TU Dortmund

Seminar

Model Averaging

Lecturers:
Prof. Dr. Paul Bürkner
Dr. Javier Aguilar
Soham Mukherjee

Author: Md Ryad Ahmed Biplob

Group number: 1

Group members: Sharmin Ahmed

September 25, 2025

Contents

1 Introduction

When faced with a challenge, we often look to create a model to understand and predict the outcome. Consider the scenario of being late for class. The exact amount of time one is late is a result of many contributing factors, and a good model should capture as many of them as possible. For example, one model might focus on how much an individual overslept, using historical data from the alarm clock habits. A second model could emphasize traffic congestion, incorporating real-time data from a mapping service. A third might be a simple function of the car's mechanical reliability.

Each of these models, when used alone, could give a different prediction for an individual's lateness. The first might say the person will be 10 minutes late, the second says 15, and the third says 5. The problem is, we can never be certain which of these is the "true" model that perfectly captures our specific situation. Choosing just one model and acting on its prediction would be shortsighted and possibly lead to further lateness. If only there were a technique to leverage all of them, to combine their strengths and account for their individual uncertainties. This is precisely the problem that Bayesian Model Averaging (BMA) seeks to solve.

2 Statistical methods

2.1 Bayesian Inference

According to (Martin et al., 2021, 1.1), Bayesian Models have two defining characteristics:

- Unknown quantities, also known as parameters, which are described using probability distributions
- The usage of Bayes' Theorem to update the values of the parameters conditioned by the data

Given some observed sample data $\boldsymbol{Y}$ and the parameter $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{Y})} \tag{1}$$

where the likelihood function $p(\boldsymbol{Y}|\boldsymbol{\theta})$ quantifies the relationship between the observed data $\boldsymbol{Y}$ and the unknown parameters $\boldsymbol{\theta}$. The prior distribution $p(\boldsymbol{\theta})$ represents our initial beliefs about these parameters before observing any data. Multiplying the likelihood and the prior yields the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$, which represents our updated beliefs about the parameters after observing the data. $p(\boldsymbol{Y})$ is the marginal likelihood and acts as a normalising constant, where $\boldsymbol{\Theta}$ is the set of all possible values of $\boldsymbol{\theta}$.

$$p(\boldsymbol{Y}) = \int_{\boldsymbol{\theta} \in \Theta} p(\boldsymbol{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

In some models, the marginal likelihood can be analytically solved. However in most cases, the numerical computation of the integral involves a high-dimensional integration over a usually complicated and highly variable function (Martin et al., 2021, 11.7). As a result, in most cases, the marginal likelihood is not generally computed and we see Equation 1 expressed as a proportionality.

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3}$$

2.2 Bayesian Model Averaging

In Equation 3, the posterior distribution can be considered combining a model with the data. So in practice, if either bad data or a bad model is chosen then the resulting posterior is a combination of both of these factors. This results in the following updated equation (Martin et al., 2021, 1.1):

$$p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{M}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{M}) p(\boldsymbol{\theta}, \boldsymbol{M}) \tag{4}$$

The specific model $\boldsymbol{M}$ under consideration is typically viewed as one of many plausible models within a set of candidate models $\mathcal{M} = \{\boldsymbol{M_1}, \ldots, \boldsymbol{M_K}\}$. The standard approach is to select a single "best" model, $\boldsymbol{M^\star}$, based on a goodness-of-fit criterion. This single-model selection ignores the inherent uncertainty in choosing one model over others. A significant issue arises when no single model adequately describes the data or when multiple models exhibit similar performance but produces substantially different predictions. Bayesian Model Averaging (BMA) addresses this by computing a marginal distribution—a mixture of the posterior distributions under each $\boldsymbol{M_K}$—where the contribution of each model is weighted by its posterior model probability (Hoeting et al., 1999, p. 383-384). According to (Hoeting et al., 1999), if $\boldsymbol{\theta}$ is the parameter of interest, given the data $\boldsymbol{Y}$ and the set of Models $\mathcal{M} = \boldsymbol{\{M_1, \ldots, M_K\}}$, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$ after considering each Model is given by,

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \sum_{\boldsymbol{k=1}}^{K} p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{M_k})p(\boldsymbol{M_k}|\boldsymbol{Y}) \tag{5}$$

where

$$p(\boldsymbol{M_k}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{M_k})p(\boldsymbol{M_k})}{\sum_{\boldsymbol{l=1}}^{K} p(\boldsymbol{Y}|\boldsymbol{M_l})p(\boldsymbol{M_l})} \tag{6}$$

and

$$p(\boldsymbol{Y}|\boldsymbol{M_k}) = \int p(\boldsymbol{Y}|\boldsymbol{\theta_k}, \boldsymbol{M_k})p(\boldsymbol{\theta_k}|\boldsymbol{M_k})d\theta_k \tag{7}$$

