and we can clearly identify that r_hat has converged to 1.

Results of the Inference:

- Upon plotting various parameters and statistics of the model and its distributions, we obtain some interesting observations that ensure that the model being trained is valid.
- **Single Point Parameters** can be obtained using Maximum a Posteriori method, which acts similar to MLE from Baye's theorem discussed in SMAI.
- It tries to find the maximum likelihood parameter set by sampling from the posteriors of each parameter, which usually gives a value close to the mean obtained of the posterior.
- The parameters obtained in our model are:

  {'b_intercept': -35.02104053, '
  b_0': 2.07642817, 'b_1': -0.0201467,
  'eps_log__': 1.10557563,
  'sg_log__': 2.45786061,
  'eps': 0.33102028,
  'sg': 11.67979715}

- For comparison, upon performing MSE on the dataset with these parameters, we get a slightly better error value than a Quadratic regression performed using sklearn, implying that the behaviour is as expected, and the modelling is fairly effective.

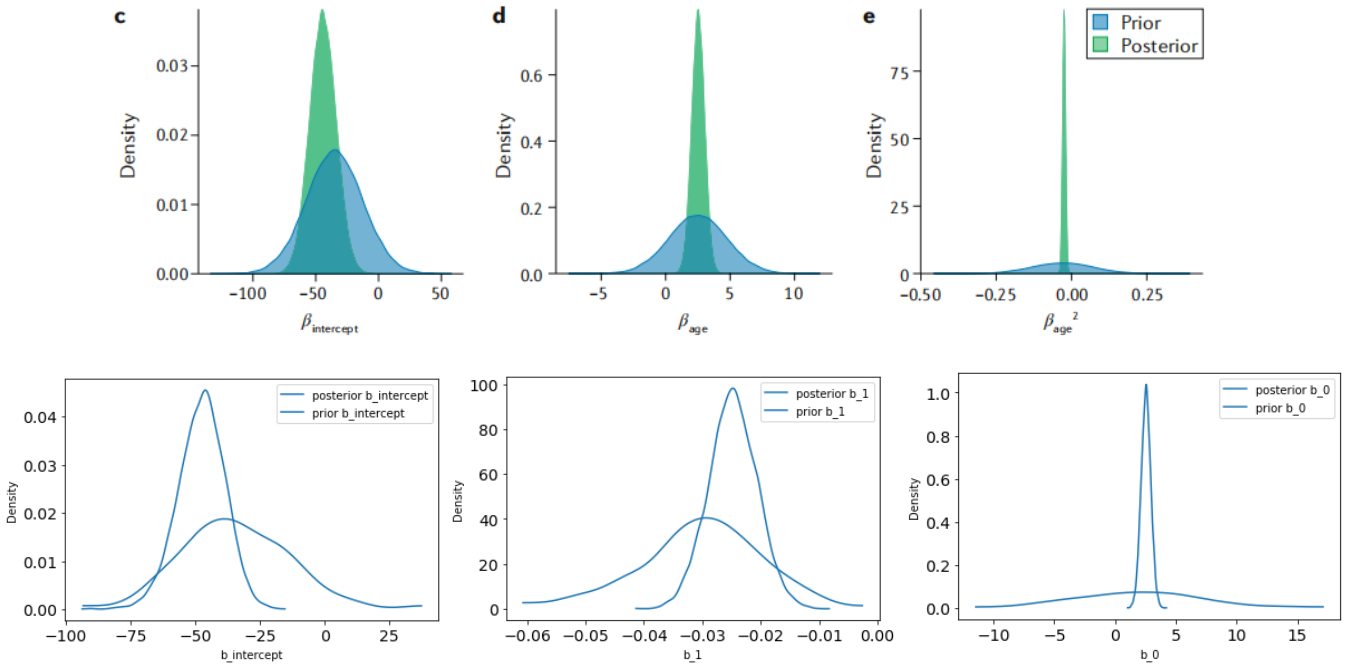


- The above plots are the trace plots obtained after the MCMC simulation performed on 4 chains.
- The right-side plots explain the values of parameters explored at each step of the random walk by the chains for each parameter, hence the hugely populated plot. The x-axis refers to the number of iterations , and y-axis refers to the value of the parameter.
- The dotted lines on the left are the individual posteriors obtained by each of the 4 chains in simulation, while the bold line is the final inference performed, and the posterior obtained as a result of all 4 chains which is essentially an intermediary of all four chains.

