STAT 153 - Introduction to Time Series Lecture Eight

Fall 2022, UC Berkeley

Aditya Guntuboyina

September 21, 2022

The question of model selection arises very frequently in time series modeling. For example, (a) should one fit a linear trend model or a quadratic trend model, (b) should one fit a single sinusoidal model or have two sinusoidal components. In the next few lectures, we shall go over some general principles for model selection. There are broadly two main approaches for model selection:

1. **Test Error Evaluation**: Here one splits the existing dataset into two parts, use one part for fitting each model and the other part for evaluating the accuracy of the fitted model. There are many ways of actually doing this. While this methodology is popular, there are some arbitrary aspects to this. For example, it is not quite clear how to precisely do the training/test data split. If the test dataset is too small, then the test error estimates will be noisy. On other other hand, if the training dataset is too small, the performance of the model fitted on the training dataset will be quite different from its performance on the full dataset. It should be noted that the popular model selection method based on AIC is also based on test error evaluation but, to derive the AIC, one uses asymptotics as opposed to mechanistically creating a test dataset. We shall not spend much time on these approaches in the next few lectures but we shall discuss them later.

2. **Bayesian Model Selection**: If the models that are being considered are "Bayesian Models" which means that they are equipped with priors, then model selection can be done by comparing the probability that each model assigns to the observed data. This approach is called "Bayesian Model Selection". It is principled and conceptually quite simple. The downside is the specification of the prior and the arbitrariness that comes along with the prior specification. Nevertheless, in standard situations, this approach can be made to work with certain default prior choices and the resulting methods work very well. We shall go over this approach in more details in the next few lectures.

1 Bayesian Model Selection

Bayesian model selection works for comparing Bayesian models. By a Bayesian model, we mean a model in which both the likelihood as well as the prior are specified. For example, given data $y_1, \dots, y_n$, consider the two models:

$$Y_1, \dots, Y_n \overset{\text{i.i.d}}{\sim} N(\theta, 1) \qquad \text{with } \theta \in [-5, 5], \tag{1}$$

and

$$Y_1, \dots, Y_n \overset{\text{i.i.d}}{\sim} N(\theta, 1) \qquad \text{with } \theta \sim \text{unif}[-5, 5]. \tag{2}$$

Model (1) is not a Bayesian model because the prior is not specified. The constraint $\theta \in [-5, 5]$ does not precisely say how θ is distributed on $[-5, 5]$. On the other hand, the model (2) is a Bayesian model.

An important advantage of Bayesian models is that they allow calculation of the probability of the observed data under the model. For example, the Bayesian model (2) would calculate the probability of the observed data $y_1, \dots, y_n$ as

$$\frac{1}{10} \int_{-5}^{5} (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \theta)^2 \right) d\theta.$$

On the other hand, the non-Bayesian model (1) would not allow computation of the probability of the observed dataset. Indeed, under the model (1), one can write the probability of the observed data as

$$(2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \theta)^2 \right)$$

for some $\theta \in [-5, 5]$. But this not give a precise answer to the probability of the observed data as it involves the unknown value θ about which we only know that $\theta \in [-5, 5]$.

Note the slight abuse of terminology here. By probability of the observed data under a model, I actually mean the joint density:

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$$

when the underlying random variables are continuous. In the case where the random variables are discrete, probability of the observed data will mean

$$\mathbb{P}\{Y_1 = y_1, \dots, Y_n = y_n\}.$$

In the continuous case, one really should think of an observation 1.29 as not being exactly equal to the number 1.29 but rather as $[1.29 - \delta, 1.29 + \delta]$ for some very small number δ which represents recording precision. The observed dataset $y_1, \dots, y_n$ is then really $[y_1 - \delta, y_1 + \delta], \dots, [y_n - \delta, y_n + \delta]$. In such a case, the probability of the observed dataset will be represented by

$$\mathbb{P}\{Y_1 \in [y_1 - \delta, y_1 + \delta], \dots, Y_n \in [y_n - \delta, y_n + \delta]\} \approx f_{Y_1, \dots, Y_n}(y_1, \dots, y_n)(2\delta)^n.$$

Thus, up to multiplication by the constant factor $(2\delta)^n$ (which will be the same across different models), the probability of the observed dataset is proportional to the joint density. This justifies the abuse of notation referring to the joint density as the probability of the observed dataset.