$p(\boldsymbol{M_k}|\boldsymbol{Y})$ is the posterior probability for the Model $\boldsymbol{M_k}$. $p(\boldsymbol{Y}|\boldsymbol{M_k})$ is the marginal likelihood of Model $\boldsymbol{M_k}$ where $\boldsymbol{\theta_k}$ is the vector of parameters for Model $\boldsymbol{M_k}$. $p(\boldsymbol{\theta_k}|\boldsymbol{M_k})$ is the prior for $\boldsymbol{\theta_k}$ under the model $\boldsymbol{M_k}$. Equation 7 is an extension of Equation 2 under the Model $\boldsymbol{M_k}$.

According to (Yao et al., 2018, p. 918), BMA is ideal for the $\mathcal{M}$-closed case, where the assumption is that the true data generating model is one of $\boldsymbol{M_k} \in \mathcal{M}$ although unknown. Some other restrictions that prevents BMA being one of the standard applications used in data analysis are as follows (Hoeting et al., 1999, p. 384):

- The number of terms in Equation 5 can become too high
- The various integrals involved require numerical approximations such as Markov chain Monte Carlo methods, since in majority of cases an analytical solution is not possible
- Specifying $p(\boldsymbol{M_k})$ is quite difficult and very little work has been done so far
- Selecting a group of models where the $\mathcal{M}$-closed case applies is also a very difficult task since there is yet to be any proper consensus in the selection process

2.3 Pseudo Bayesian Model Averaging

(Yao et al., 2018) reiterates the difficulty in calculating the Marginal Likelihood $p(\boldsymbol{Y}|\boldsymbol{M_k})$ of Model $\boldsymbol{M_k}$ and also mentions that it is very sensitive to the priors $p(\boldsymbol{\theta_k}|\boldsymbol{M_k})$ used in each model. An alternative to this is to compute the expected log pointwise predictive density (ELPD). It is more robust because it takes into consideration that BMA fails in the $\mathcal{M}$-open setting as mentioned in (Martín, 2021) and (Yao et al., 2018).

$$\text{ELPD} = \sum_{i=1}^{N} \log \int p(\boldsymbol{y_i}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\theta \tag{8}$$

According to (Martín, 2021), N is the number of data points, $\boldsymbol{y_i}$ is the $\boldsymbol{i}$-th data point, $\boldsymbol{\theta}$ are the parameters of the model, $p(\boldsymbol{y_i}|\boldsymbol{\theta})$ is the likelihood of the $\boldsymbol{i}$-th data point given the parameters, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution. Given the ELPD value for each model, the dELPD$_i$ is calculated which is the difference between the model with the best ELPD(higher is better) and the $\boldsymbol{i}$-th model. The weight $\boldsymbol{w_i}$ for the $\boldsymbol{i}$-th Model is the normalized dELPD$_i$ value.

$$w_i = \frac{\exp^{dELPD_i}}{\sum_j^M \exp^{dELPD_i}} \tag{9}$$

Unfortunately the ELPD is a theoretical quantity and also needs to be approximated. One of the methods used to approximate this values is the LOO, Pareto-Smooth-Leave-One-Out-Cross-Validation.

3 Statistical analysis

The analysis was conducted using the Python programming language (v3.11.5). The key libraries used include: `PyMC` (v5.25.1; Salvatier et al. (2016)), `ArviZ` (Kumar et al. (2019)), `Matplotlib` (Harris et al. (2020a)), `NumPy` (Harris et al. (2020b)), `Xarray` (Hoyer and Hamman (2017)), and `pandas` (McKinney (2010)).

Random seed was fixed to ensure reproducibility. Data preprocessing and visualization were done using `pandas` and `Matplotlib`, while Bayesian modeling was carried out using `pymc`, and inference diagnostics via `ArviZ`.

3.1 About Dataset

The dataset Soriano (2017) details the estimate of body fat percentage and various body circumference measurements for a total of 252 men. The data was provided by Dr. A. Garth Fisher. The dataset contains 14 variables, with the percentage body fat (calculated using the Siri equation) as the response variable. The covariates include age (years), weight (lbs), and height (inches), alongside 10 circumference measurements (in cm) across the neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist. The goal of our analysis is to create several Bayesian models and carry out model averaging. We aim to demonstrate how the resulting predictions are improved compared to using a single individual model selected based solely on goodness-of-fit.