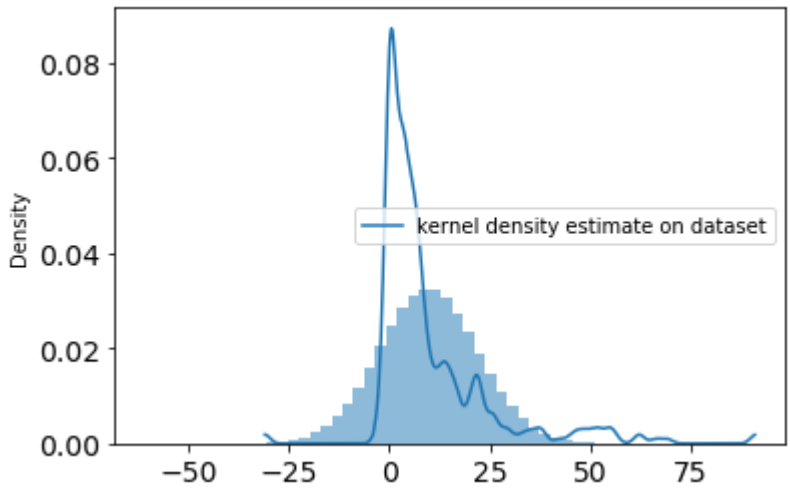
Comparison with the Paper:

- The below three are plots given in the paper, that they have obtained after sampling the posterior for each parameter from MCMC , compared with the prior taken.

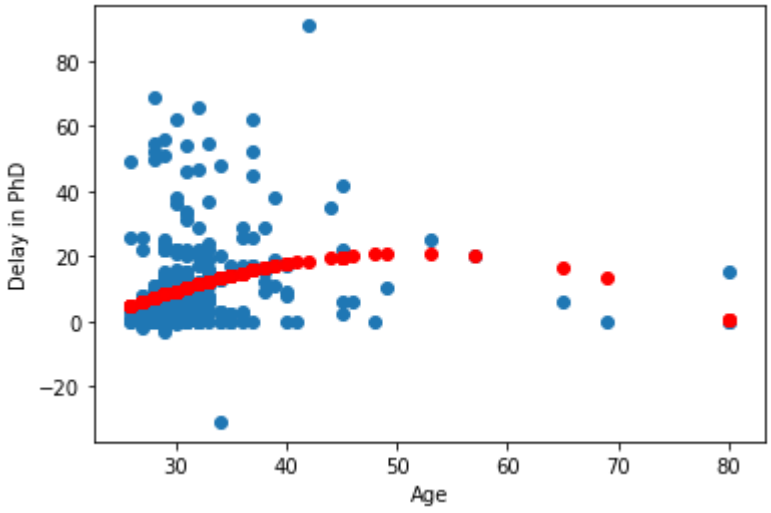
- Below them are the graphs obtained from our model, upon performing sampling on the posteriors obtained and plotting it against the prior we have considered. It is quite clear that trends are similar if we compare the plots.
- Although not perfectly the same, which is understandable, as keep in mind - we perform a RANDOM walk as a part of MCMC chain simulation hence the final values can be slightly different, but the trend is fairly similar, establishing the modelling has been a success.



- Some important implications of the plots are that there is a clear shortening in the variance of the posterior distributions of the parameters compared to the priors considered, implying there has been significant learning that the model underwent during the fitting.
- Another way to compare the obtained posterior with the given dataset is as mentioned before, comparing the KDE (Kernel Density Estimation) with the Posterior we have obtained, for the model.



- Below is the expected line fit of regression obtained upon considering the MAP (Maximum a posteriori) parameter sample plotted against the Observed sample.
- Do note, that this regression is not expected to be perfectly fitting of the data, as the data is very inconsistent and is not likely a poly linear regression in the age parameter alone.



Team Members and Work Division:

- Pothuri Praneeth Varma
    - MCMC, Modelling and Testing with poisson models and Normal Models, paper, Documentation
- Polakampalli Sai Namrath
    - MCMC modelling for Bayesian, Inference & Trace plots and statistics, Comparison with Linear regression, paper, Documentation
- Boyina Vamsi Krishna
    - Checking prior elicitation, Posterior checks and Likelihood with different distributions and regressions, paper
- Thakkallapally Rohan Rao
    - Checking Prior elicitations and likelihood with different distributions, paper, Documentation