STAT 153 - Introduction to Time Series
Lecture Nine

Fall 2022, UC Berkeley

Aditya Guntuboyina

September 23, 2022

1 Recap: Bayesian Model Selection

We started discussing Bayesian Model Selection in the last class. Consider a generic dataset y (y could be a vector or matrix or something even more general). We have K Bayesian models for y: $M_1, \ldots, M_K$. The model M_i involves parameters denoted by θ_i and is given by

$$\text{Model } M_i : Y \mid \theta_i \sim f_{Y|\theta_i}(y) \qquad \text{with the prior } \theta_i \sim f_{\theta_i}(\theta_i). \tag{1}$$

For example, suppose we have a time series dataset $y_1, \ldots, y_n$ and we are deciding between a linear trend model and a quadratic trend model. We then have $K = 2$. Model M_1 is given by

$$Y_i \mid \alpha_0, \alpha_1, \sigma \overset{\text{independent}}{\sim} N(\alpha_0 + \alpha_1 t_i, \sigma^2) \qquad \text{for } i = 1, \ldots, n$$

with the prior

$$\alpha_0, \alpha_1, \log \sigma \sim \text{unif}(-C, C).$$

So the parameter vector θ_1 in (1) for model M_1 is $\theta_1 = (\alpha_0, \alpha_1, \sigma)$. Model M_2 is given by

$$Y_i \mid \beta_0, \beta_1, \tau \overset{\text{independent}}{\sim} N(\beta_0 + \beta_1 t_i + \beta_2 t_i^2, \tau^2) \qquad \text{for } i = 1, \ldots, n$$

with the prior

$$\beta_0, \beta_1, \beta_2, \log \tau \sim \text{unif}(-C, C).$$

So the parameter vector θ_2 in (1) for model M_2 is $\theta_2 = (\beta_0, \beta_1, \beta_2, \tau)$.

Bayesian Model Selection compares the models $M_1, \ldots, M_K$ via their average density where the average is with respect to the prior. More specifically, for each model M_i, define

$$f_{Y|M_i}(y) = \int f_{Y|\theta_i}(y) f_{\theta_i}(\theta_i) d\theta_i$$

Preference will be given to the model for which $f_{Y|M_i}(y)$ is higher (note that y is the observed data). The following are alternative names for $f_{Y|M_i}(y)$:

1. **Marginal or Integrated Likelihood**: $f_{Y|M_i}(y)$ is simply the integration of the likelihood $f_{Y|\theta_i}(y)$ with respect to the prior density $f_{\theta_i}(\theta_i)$.
2. **Evidence**: $f_{Y|M_i}(y)$ is often referred to as the Evidence of the model M_i under the observed data y.
3. **Total Probability of the Observed Data under the model** M_i.

Thus Bayesian Model Selection compares the Integrated Likelihoods or Evidences of models.

2 Hierarchical Modeling

Bayesian model selection can also be understood from the perspective of hierarchical modeling. Specifically consider the following hierarchical model which combines the K models $M_1, \ldots, M_K$ into a *single Bayesian model*:

$$\begin{aligned} &\text{This model has parameters } \mathcal{M}, \theta_1, \ldots, \theta_K \\ &\text{Likelihood is given by } Y \mid \mathcal{M} = i, \theta_1, \ldots, \theta_K \sim f_{Y|\theta_i} \text{ for } i = 1, \ldots K \\ &\text{Prior} : \mathcal{M} \sim \text{unif}\{1, \ldots, K\}, \quad \theta_1 \sim f_{\theta_1} \quad \ldots \quad \theta_K \sim f_{\theta_K} \text{ are independent.} \end{aligned} \tag{2}$$