Consider now a generic dataset y (y could be a vector or matrix or something even more general). We have two Bayesian models for y:

$$M_1 : Y \mid \theta \sim p_\theta \qquad \text{with } \theta \sim f_\theta(\cdot) \tag{3}$$

and

$$M_2 : Y \mid \alpha \sim q_\alpha \qquad \text{with } \alpha \sim f_\alpha(\cdot) \tag{4}$$

Bayesian Model Selection compares M_1 and M_2 by simply calculating the probabiliity of the observed data y under both M_1 and M_2. Specifically, we compare

$$f_{Y|M_1}(y) = \int p_\theta(y) f_\theta(\theta) d\theta \quad \text{and} \quad f_{Y|M_2}(y) = \int q_\alpha(y) f_\alpha(\alpha) d\alpha.$$

Preference will be given to the model for which the probability of observed data is higher. The following are alternative terms for $f_{Y|M_1}(y)$:

1. **Marginal or Integrated Likelihood**: $f_{Y|M_1}(y)$ is simply the integration of the likelihood $p_\theta(y)$ with respect to the prior density $f_\theta(\theta)$.
2. **Evidence**: $f_{Y|M_1}(y)$ is often referred to as the Evidence of the model M_1 under the observed data y.

Thus Bayesian Model Selection compares the Integrated Likelihoods or Evidences of models. The following simple example is a good illustration of the basic idea behind Bayesian Model Selection.

Example 1.1 (MacKay)**.** *This example is from Chapter 28 of David MacKay's book titled Information Theory, Inference, and Learning Algorithms. We have the dataset* $-1, 3, 7, 11$*. Consider the following two Bayesian models for this dataset:*

1. ***Model 1 (linear):*** $Y_1 = \alpha$ *and* $Y_{n+1} = Y_n + \beta$ *for* $n \geq 1$*. This model has the two parameters* α *and* β*. We assume that* α *and* β *are integer-valued that they are independently uniformly distributed over the set* $\{-50, -49, \dots, 49, 50\}$ *which has cardinality 101.*
2. ***Model 2 (cubic):*** $Y_1 = a$ *and* $Y_{n+1} = bY_n^3 + cY_n^2 + d$*. This model has the four parameters* a, b, c, d*. We assume that these four parameters are independent with* a *having the uniform on* $\{-50, -49, \dots, 49, 50\}$ *and* b, c, d *each having the distribution of* x/y *where* $x \sim$ *Unif*$\{-50, -49, \dots, 49, 50\}$ *and* $y \sim$ *Unif*$\{1, \dots, 50\}$ *are independent.*

Which of these two models would you use for the data? Bayesian model selection is readily applicable here as both the models are Bayesian. We only need to calculate the probability of the observed data for the two models. For the linear model (M1):

$$\begin{aligned}
&\mathbb{P}\{Y_1 = -1, Y_2 = 3, Y_3 = 7, Y_4 = 11 \mid \mathrm{M1}\} \\
&= \sum_{i,j} \mathbb{P}\{Y_1 = -1, Y_2 = 3, Y_3 = 7, Y_4 = 11 \mid \alpha = i, \beta = j, \mathrm{M1}\}\mathbb{P}\{\alpha = i, \beta = j \mid \mathrm{M1}\} \\
&= \mathbb{P}\{Y_1 = -1, Y_2 = 3, Y_3 = 7, Y_4 = 11 \mid \alpha = -1, \beta = 4, \mathrm{M1}\}\mathbb{P}\{\alpha = -1, \beta = 4 \mid \mathrm{M1}\} \\
&= (1)\mathbb{P}\{\alpha = -1 \mid \mathrm{M1}\}\mathbb{P}\{\beta = 4 \mid \mathrm{M1}\} = \left(\frac{1}{101}\right)^2 = 9.8 \times 10^{-5}.
\end{aligned}$$

For the cubic model:

$$\begin{aligned}
&\mathbb{P}\{Y_1 = -1, Y_2 = 3, Y_3 = 7, Y_4 = 11 \mid \mathrm{M2}\} \\
&= \sum_{a,b,c,d} \mathbb{P}\{Y_1 = -1, Y_2 = 3, Y_3 = 7, Y_4 = 11 \mid a, b, c, d, \mathrm{M2}\}\mathbb{P}\{a = a, b = b, c = c, d = d \mid \mathrm{M2}\}
\end{aligned}$$

