STAT 153 - Introduction to Time Series Lecture Ten

Fall 2022, UC Berkeley

Aditya Guntuboyina

September 28, 2022

1 A Model Selection Problem

We have a time series dataset $(t_1, y_1), \dots, (t_n, y_n)$. Consider K possible models $M_1, \dots, M_K$ for this dataset where M_k is given by:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \left[\beta_{2j-1} \cos\left(\frac{2\pi j t_i}{n}\right) + \beta_{2j} \sin\left(\frac{2\pi j t_i}{n}\right) \right] + Z_i \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

In other words, the model M_k uses k sinusoids at the Fourier frequencies $1/n, 2/n, \dots, k/n$.

How do we decide which model among $M_1, \dots, M_K$ is best for this dataset? Bayesian model selection solves this problem by comparing the Evidences of each model M_k for the given dataset. The evidence is simply the averaged likelihood where the average is with respect a prior. The likelihood of model M_k is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \left[\beta_{2j-1} \cos\left(\frac{2\pi j t_i}{n}\right) + \beta_{2j} \sin\left(\frac{2\pi j t_i}{n}\right)\right]\right)^2\right).$$

A more compact expression for the likelihood can be written in matrix form as:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) \tag{1}$$

where Y is the column vector containing $y_1, \dots, y_n$, X is the $n \times p$ matrix ($p = 2k+1$) whose i^{th} row is given by

$$1, \cos\left(\frac{2\pi j t_i}{n}\right), \sin\left(\frac{2\pi j t_i}{n}\right), j = 1, \dots, k$$

and β is the $p \times 1$ vector with entries $\beta_0, \beta_1, \dots, \beta_p$. To calculate the evidence of M_k, we need a specification of the prior. Consider the prior:

$$\beta_0, \beta_1, \dots, \beta_{2k}, \log\sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C) \tag{2}$$

for a large C. The Evidence is then given by

$$\begin{aligned}
\text{Evidence}(M_k) &= \int\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) \frac{I\{-C < \beta_j, \log\sigma < C\}}{\sigma(2C)^{p+1}} d\beta d\sigma \\
&\approx \left(\frac{1}{2C}\right)^{p+1} \left(\frac{1}{2\pi}\right)^{n/2} \int_0^\infty \int_{\mathbb{R}^p} \sigma^{-n-1} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) d\beta d\sigma.
\end{aligned}$$

For the integral over β, we write

$$\begin{aligned}\|Y - X\beta\|^2 &= \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\beta\|^2 \\ &= \|Y - X\hat{\beta}\|^2 + \left(\beta - \hat{\beta}\right)^T X^T X \left(\beta - \hat{\beta}\right).\end{aligned}$$

The Evidence of M_k is then

$$\left(\frac{1}{2C}\right)^{p+1} \left(\frac{1}{2\pi}\right)^{n/2} \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{\|Y - X\hat{\beta}\|^2}{2\sigma^2}\right) \int_{\mathbb{R}^p} \exp\left(-\frac{(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}))}{2\sigma^2}\right) d\beta d\sigma$$

$$= \left(\frac{1}{2C}\right)^{p+1} \left(\frac{1}{2\pi}\right)^{n/2} \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{\|Y - X\hat{\beta}\|^2}{2\sigma^2}\right) (2\pi)^{p/2} \sqrt{\det(\sigma^2 (X^T X)^{-1})} d\sigma$$

$$= \left(\frac{1}{2C}\right)^{p+1} \left(\frac{1}{2\pi}\right)^{(n-p)/2} |X^T X|^{-1/2} \int_0^\infty \sigma^{-n+p-1} \exp\left(-\frac{\|Y - X\hat{\beta}\|^2}{2\sigma^2}\right) d\sigma.$$

The integral over σ can be related to the Gamma function by the change of variable:

$$t = \frac{\|Y - X\hat{\beta}\|^2}{2\sigma^2} \quad \text{or} \quad \sigma = \frac{\|Y - X\hat{\beta}\|}{\sqrt{2t}} \quad \text{so} \quad d\sigma = \frac{-\|Y - X\hat{\beta}\|}{2\sqrt{2}} t^{-3/2} dt.$$

As a result,

$$\begin{aligned}&\int_0^\infty \sigma^{-n+p-1} \exp\left(-\frac{\|Y - X\hat{\beta}\|^2}{2\sigma^2}\right) d\sigma \\ &= \frac{2^{(n-p-2)/2}}{\|Y - X\hat{\beta}\|^{n-p}} \int_0^\infty t^{(n-p-2)/2} e^{-t} dt = \frac{2^{(n-p-2)/2}}{\|Y - X\hat{\beta}\|^{n-p}} \Gamma\left(\frac{n-p}{2}\right).\end{aligned}$$

Therefore

$$\begin{aligned}\text{Evidence}(M_k) &= \left(\frac{1}{2C}\right)^{p+1} \left(\frac{1}{2\pi}\right)^{(n-p)/2} |X^T X|^{-1/2} \frac{2^{(n-p-2)/2}}{\|Y - X\hat{\beta}\|^{n-p}} \Gamma\left(\frac{n-p}{2}\right) \\ &= \frac{1}{2} \left(\frac{1}{2C}\right)^{p+1} \frac{|X^T X|^{-1/2}}{\pi^{(n-p)/2}} \frac{1}{\|Y - X\hat{\beta}\|^{n-p}} \Gamma\left(\frac{n-p}{2}\right). \qquad (3)\end{aligned}$$

It should be noted that the above calculation is very similar to an integral calculation in Lecture 4. But there the matrix X depended on another parameter (such as frequency f), and we could afford to treat p as a constant so that the above expression can be written as being proportional to $|X'X|^{-1/2} \|Y - X\hat{\beta}\|^{p-n}$. Now the dependence on p is important as the different models $M_1, \ldots, M_K$ have different values of p (in fact, $p = 2k + 1$ in model M_k).

