STAT 153 - Introduction to Time Series Lecture Twelve

Fall 2022, UC Berkeley

Aditya Guntuboyina

October 6, 2022

1 Model Selection in some Nonlinear Regression Models

In the last lecture, we studied Bayesian Model Selection in linear regression models. Today, we shall extend the analysis to certain nonlinear regression models. Our prototypical example is the problem of determining the number of sinusoids (with unknown frequencies) to fit to the data. For instance, consider the problem of choosing between the two models M_1 and M_2 where M_1 is given by

$$Y_i = \beta_0 + \beta_1 \cos(\omega_1 t_i) + \beta_2 \sin(\omega_1 t_i) + Z_i \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

and M_2 is given by

$$Y_i = \beta_0 + \beta_1 \cos(\omega_1 t_i) + \beta_2 \sin(\omega_1 t_i) + \beta_3 \cos(\omega_2 t_i) + \beta_4 \sin(\omega_2 t_i) + Z_i \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

The model M_1 fits a single sinuosoid at an unknown angular frequency ω_1 while M_2 fits two sinusoids at unknown angular frequencies ω_1 and ω_2. You may recall that an angular frequency ω is related to the usual frequency f via $\omega = 2\pi f$. If ω_1 in model M_1 and (ω_1, ω_2) in model M_2 are known, then this reduces to comparing between two linear models. But now we consider them as unknown parameters.

As we have seen previously (say in Lecture 7), the models M_1 and M_2 can be represented in general form as:

$$Y = X(\omega)\beta + Z \tag{1}$$

where $X(\omega)$ is the design matrix. Here ω is the parameter representing the nonlinear parameters. For example, in model M_1, we have $\omega = \omega_1$ and in model M_2, we have $\omega = (\omega_1, \omega_2)$. The number of columns of $X(\omega)$ is p and this is the number of β-parameters in the model. For example, the value of p for the model M_1 equals 3 and the value of p in model M_2 equals 5. In general, if we are considering a model with k sinusoids, we would have $p = 2k + 1$.

In order to compare these models, we would need to calculate their Evidences. Here is how the Evidence of model (1) can be calculated in the general case. We use the formula

$$\text{Evidence} = \int \text{likelihood}(\theta) \times \text{prior}(\theta) d\theta \approx \text{prior}(\hat{\theta}) \times \int \text{likelihood}(\theta) d\theta. \tag{2}$$

Here $\theta = (\omega, \beta, \sigma)$ and $\hat{\theta}$ is the Maximum Likelihood Estimator. The likelihood is given by

$$\text{likelihood}(\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right)$$

Thus

$$\int \text{likelihood}(\theta) d\theta = \int\int\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right) d\beta d\omega d\sigma.$$

Integration over β (for fixed ω and σ) is just as in the case of the linear regression model. We write

$$\|Y - X(\omega)\beta\|^2 = \|Y - X(\omega)\hat{\beta}(\omega)\|^2 + \left(\beta - \hat{\beta}(\omega)\right)^T X(\omega)^T X(\omega) \left(\beta - \hat{\beta}(\omega)\right)$$

where $\hat{\beta}(\omega)$ is the least squares estimator for fixed ω:

$$\hat{\beta}(\omega) = \text{minimizer of } \|Y - X(\omega)\beta\|^2 = \left(X(\omega)^T X(\omega)\right)^{-1} X(\omega)^T Y.$$

We then get

$$\begin{aligned}
&\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right) d\beta \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \int \exp\left(-\frac{\left(\beta - \hat{\beta}(\omega)\right)^T X(\omega)^T X(\omega) \left(\beta - \hat{\beta}(\omega)\right)}{2\sigma^2}\right) d\beta \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \left(\sqrt{2\pi}\sigma\right)^p |X(\omega)^T X(\omega)|^{-1/2} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) |X(\omega)^T X(\omega)|^{-1/2}.
\end{aligned}$$

