Model Evaluation and Selection the Bayesian Way

Thu Nguyen

Stat 520C Final Project

April 28, 2023

1 Introduction

The report studies how models are evaluated and selected within the Bayesian framework. The reference text is *Bayesian Data Analysis* by Gelman et al [Gel+13], specifically chapter 7. Section 2 gives an overview of the different metrics used in evaluating fitted models and selecting the best fitted model. Section 3 provides a simple example of fitting a linear regression model with a non-informative prior and applies the metrics in section 2. Section 4 concludes with some remarks. The code is available at github.com/ngthu003/stat520C_finalProject.

2 Model Evaluation and Selection

2.1 Model Evaluation

A regular workflow in statistical learning often starts with designing and fitting a model and ends with evaluating the model. Model evaluation is assessed via the accuracy of the fitted model's prediction. Under Bayesian inference, and in particular probabilistic prediction, the goal is to provide inferences about some data Y that can capture the full uncertainty about Y. This predictive accuracy is usually quantified through some scoring rules. A commonly used rule is the log predictive density $\log p(Y|\theta)$, where Y is some data point(s) and θ the parameter(s) of interest.

We make a first remark that when it comes to assessing a model's accuracy, we will work with the log predictive density instead of the log posterior density since the latter is only relevant for the purpose of estimating the parameters and not for the task at hand.

Let θ be the parameter(s) of interest. Let $p(\theta)$ be the prior. Given observed data Y generated from some true model f, let $p_{\text{post}}(\theta) = p(\theta|Y)$ be the posterior. Let $\tilde{Y}_i$ be a new data point, we define the out-of-sample predictive fit as

$$\log p_{\text{post}}(\tilde{Y}_i) = \log \mathbb{E}_{\text{post}}[p(\tilde{Y}_i|\theta)] = \log \int p_{\text{post}}(\tilde{Y}_i|\theta)\ p_{\text{post}}(\theta)\ d\theta.$$

We can now introduce a few measures of predictive accuracy:

1. Expected log predictive density for a new data point $\tilde{Y}_i$ given the true model f

$$\text{elpd}(\tilde{Y}_i) = \mathbb{E}_f\left[\log p_{\text{post}}(\tilde{Y}_i)\right] = \int \log p_{\text{post}}(\tilde{Y}_i)\ f(\tilde{Y}_i)\ d\tilde{Y}.$$

2. Expected log pointwise predictive density for a new dataset

$$\text{elppd} = \sum_{i=1}^{n} \text{elpd}(\tilde{Y}_i) = \sum_{i=1}^{n} \mathbb{E}_f \left[\log p_{\text{post}}(\tilde{Y}_i) \right]$$

3. Log pointwise predictive density for a fitted model from the training data Y

$$\text{lppd} = \log \prod_{i=1}^{n} p_{\text{post}}(Y_i) = \sum_{i=1}^{n} \log \int p(Y_i|\theta)\ p_{\text{post}}(\theta)\ d\theta.$$

In practice, to compute lppd, specifically the integral which we note is $\mathbb{E}_{\text{post}}[p(Y_i|\theta)]$, we typically draw a large number of samples from the posterior $p_{\text{post}}(\theta)$:

$$\text{computed lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta^s) \right).$$

This idea of repeated sampling as a way to estimate integrals and expectations is a common theme in Bayesian statistics, and will in fact be used in evaluating the various metrics in subsection 2.2.

2.2 Model Selection

In practice, we often try out a variety of model designs. Once the models have been fitted, the next natural task is to select the *best fitted model*. We introduce here various metrics to evaluate and select models. These metrics are usually referred to as *information criteria* in literature.

1. AIC - Akaike information criteria

   - to estimate the out-of-sample predictive accuracy, measured by the expected log predictive density, conditional on the MLEs (maximum likelihood estimates):

$$\text{AIC} = -2\ \hat{\text{elpd}}_{\text{AIC}} = -2 \left(\log p \left(Y|\hat{\theta}_{MLE} \right) - k \right).$$

2. DIC - Deviance information criteria

   - a *somewhat* Bayesian version of AIC, where the MLEs are replaced with the posterior mean $\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|Y]$:

$$\text{DIC} = -2\ \hat{\text{elpd}}_{\text{DIC}} = -2 \left(\log p \left(Y|\hat{\theta}_{\text{Bayes}} \right) - p_{\text{DIC}} \right),$$

   where $p_{\text{DIC}} = 2 \left(\log p \left(Y|\hat{\theta}_{\text{Bayes}} \right) - \mathbb{E}_{\text{post}} \left[\log p(Y|\theta) \right] \right)$.