The random variable $\mathcal{M}$ represents the models $M_1, \ldots, M_K$. More precisely $\mathcal{M} = i$ represents the model M_i. The likelihood in the above model specifies the distribution of Y given all the parameters $\mathcal{M}, \theta_1, \ldots, \theta_K$. If we want only the distribution of Y given $\mathcal{M}$, we have to integrate the other parameters $\theta_1, \ldots, \theta_K$ as follows:

$$\begin{aligned} f_{Y|\mathcal{M}=i}(y) &= \int \cdots \int f_{Y|\mathcal{M}=i,\theta_1,\ldots,\theta_K}(y) f_{\theta_1,\ldots,\theta_K|\mathcal{M}=i}(\theta_1, \ldots, \theta_K) d\theta_1 \ldots d\theta_K \\ &= \int \cdots \int f_{Y|\theta_i}(y) f_{\theta_1}(\theta_1) \ldots f_{\theta_K}(\theta_K) d\theta_1 \ldots d\theta_K \\ &= \left(\int f_{Y|\theta_i}(y) f_{\theta_i}(\theta_i) d\theta_i \right) \left[\prod_{j:j\neq i} \int f_{\theta_j}(\theta_j) d\theta_j \right] \\ &= \int f_{Y|\theta_i}(y) f_{\theta_i}(\theta_i) d\theta_i. \end{aligned}$$

Therefore in this hierarchical model, the density of Y conditional on $\mathcal{M} = i$ is simply the Evidence or Integrated Likelihood $f_{Y|M_i}(y)$ for the model M_i. The posterior of $\mathcal{M}$ given $Y = y$ is given by:

$$\begin{aligned} \mathbb{P}\{\mathcal{M} = i \mid Y = y\} &= \frac{f_{Y|\mathcal{M}=i}(y)\mathbb{P}\{\mathcal{M} = i\}}{f_Y(y)} \\ &= \frac{f_{Y|\mathcal{M}=i}(y)\mathbb{P}\{\mathcal{M} = i\}}{f_{Y|\mathcal{M}=1}(y)\mathbb{P}\{\mathcal{M} = 1\} + \cdots + f_{Y|\mathcal{M}=K}(y)\mathbb{P}\{\mathcal{M} = K\}} \\ &= \frac{f_{Y|\mathcal{M}=i}(y) \times \frac{1}{K}}{f_{Y|\mathcal{M}=1}(y) \times \frac{1}{K} + \cdots + f_{Y|\mathcal{M}=K}(y) \times \frac{1}{K}} \\ &= \frac{f_{Y|\mathcal{M}=i}(y)}{f_{Y|\mathcal{M}=1}(y) + \cdots + f_{Y|\mathcal{M}=K}(y)} \\ &= \frac{f_{Y|M_i}(y)}{f_{Y|M_1}(y) + \cdots + f_{Y|M_K}(y)}. \end{aligned}$$

Therefore the posterior probability of the model parameter $\mathcal{M}$ is simply the given by the Evidences normalized to sum to one.

To summarize, Bayesian model selection can be viewed as usual Bayesian inference in this Hierarchical Bayesian Model.

3 Connection to AIC and BIC

Model selection is usually done using criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). These are defined, for a model M, as follows:

$$AIC(M) := -2 \times (\text{maximum loglikelihood for } M) + 2 \times (\text{number of parameters in } M) \tag{3}$$

and

$$BIC(M) := -2 \times (\text{maximum loglikelihood for } M) + (\log n) \times (\text{number of parameters in } M) \tag{4}$$

Models with smaller AIC/BIC are preferred to models with larger AIC/BIC. The only difference between AIC and BIC is in the second term where AIC uses the factor 2 for the number of parametes while BIC uses the factor $\log n$ (here n is the number of observations). Because the BIC uses a more stringent penalty for model complexity, models selected using the BIC are often smaller compared to models selected using the AIC.

These model selection criteria tradeoff goodness of fit of the model with the complexity of the model. Indeed the first term measures goodness of fit to the data (it is smaller for more complex models) while the second term measures the complexity of the model (it is smaller for simpler models).