As a result

$$\int \text{likelihood}(\theta) d\theta = \int\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) |X(\omega)^T X(\omega)|^{-1/2} d\omega d\sigma.$$

The integrand above will be highly concentrated in a small region around $\hat{\omega}$ where

$$\hat{\omega} = \text{minimizer of } \|Y - X(\omega)\hat{\beta}(\omega)\|^2.$$

This is because the term

$$\exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \tag{3}$$

will decay rapidly as ω moves away from $\hat{\omega}$. Further, for ω close to $\hat{\omega}$, the determinant term $|X(\omega)^T X(\omega)|^{-1/2}$ will generally be quite close to $|X(\hat{\omega})^T X(\hat{\omega})|^{-1/2}$. We can therefore write

$$\begin{aligned}
\int \text{likelihood}(\theta) d\theta &= \int\int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) |X(\omega)^T X(\omega)|^{-1/2} d\omega d\sigma \\
&\approx |X(\hat{\omega})^T X(\hat{\omega})|^{-1/2} \int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p} \left[\int \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) d\omega\right] d\sigma.
\end{aligned}$$

To evaluate the integral over ω above, we use the fact that the quantity (3) is highly concentrated around $\hat{\omega}$ so that we can approximate it by a quadratic in the exponent. Specifically, let

$$S(\omega) := \|Y - X(\omega)\hat{\beta}(\omega)\|^2.$$

Note that $\hat{\omega}$ minimizes $S(\omega)$ so that $\nabla S(\hat{\omega}) = 0$. The second order Taylor approximation for $S(\omega)$ around $\hat{\omega}$ is thus

$$\begin{aligned} S(\omega) &\approx S(\hat{\omega}) + \langle \omega - \hat{\omega}, \nabla S(\hat{\omega}) \rangle + (\omega - \hat{\omega})^T \frac{HS(\hat{\omega})}{2} (\omega - \hat{\omega}) \\ &= S(\hat{\omega}) + (\omega - \hat{\omega})^T \frac{HS(\hat{\omega})}{2} (\omega - \hat{\omega}) \end{aligned}$$

where $HS(\hat{\omega})$ is the Hessian (matrix of second derivatives) of $S(\omega)$ evaulated at $\omega = \hat{\omega}$. Because $\hat{\omega}$ minimizes $S(\omega)$, the Hessian matrix $HS(\hat{\omega})$ is positive semi-definite. By this approximation, we get

$$\begin{aligned} \int \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) d\omega &= \int \exp\left(-\frac{S(\omega)}{2\sigma^2}\right) d\omega \\ &= \int \exp\left(-\frac{S(\hat{\omega}) + (\omega - \hat{\omega})^T \frac{HS(\hat{\omega})}{2} (\omega - \hat{\omega})}{2\sigma^2}\right) d\omega \\ &= \exp\left(-\frac{S(\hat{\omega})}{2\sigma^2}\right) \int \exp\left(-\frac{(\omega - \hat{\omega})^T \frac{HS(\hat{\omega})}{2} (\omega - \hat{\omega})}{2\sigma^2}\right) d\omega \\ &= \exp\left(-\frac{S(\hat{\omega})}{2\sigma^2}\right) \left(\sqrt{2\pi}\sigma\right)^k \left|\frac{1}{2} HS(\hat{\omega})\right|^{-1/2} \end{aligned}$$

where $|HS(\hat{\omega})/2|$ is the determinant of the Hessian matrix $HS(\hat{\omega})$.