It turns out that the cubic model explains the given data perfectly if and only if its four parameters a, b, c, d *are chosen as* $a = -1, b = -1/11, c = 9/11, d = 23/11$*. As a result*

$$\begin{aligned}
&\mathbb{P}\{Y_1 = -1, Y_2 = 3, Y_3 = 7, Y_4 = 11 \mid \mathrm{M2}\} \\
&= \mathbb{P}\left\{a = -1, b = -1/11, c = 9/11, d = 23/11 \mid M2\right\} \\
&= \mathbb{P}\{a = -1\}\mathbb{P}\{b = -1/11\}\mathbb{P}\{c = 9/11\}\mathbb{P}\{d = 23/11\} \\
&= \left(\frac{1}{101}\right)\left(4 \cdot \frac{1}{101} \cdot \frac{1}{50}\right)\left(4 \cdot \frac{1}{101} \cdot \frac{1}{50}\right)\left(2 \cdot \frac{1}{101} \cdot \frac{1}{50}\right) = 2.5 \times 10^{-12}.
\end{aligned}$$

Clearly the probability of the observed data is much smaller for the cubic model compared to the simpler linear model. Bayesian model selection here will prefer the linear model and this would align with common sense. Note here both the models explain the data equally well. The cubic model gets downgraded however because the prior in the cubic model gives a much smaller probability to the correct parameter values compared to the linear model.

Example 1.2. We have the dataset consisting of the $n = 6$ observations 26.6, 38.5, 34.4, 34, 31, 23.6. Consider the following two Bayesian models for this dataset:

1. ***Model 1 (Normal)***: $X_1, \dots, X_n \overset{i.i.d}{\sim} N(\theta, \sigma^2)$. This mdoel has the two parameters θ and σ. For the prior, we assume
$$\theta \sim Unif(-C_1, C_1) \quad \text{and} \quad \log \sigma \sim Unif(-C_2, C_2). \tag{5}$$
Here C_1 and C_2 are large constants.

2. ***Model 2 (Laplace)***: $X_1, \dots, X_n \overset{i.i.d}{\sim} Lap(\theta, \sigma)$. Recall that the $Lap(\theta, \sigma)$ distribution is given by the density
$$x \mapsto \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right) \tag{6}$$
This model also has the two parameters θ and σ and we use exactly the same prior (5).

Bayesian model selection then proceeds by calculating the **integrated** joint density at the observed data under each of the two models. For the first model, we need to calculate

$$\begin{aligned} E_1 &:= \int\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\right) \frac{I\{-C_1 < \theta < C_1\}}{2C_1} \frac{I\{e^{-C_2} < \sigma < e^{C_2}\}}{2C_2\sigma} d\theta d\sigma \\ &= \frac{1}{4C_1C_2} \int_{-C_1}^{C_1} \int_{e^{-C_2}}^{e^{C_2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\right) \frac{1}{\sigma} d\sigma d\theta \\ &\approx \frac{1}{4C_1C_2} \int_{-C_1}^{C_1} \int_{0}^{\infty} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\right) \frac{1}{\sigma} d\sigma d\theta \end{aligned}$$

Both these integrals can be calculated in closed form. For correspondence with the calculation for the second model however, let us only evaluate the integral over σ in closed form. By the change of variable:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \theta)^2}{2v}},$$

the inner integral can be written in terms of the Gamma function $\Gamma(n/2) := \int_0^\infty v^{(n/2)-1} e^{-v} dv$. This gives

$$E_1 = \frac{1}{4C_1C_2} \frac{\pi^{-n/2}\Gamma(n/2)}{2} \int_{-C_1}^{C_1} \frac{d\theta}{\left(\sum_{i=1}^n (x_i - \theta)^2\right)^{n/2}}$$

The one-dimensional integral above can be calculated numerically (it can also be evaluated exactly when $C_1 = \infty$).