3.2 Scatterplot Analysis

Figure 1 show the scatter plots of the variables abdomen, wrist, height and weight against siri. Some of the observations are pretty obvious. The abdomen, wrist and weight has a linear relationship with the percentage body fat variable. Height is a unique variable because people can come in different body shapes regardless of their age. Depending on theri body shape their fat percentage might vary as well.

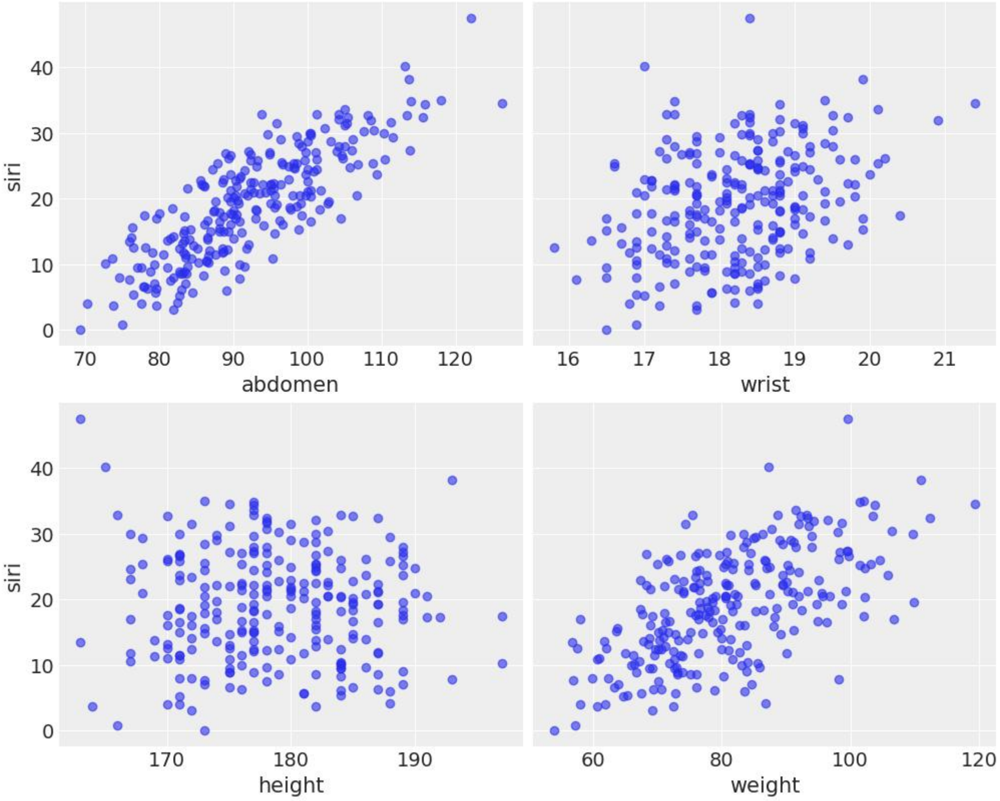

Figure 1: Scatter plots showing how `siri` varies with different body metrics.

3.3 Model Building

All models that we are testing are simple linear regression models where the response variable is `siri` and the independent covariates are individuals or a combination of the variables available in the dataset. We will be investigating how using a series of models and model averaging can produce improved results. All priors and linear transformations are kept simple to avoid compleximity of models. The parameters are α - the intercept for the linear equation, σ - the standard deviation and β a vector of coefficients for the covariates used in the linear regression models.

3.3.1 Prior Distributions

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = \alpha + \beta X$$

$$\alpha \sim \mathcal{N}(0, 1) \quad \text{with } \sigma > 0$$

$$\beta \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{N}(0, 5)$$

Here, α and β are given standard normal priors with a mean of 1 and a variance of 1 for both. The standard deviation of the residuals, σ, is also assigned a normal prior with the constraint of being positive only and the variance of 5.

3.3.2 Models and their covariates

Model 0 is defined with only the covariate `abdomen`. Model 1 is defined with the covariates `wrist`, `height` and `weight`.

3.3.3 Model Posteriors and Comparisons

Figure 2 and 3 shows posterior predictive against the sample data. In Model 0 seems to be performing worse when compared to Model 1. In Table 1 we can see the ELPD scores for each model alongside with their respective weights. Based on their respective posterior predictive plots, the lower weight for Model 0 seems evident. Figure 4 shows all of posterior means together for all three models including the average model. While comparing , we can see that the average model is basically passing between the two models.