What is the connection of the Evidence-based Bayesian Model Selection to criteria such as AIC and BIC? It turns out that the Evidence $f_{Y|M}(y)$ for a model M can be written in a form that bears some similarities to (3) and (4). To obtain this alternative formula, we use the following expression for the posterior density of the parameter θ in the model M:

$$\text{posterior}(\theta) = \frac{\text{prior}(\theta) f_{Y|\theta}(y)}{f_{Y|M}(y)} \qquad \text{for every } \theta.$$

Here $\text{prior}(\theta) = f_\theta(\theta)$ and $\text{posterior}(\theta)$ is the density of θ conditional on $Y = y$ in the model M. As a result, we have

$$f_{Y|M}(y) = \frac{\text{prior}(\theta) f_{Y|\theta}(y)}{\text{posterior}(\theta)} \qquad \text{for every } \theta. \tag{5}$$

Note that the left hand side of the expression does not depend on θ while the right hand side depends on θ. It is true for every choice of θ. Taking θ to be the MLE $\hat{\theta}$ in the model M, we obtain

$$f_{Y|M}(y) = \frac{\text{prior}(\hat{\theta}) f_{Y|\hat{\theta}}(y)}{\text{posterior}(\hat{\theta})}.$$

This immediately gives the formula:

$$-2 \log f_{Y|M}(y) = -2 \log f_{Y|\hat{\theta}}(y) + 2 \log \left[\frac{\text{posterior}(\hat{\theta})}{\text{prior}(\hat{\theta})} \right]$$

Observe that $\log f_{Y|\hat{\theta}}(y)$ is simply the maximum log-likelihood for the model M. Thus

$$-2 \log (\text{Evidence}(M)) = -2 \times (\text{Maximum log-likelihood for } M) + 2 \log \left[\frac{\text{posterior}(\hat{\theta})}{\text{prior}(\hat{\theta})} \right] \tag{6}$$

Note the similarity of (6) with (3) and (4). The first term above measures the fit of the best parameter configuration in M to the observed data, while the second term measures model complexity. The model complexity term is more complicated compared to (3) and (4). The larger the posterior at the MLE is compared to the prior at the MLE, the more the complexity term. For complex models, the term $\text{prior}(\hat{\theta})$ will be quite small which contributes to the second (model complexity) term in (6) being large.

3.1 The BIC as an approximation of (6)

The BIC (Bayesian Information Criterion) is derived as an approximation for (6) when the posterior is replaced by a normal approximation. In many cases, the posterior distribution is well approximated by a normal distribution $N_p(\hat{\theta}, \Sigma/n)$ where $\hat{\theta}$ is the MLE, n denotes sample size and Σ is a $p \times p$ covariance matrix (generally Σ is related to the Hessian of the log-likelihood evaluated at $\hat{\theta}$). Therefore

$$\text{posterior}(\theta) \approx (2\pi)^{-p/2} \left(\det(\Sigma/n)\right)^{-1/2} \exp\left(-\frac{n}{2}(\theta - \hat{\theta})'\Sigma^{-1}(\theta - \hat{\theta})\right)$$

which implies that

$$\text{posterior}(\hat{\theta}) = (2\pi)^{-p/2} \left(\det(\Sigma/n)\right)^{-1/2}.$$

As a result

$$\begin{aligned}
\log\left[\frac{\text{posterior}(\hat{\theta})}{\text{prior}(\hat{\theta})}\right] &= \log \frac{(2\pi)^{-p/2} \left(\det(\Sigma/n)\right)^{-1/2}}{\text{prior}(\hat{\theta})} \\
&= \log \frac{(2\pi)^{-p/2} n^{p/2} \left(\det(\Sigma)\right)^{-1/2}}{\text{prior}(\hat{\theta})} \\
&= \frac{p}{2}(\log n) - \frac{p}{2}(\log(2\pi)) - \frac{1}{2}\log\det\Sigma - \log\text{prior}(\hat{\theta}) \\
&= \frac{p}{2}(\log n)\left\{1 - \frac{\frac{p}{2}(\log(2\pi)) + \frac{1}{2}\log\det\Sigma + \log\text{prior}(\hat{\theta})}{\frac{p}{2}(\log n)}\right\}.
\end{aligned}$$