We have therefore deduced that

$$\int \text{likelihood}(\theta) d\theta \approx |X(\hat{\omega})^T X(\hat{\omega})|^{-1/2} \left|\frac{1}{2} HS(\hat{\omega})\right|^{-1/2} \int_0^\infty \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p-k} \exp\left(-\frac{S(\hat{\omega})}{2\sigma^2}\right) d\sigma.$$

The integral over σ can be computed in closed form (as we previously did, say, in Lecture 10) to get:

$$\begin{aligned} \int_0^\infty \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-p-k} \exp\left(-\frac{S(\hat{\omega})}{2\sigma^2}\right) d\sigma &= \left(\frac{1}{\sqrt{2\pi}}\right)^{n-p-k} \frac{\frac{1}{2}\Gamma\left(\frac{n-p-k-1}{2}\right)}{\left(\frac{S(\hat{\omega})}{2}\right)^{(n-p-k-1)/2}} \\ &= \frac{1}{2\sqrt{2}} \frac{1}{\pi^{(n-p-k)/2}} \frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k-1}}. \end{aligned}$$

We have thus proved that

$$\int \text{likelihood}(\theta) d\theta \approx |X(\hat{\omega})^T X(\hat{\omega})|^{-1/2} \left|\frac{1}{2} HS(\hat{\omega})\right|^{-1/2} \frac{1}{2\sqrt{2}} \frac{1}{\pi^{(n-p-k)/2}} \frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k-1}}.$$

Combining with (2), we get the following formula for the Evidence:

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

Now we have to choose the prior. In the case of the linear regression model, we used the Zellner prior for β. We shall do the same here also. Specifically, conditional on ω, we shall assume that β and σ are independent with

$$\beta \mid \omega \sim N_p\left(0, \tau^2 (X(\omega)^T X(\omega))^{-1}\right) \quad \text{and} \quad (\log \sigma) \mid \omega \sim \text{Unif}(-C, C).$$

We also need to choose a prior for the parameter ω. For this, we shall use the choice:

$$\omega \sim N_k(0, \gamma^2 I_k).$$

In other words, we are assuming that the k components $\omega_1, \dots, \omega_k$ of ω are i.i.d $N(0, \gamma^2)$. With these prior choices, we can evaluate the prior density at the MLE $\hat{\theta}$. Note here that the MLE for ω is given by $\hat{\omega}$, the MLE for β is given by $\hat{\beta}(\hat{\omega})$ and the MLE for σ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2} = \frac{\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|}{\sqrt{n}}. \tag{4}$$

Plugging in the prior choices, we obtain

$$\begin{aligned}
&\text{prior}(\hat{\theta}) \\
&= \text{prior}(\hat{\beta}(\hat{\omega}), \hat{\sigma} \mid \hat{\omega})\text{prior}(\hat{\omega}) \\
&= \text{prior}(\hat{\beta}(\hat{\omega}) \mid \hat{\omega})\text{prior}(\hat{\sigma} \mid \hat{\omega})\text{prior}(\hat{\omega}) \\
&= \left(\frac{1}{\sqrt{2\pi}\tau}\right)^p |X(\hat{\omega})^T X(\hat{\omega})|^{1/2} \exp\left(-\frac{\hat{\beta}(\hat{\omega})^T X(\hat{\omega})^T X(\hat{\omega})\hat{\beta}(\hat{\omega})}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{2C\hat{\sigma}} \left(\frac{1}{\sqrt{2\pi}\gamma}\right)^k \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^{p+k} |X(\hat{\omega})^T X(\hat{\omega})|^{1/2} \tau^{-p} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{2C\hat{\sigma}} \gamma^{-k} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right)
\end{aligned}$$

Plugging this term in the derived formula for Evidence, we obtain the following formula for the evidence

$$\begin{aligned}
&\left(\frac{1}{\sqrt{2\pi}}\right)^{p+k} \tau^{-p} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{2C\hat{\sigma}} \gamma^{-k} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right) \\
&\times \left|\frac{1}{2}HS(\hat{\omega})\right|^{-1/2} \frac{1}{2\sqrt{2}} \frac{1}{\pi^{(n-p-k)/2}} \frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k-1}}.
\end{aligned}$$