3. WAIC - Watanabe-Akaike information criteria

   - a fully Bayesian approach using the computed log pointwise posterior predictive density:

$$\text{WAIC} = -2\ \hat{\text{elppd}}_{\text{WAIC}} = -2 \left(\text{lppd} - p_{\text{WAIC}} \right),$$

   where $p_{\text{WAIC}} = 2 \sum_{i=1}^{n} \left(\log \mathbb{E}_{\text{post}}[p(Y_i|\theta)] - \mathbb{E}_{\text{post}} \left[\log p(Y_i|\theta) \right] \right)$.

4. LOO-CV - Leave-one-out cross validation
    - the usual LOO-CV where in every i^{th} iteration (n in total) a model is fitted on all but the i^{th} observation, after which the predictive fit is computed on that i^{th} observation:

$$\text{LOO-CV} = -2\text{lppd}_{\text{LOO-CV}} = -2\sum_{i=1}^{n} \log p_{\text{post}(-i)}(Y_i).$$

Since these information criteria are for model selection, it matters less the magnitude but rather the ranking of the different models. Ideally we hope to see the similar ranking across the different criteria, in which case we will choose the model with the lowest value.

3 Simulations

3.1 Estimating Linear Regression Models the Bayesian Way

Let us consider a simple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, and $\mathbf{X}$ is the design matrix (including the intercept column) of dimension $n \times J$ (e.g. $J = 3$ for the full model). We note that this is equivalent to specifying

$$Y \sim \mathcal{N}\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2\right).$$

Instead of the usual Ordinary Least Squares estimates, we will apply the Bayesian approach. Following [Gel+13], consider a non-informative prior

$$p(\boldsymbol{\beta}, \log\sigma) \propto 1.$$

This gives the posterior distribution as

$$f(\boldsymbol{\beta}, \sigma^2|Y) = \frac{f(\boldsymbol{\beta}, \sigma^2|Y)}{f(\boldsymbol{\beta}|\sigma^2, Y)}\ f(\boldsymbol{\beta}|\sigma^2, Y) = f(\sigma^2|Y)\ f(\boldsymbol{\beta}|\sigma^2, Y).$$

These are:

- The marginal posterior distribution of the variance σ^2:

$$\sigma^2|Y \sim \text{Inv} - \chi^2(n - J, s^2),$$

  where s^2 is the usual unbiased estimator in OLS:

$$s^2 = \frac{1}{n-J}(Y - \mathbf{X}\hat{\boldsymbol{\beta}})^T(Y - \mathbf{X}\hat{\boldsymbol{\beta}})$$

- The conditional posterior distribution of the coefficients $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\sigma^2, Y \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, V_\beta\sigma^2),$$

  where $\hat{\boldsymbol{\beta}}$ is the usual OLS estimate:

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TY$$
$$V_\beta = (X^TX)^{-1}.$$

The detailed derivatives of these posterior distributions are available in chapter 14 of [Gel+13].

3.2 Set-up

In our simulation, let the true model be

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} .5 \\ 4 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \sigma^2 = 1.5,$$

i.e. the true model only depends on X_1 and not X_2 as in 1. We will consider two scenarios of

$$(1)\ n = 15 \qquad \text{and} \qquad (2)\ n = 50.$$

In each of the two scenarios, we will first fit the following model choices:

$$\begin{aligned} &\text{Model 1 (M1):} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \\ &\text{Model 2 (M2):} \quad Y = \beta_0 + \beta_1 X_1 + \epsilon, \\ &\text{Model 3 (M3):} \quad Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon, \\ &\text{Model 4 (M4):} \quad Y = \beta_1 X_1 + \epsilon. \end{aligned}$$

Once fitted, we will use the information criteria in section 2 to select the best model. Given that the true model is $Y = .5 + 4X_1 + \epsilon$, we hope that *Model* 2 will be chosen in both scenarios.

3.3 Results

Figure 1 shows the posterior distribution of the log predictive density $\log p(Y|\boldsymbol{\beta}, \sigma^2)$ for each of the 4 fitted models. Table 1 shows the maxima of of the density, which are obtained when $\boldsymbol{\beta}$ and σ^2 are the MLEs (maximum likelihood estimates). We note an interesting finding that the (incorrect) full model (M1) is not doing significant worse than the true model (M2), and even better in the case of $n = 50$ when it comes to predicting the posterior likelihood of the data given the MLEs.