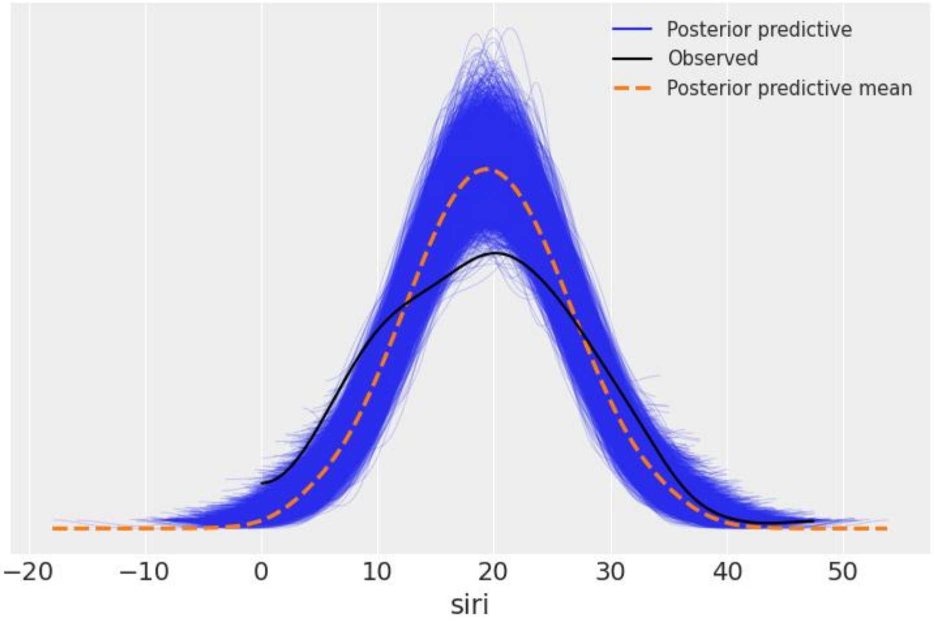

Figure 2: Comparison between observed distribution of `siri` and posterior predictive samples for Model-0

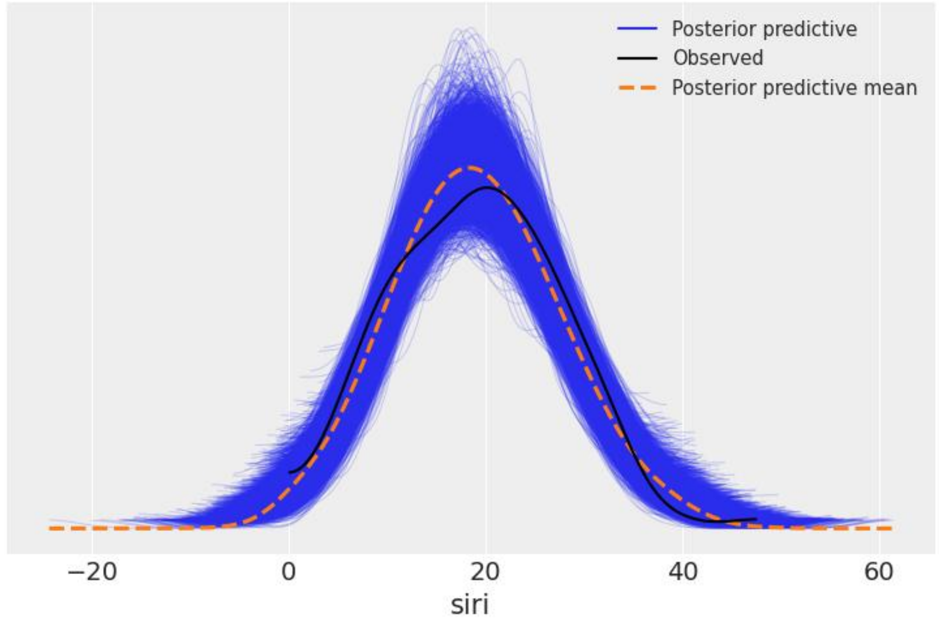

Figure 3: Comparison between observed distribution of `siri` and posterior predictive samples for Model-1

Model	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning
model_1	0	-817.30	3.70	0.00	0.64	10.46	0.00	False
model_0	1	-825.36	1.88	8.05	0.36	9.98	8.66	False

Table 1: Comparison of models using Leave-One-Out cross-validation metrics.