We calculate Evidence(M_k) for each $k = 1, \ldots, K$, and then normalize the evidences so that they sum to one. These normalized evidences should be viewed as posterior probabilities for the models $M_1, \ldots, M_K$ given the observed data.

The choice of C (in the prior) is quite crucial for this method. While C will never be chosen to be small, values of C that are too large will heavily penalize larger models (i.e., models with large k). We will say more about the choice of C later.

2 AIC and BIC for the model M_k

A different way of selecting among the models $M_1, \ldots, M_K$ is via the AIC and BIC. As we saw in the last class, for a model M, AIC and BIC are defined as follows:

$$AIC(M) := -2 \times (\text{maximum loglikelihood for } M) + 2 \times (\text{number of parameters in } M) \quad (4)$$

and

$$BIC(M) := -2 \times (\text{maximum loglikelihood for } M) + (\log n) \times (\text{number of parameters in } M) \tag{5}$$

Models with smaller AIC/BIC are preferred to models with larger AIC/BIC. It is quite easy to calculate AIC and BIC for the models M_k of the last section. We only need to calculate the maximum likelihood for M_k. The likelihood is given by (1) so the log-likelihood is:

$$\text{log-likelihood}(M_k) = -\frac{\|Y - X\beta\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2).$$

By minimizing this function with respect to β and σ (just take derivatives with respect to β and σ^2, set them to zero and solve the resulting equations), one can find the Maximum Likelihood Estimators:

$$\hat{\beta} = \text{least squares estimate} = (X^TX)^{-1}X^TY \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\|Y - X\hat{\beta}\|^2.$$

Thus the maximum log-likelihood is given by

$$\begin{aligned}\text{maximum log-likelihood}(M_k) &= -\frac{\|Y - X\hat{\beta}\|^2}{2\hat{\sigma}^2} - \frac{n}{2}\log(2\pi\hat{\sigma}^2)\\ &= -\frac{n}{2} - \frac{n}{2}\log\left(\frac{2\pi}{n}\|Y - X\hat{\beta}\|^2\right).\end{aligned}$$

Then

$$(-2) \times \text{maximum log-likelihood}(M_k) = n + n\log\left(\frac{2\pi}{n}\|Y - X\hat{\beta}\|^2\right).$$

AIC and BIC are thus given by (note that the number of parameters in M_k is $p+1$ where $p = 2k+1$)

$$AIC(M_k) = n + n\log\left(\frac{2\pi}{n}\|Y - X\hat{\beta}\|^2\right) + 2(p+1)$$

and

$$BIC(M_k) = n + n\log\left(\frac{2\pi}{n}\|Y - X\hat{\beta}\|^2\right) + (\log n)(p+1).$$

3 More on Evidence Calculation

Let us get back to the Evidence calculation for the model M_k in Section 1. Here is a slightly different way of doing this calculation that is more insightful and it is also applicable for priors different from (2). Let us start with

$$\text{Evidence} = \int \text{likelihood}(\theta) \times \text{prior}(\theta) d\theta$$

Generally, the likelihood will be peaked in a small region of the parameter space. On the other hand, the prior will be chosen to be uninformative so it will be comparatively flat in the region of concentration of the likelihood. As a result,

$$\text{Evidence} \approx \text{prior}(\hat{\theta}) \int \text{likelihood}(\theta) d\theta$$

where $\hat{\theta}$ is the MLE. For the model M_k, we have $\theta = (\beta, \sigma)$ and the likelihood is given by (1) so we can write

$$\text{Evidence}(M_k) \approx \text{prior}(\hat{\theta}) \int\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) d\beta d\sigma.$$

This integral can be written in closed form (and the derivation of this expression follows essentially the same argument that we used to derive (3)) to get

$$\text{Evidence}(M_k) \approx \text{prior}(\hat{\theta}) \frac{1}{2\sqrt{2}} \frac{|X^T X|^{-1/2}}{\pi^{(n-p)/2}} \frac{1}{\|Y - X\hat{\beta}\|^{n-p-1}} \Gamma\left(\frac{n-p-1}{2}\right). \tag{6}$$

This formula is very similar to (3) and they both lead to almost the same answers.

From the formula (6), it is clear that the dependence of the prior on the evidence is only through the value of the prior density at the MLE $\hat{\theta}$. For the case of the uniform prior, this is given by $\left(\frac{1}{2C}\right)^{p+1}$. In the next class, we shall work with alternative priors that better take into account the inherent scaling present in the problem; this will also give us a solution which does not depend on the somewhat arbitrary constant C.

4 Recommended List of Readings for Today

Two very good references for Bayesian model selection in the context of linear models are

1. Chapter 5 of the book *Bayesian spectrum analysis and parameter estimation* by Larry Bretthorst, and

2. Section 9.3.1 of the book *A first course in Bayesian Statistical Methods* by Peter Hoff.

Both these references do not work with the uniform prior and instead work with Gaussian priors. We shall discuss this in the next class.