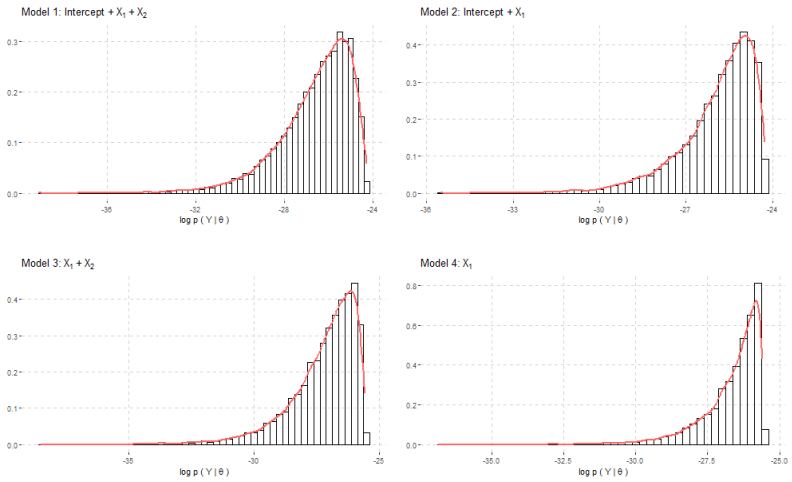

Figure 1: Posterior distribution of the log predictive density $\log p(Y|\boldsymbol{\beta}, \sigma^2)$ when $n = 15$.

Figures 2 and 3 compare the various information criteria introduced in section 2 from the 4 fitted models. We note some findings:

1. In both scenarios (of $n = 15$ and $n = 50$), the best fitted models are (M2) Intercept $+X_1$, which is indeed the true model.

Model	n	$\max \log p(Y\|\boldsymbol{\beta}, \sigma^2)$	$\mathbb{E}[\log p(Y\|\boldsymbol{\beta}, \sigma^2)]$
(M1) Intercept + X_1 + X_2	15	-24.27	-26.61
(M2) Intercept + X_1	15	-24.26	-25.92
(M3) X_1 + X_2	15	-25.57	-27.23
(M4) X_1	15	-25.60	-26.67
(M1) Intercept + X1 + X_2	50	-69.62	-71.74
(M2) Intercept + X1	50	-69.71	-71.25
(M3) X_1 + X_2	50	-73.78	-75.33
(M4) X_1	50	-73.80	-74.81

Table 1: The maxima and means of the log predictive density from the 4 fitted models.

2. The next best fitted model is now different for each scenario: when $n = 15$, the next best is (M4) X_1 while for $n = 50$, it is (M1) Intercept $+X_1 + X_2$, the full model. It appears as though the two models are *swapped* in terms of how they trail behind the best fitted model (M2).

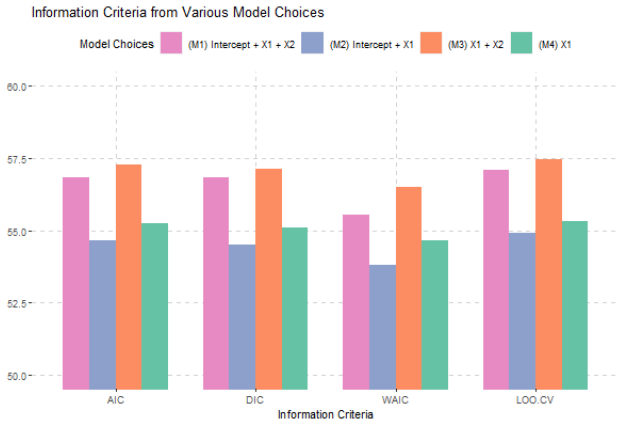

Figure 2: Various information criteria from the fitted models, when $n = 15$.

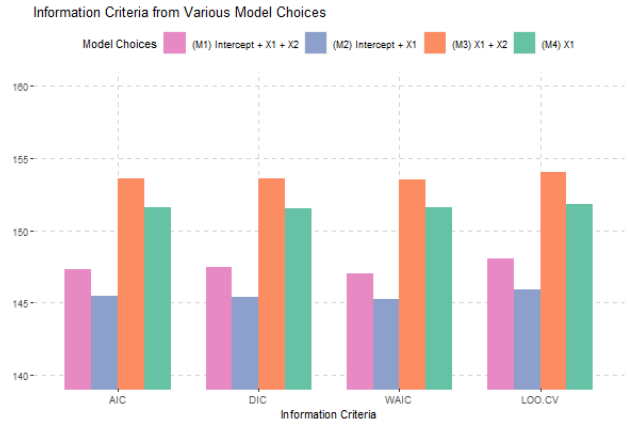

Figure 3: Various information criteria from the fitted models, when $n = 50$.

4 Conclusion

This report serves as a brief overview of how models can be evaluated and compared under the Bayesian framework. All of these are just one way to do so, and there are certainly other approaches to evaluate and compare models. We emphasize again that the metrics in section 2 are not absolute, but rather should be compared relatively between models. The ideal situation is when the ranking of models is similar across the different metrics, in which case we can be highly confident about how each model performs relative to the other. Section 3 provides an application of such metrics in selecting the best fitted linear regression model. We note though that while the chosen best model is indeed the true model, the next best model appears to be susceptible to the data size.

References

[Gel+13] A. Gelman et al. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN: 9781439840955. URL: https://books.google.ca/books?id=ZXL6AQAAQBAJ.