Now if the sum of the terms $\frac{p}{2}(\log(2\pi))$, $\frac{1}{2}\log\det\Sigma$ and $\log\text{prior}(\hat{\theta})$ is small compared to $\frac{p}{2}\log n$:

$$\left|\frac{\frac{p}{2}(\log(2\pi)) + \frac{1}{2}\log\det\Sigma + \log\text{prior}(\hat{\theta})}{\frac{p}{2}(\log n)}\right| \ll 1, \tag{7}$$

then we can approximate the term in the parantheses by just 1 leading to

$$\log\left[\frac{\text{posterior}(\hat{\theta})}{\text{prior}(\hat{\theta})}\right] \approx \frac{p}{2}(\log n).$$

The formula (6) then simplifies to

$$-2\log\left(\text{Evidence}(M)\right) \approx -2 \times (\text{Maximized log-likelihood for } M) + p\log n \tag{8}$$

The right hand side above is the BIC (see (4)). Also note that because of the assumption (7), the formula (8) does not depend on the prior making this convenient to use in practice.

4 Connection to Generalization Accuracy

The Evidence for a model also has some connection to estimates of generalization accuracy of the model. To see this, suppose that the dataset Y is made of n observations $Y_1, \ldots, Y_n$. Then we can decompose the Evidence as

$$\begin{aligned}
\text{Evidence}(M) &:= f_{Y_1,\ldots,Y_n|M}(y_1, \ldots, y_n) \\
&= f_{Y_1|M}(y_1) f_{Y_2|Y_1=y_1,M}(y_2) f_{Y_3|Y_1=y_1,Y_2=y_2,M}(y_3) \ldots f_{Y_n|Y_1=y_1,\ldots,Y_{n-1}=y_{n-1},M}(y_n).
\end{aligned}$$

Thus the Evidence is simply the product of all the predictive probabilities for each of the data points, using the model "trained" on the previous data points. Note that

$$f_{Y_i|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},M}(y_i) = \int f_{Y_i|\theta}(y_i) f_{\theta|Y_1=y_1\ldots,Y_{i-1}=y_{i-1}}(\theta) d\theta$$

assuming that $Y_1, \ldots, Y_n$ are independent conditional on θ. When i is not too small, the posterior density $f_{\theta|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}}(\theta)$ should be peaked near the MLE based on the data $Y_1, \ldots, Y_{i-1}$. In this case,

$$f_{Y_i|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},M}(y_i) = \int f_{Y_i|\theta}(y_i) f_{\theta|Y_1=y_1\ldots,Y_{i-1}=y_{i-1}}(\theta) d\theta \approx f_{Y_i|\hat{\theta}_i}(y_i)$$

where $\hat{\theta}_i$ is the MLE of θ based on $y_1, \ldots, y_{i-1}$. Therefore, $f_{Y_i|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},M}(y_i)$ is related to the Generalization Accuracy of the MLE based on $y_1, \ldots, y_{i-1}$ at the new data point y_i. The Evidence therefore is the product of these of these generalization accuracies. For more on the connection between the Evidence and Test Error methods such as Cross-Validation, see `http://www.inference.org.uk/mackay/Bayes_FAQ.html`.

5 Recommended List of Readings for Today

1. For a very good treatment of Bayesian Model Comparison, see Chapter 28 of the book *Information Theory, Inference and Learning Algorithms* by David MacKay, or Chapter 20 of the book *Probability Theory: the logic of science* by E. T. Jaynes.

2. The formula (6) can be found in the 2010 paper titled *Bayesian system identification based on probability logic* by James L. Beck.