For the second model, we need to calculate

$$\begin{aligned} E_2 &:= \int\int \left(\frac{1}{2\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n |x_i - \theta|}{\sigma}\right) \frac{I\{-C_1 < \theta < C_1\}}{2C_1} \frac{I\{e^{-C_2} < \sigma < e^{C_2}\}}{2C_2\sigma} d\theta d\sigma \\ &= \frac{1}{4C_1C_2} \int_{-C_1}^{C_1} \int_{e^{-C_2}}^{e^{C_2}} \left(\frac{1}{2\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n |x_i - \theta|}{\sigma}\right) \frac{1}{\sigma} d\sigma d\theta \\ &\approx \frac{1}{4C_1C_2} \int_{-C_1}^{C_1} \int_{0}^{\infty} \left(\frac{1}{2\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n |(x_i - \theta|}{\sigma}\right) \frac{1}{\sigma} d\sigma d\theta. \end{aligned}$$

The inner integral over σ can be evaluated exactly by the change of variable

$$v = \frac{\sum_{i=1}^n |(x_i - \theta|}{\sigma}$$

and then relating to the Gamma function. This gives

$$E_2 = \frac{1}{4C_1C_2} \left(2^{-n}\Gamma(n)\right) \int_{-C_1}^{C_1} \frac{d\theta}{\left(\sum_{i=1}^n |x_i - \theta|\right)^n}.$$

The one-dimensional integral above has to be evaluated numerically (it may not be possible to evaluate it in closed form).

The integrated likelihoods of the two models E_1 and E_2 will be compared with each other. The model with the higher value will be preferred.

Note that instead of calculating the integral over σ as above, we could have directly evaluated the double integral numerically by gridding the range of both σ and θ. This can be done although it will slightly more computationally intensive and should give the same answers as the method involving numerical integration over only θ.

2 Hierarchical Modeling

Bayesian model selection can also be understood from the perspective of hierarchical modeling. Specifically consider the following hierarchical model which converts the two models M_1 and M_2 (defined as in (3) and (4) respectively) into a *single Bayesian model.*

$$\begin{aligned} &\mathcal{M} \text{ takes the values 1 and 2 with probabilities } \rho \text{ and } 1-\rho \\ &Y \mid \mathcal{M} = 1, \theta \sim p_\theta \quad \text{and} \quad \theta \mid \mathcal{M} = 1 \sim f_\theta \\ &Y \mid \mathcal{M} = 2, \theta \sim q_\alpha \quad \text{and} \quad \alpha \mid \mathcal{M} = 2 \sim f_\alpha \end{aligned} \tag{7}$$

The random variable $\mathcal{M}$ represents one of the two models M_1 and M_2. More precisely $\mathcal{M} = 1$ represents M_1 and $\mathcal{I} = 2$ represents M_2. ρ and $1-\rho$ represent the prior probabilities of M_1 and M_2. Under this single Bayesian model, we can calculate the posterior distribution of $\mathcal{M}$ given the data $Y = y$ as:

$$\mathbb{P}\left\{\mathcal{M} = 1 \mid Y = y\right\} = \frac{f_{Y|M_1}(y)\mathbb{P}\{\mathcal{M} = 1\}}{f_{Y|M_1}(y)\mathbb{P}\{\mathcal{M} = 1\} + f_{Y|M_2}(y)\mathbb{P}\{\mathcal{M} = 2\}}$$

and

$$\mathbb{P}\left\{\mathcal{M} = 2 \mid Y = y\right\} = \frac{f_{Y|M_2}(y)\mathbb{P}\{\mathcal{M} = 2\}}{f_{Y|M_1}(y)\mathbb{P}\{\mathcal{M} = 1\} + f_{Y|M_2}(y)\mathbb{P}\{\mathcal{M} = 2\}}.$$

These are the posterior probabilities of the two models given the data $Y = y$. Model M_1 will be preferred compared to Model M_2 if and only if

$$\mathbb{P}\left\{\mathcal{M} = 1 \mid Y = y\right\} > \mathbb{P}\left\{\mathcal{M} = 2 \mid Y = y\right\}.$$

As the denominators of the above probabilities are the same, this is equivalent to

$$f_{Y|M_1}(y)\mathbb{P}\{\mathcal{M} = 1\} > f_{Y|M_2}(y)\mathbb{P}\{\mathcal{M} = 2\}.$$

Now if $\mathbb{P}\{\mathcal{M} = 1\} = \mathbb{P}\{\mathcal{M} = 2\}$ i.e., if the two models are *a priori* equally likely, then the above comparison is equivalent to comparing $f_{Y|M_1}(y)$ and $f_{Y|M_2}(y)$. Thus Bayesian model selection in terms of evidences is equivalent to looking at posterior probabilities of the two models in a hierarchical model where the prior probabilities are the same. When the prior model probabilities are not the same, we need to multiply the evidences by the prior probabilities before evaluating the models.

3 Recommended List of Readings for Today

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.