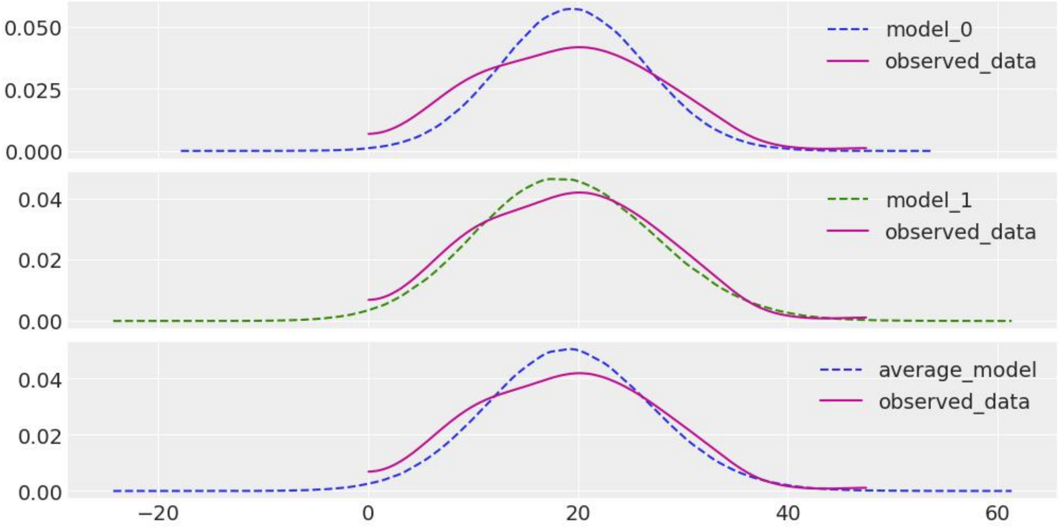

Figure 4: Comparison of model predictions with observed data using posterior predictive distributions.

4 Summary

Bayesian Model Averaging (BMA) is a sophisticated approach to inference and prediction that uniquely addresses the fundamental uncertainty inherent in model selection. Instead of committing to a single "best" model, BMA operates on the idea that the true data-generating process is probably represented by a combination of probable models.

BMA achieves this by constructing a weighted average of the predictions made by a set of candidate models. The weights are determined by the posterior model probabilities—a measure of each model's certainty given the observed data and prior beliefs. This results in predictions that are more robust and provide a more accurate quantification of predictive uncertainty compared to predictions derived from any single model.

Although BMA is a mathematically elegant and a theoretically sound technique, its practical application faces strict requirements. It performs optimally under M-closed cases (where the true model is assumed to be in the candidate set), and its performance can degrade otherwise. Furthermore, the wider adoption of BMA is often hampered by the lack of streamlined industry standards and consensus on implementation. Nonetheless, BMA can achieve excellent predictive performance, provided the researcher makes a proper choice of a diverse set of candidate models that collectively capture unique and relevant aspects of the dataset.

Bibliography

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585:357–362, 2020a. doi: 10.1038/s41586-020-2649-2.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020b. URL `https://www.nature.com/articles/s41586-020-2649-2`.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999. ISSN 08834237. URL `http://www.jstor.org/stable/2676803`.

Steven Hoyer and Joseph J Hamman. xarray: N-d labeled arrays and datasets in python. *Journal of Open Research Software*, 5(1), 2017. doi: 10.5334/jors.148.

Ravin Kumar, Christopher Carroll, Ari Hartikainen, and Osvaldo Martin. Arviz a unified library for exploratory analysis of bayesian models in python. *Journal of Open Source Software*, 4(33):1143, 2019. doi: 10.21105/joss.01143.

Osvaldo A. Martin, Ravin Kumar, and Junpeng Lao. *Bayesian Modeling and Computation in Python*. Boca Raton, December 2021. ISBN 978-0-367-89436-8.

Osvaldo Martín. Model averaging. In: PyMC Examples, 2021.

Wes McKinney. *Data Structures for Statistical Computing in Python*. 2010. URL `https://conference.scipy.org/proceedings/scipy2010/mckinney.html`.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016. doi: 10.7717/peerj-cs.55.

W. E. Siri. Gross composition of the body. In *Advances in Biological and Medical Physics*, volume IV, page ??

Fede Soriano. Body fat prediction dataset, 2017. URL `https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset`. Kaggle Dataset.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Anal.*, 13:917–1007, September 2018. doi: 10.1214/17-BA1091. URL `https://doi.org/10.1214/17-BA1091`.