Plugging in the formula (4) for $\hat{\sigma}$ gives

$$\begin{aligned}
&\sqrt{n}\left(\frac{1}{\sqrt{2\pi}}\right)^{p+k} \tau^{-p} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{2C} \gamma^{-k} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right) \\
&\times \left|\frac{1}{2}HS(\hat{\omega})\right|^{-1/2} \frac{1}{2\sqrt{2}} \frac{1}{\pi^{(n-p-k)/2}} \frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k}}.
\end{aligned}$$

This Evidence formula depends on the two quantities τ and γ (which appeared in the prior specification for $\beta \mid \omega$ and ω respectively). As in the last class (in the case of linear regression), we eliminate this dependence by integrating the above Evidence with respect to the following priors for τ and γ:

$$\log \tau \sim \text{Unif}(-C_1, C_1) \quad \text{and} \quad \log \gamma \sim \text{Unif}(-C_2, C_2).$$

Our final formula for the Evidence of the model (1) is thus:

$$\begin{aligned}
&\sqrt{n}\left(\frac{1}{\sqrt{2\pi}}\right)^{p+k} \frac{1}{4C_1C_2} \left[\int_{e^{-C_1}}^{e^{C_1}} \tau^{-p} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right) \frac{d\tau}{\tau}\right] \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{2C} \\
&\times \left[\int_{e^{-C_2}}^{e^{C_2}} \gamma^{-k} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right) \frac{d\gamma}{\gamma}\right] \left|\frac{1}{2}HS(\hat{\omega})\right|^{-1/2} \frac{1}{2\sqrt{2}} \frac{1}{\pi^{(n-p-k)/2}} \frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y - X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k}}.
\end{aligned}$$

Assuming that C_1 and C_2 are large, we can explicitly evaluate the integrals appearing above as

$$\begin{aligned}\int_{e^{-C_1}}^{e^{C_1}} \tau^{-p} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right)\frac{d\tau}{\tau} &\approx \int_0^\infty \tau^{-p} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right)\frac{d\tau}{\tau} \\ &= \int_0^\infty \tau^{-p-1} \exp\left(-\frac{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^2}{2\tau^2}\right) d\tau \\ &= 2^{(p-2)/2}\frac{\Gamma(p/2)}{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^p}\end{aligned}$$

and similarly

$$\begin{aligned}\int_{e^{-C_2}}^{e^{C_2}} \gamma^{-k} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right)\frac{d\gamma}{\gamma} &\approx \int_0^\infty \gamma^{-k} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right)\frac{d\gamma}{\gamma} \\ &= \int_0^\infty \gamma^{-k-1} \exp\left(-\frac{\|\hat{\omega}\|^2}{2\gamma^2}\right) d\gamma = 2^{(k-2)/2}\frac{\Gamma(k/2)}{\|\hat{\omega}\|^k}.\end{aligned}$$

Our final fully explicit formula for the Evidence is thus:

$$\begin{aligned}&\sqrt{n}\left(\frac{1}{\sqrt{2\pi}}\right)^{p+k}\frac{1}{4C_1C_2}2^{(p-2)/2}\frac{\Gamma(p/2)}{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^p}\frac{I\{e^{-C}<\hat{\sigma}<e^{C}\}}{2C} \\ &\times 2^{(k-2)/2}\frac{\Gamma(k/2)}{\|\hat{\omega}\|^k}\left|\frac{1}{2}HS(\hat{\omega})\right|^{-1/2}\frac{1}{2\sqrt{2}}\frac{1}{\pi^{(n-p-k)/2}}\frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y-X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k}} \\ &= \frac{\sqrt{n}}{64\sqrt{2}}\frac{I\{e^{-C}<\hat{\sigma}<e^{C}\}}{CC_1C_2}\frac{\Gamma(p/2)}{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^p}\frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y-X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k}}\frac{\Gamma(k/2)}{\|\hat{\omega}\|^k}\left|\frac{1}{2}HS(\hat{\omega})\right|^{-1/2}.\end{aligned}$$

The quantities C, C_1, C_2 will be the same across the different models being compared so they can be assumed to be proportionality constants for the Evidence and dropped. We can also drop the indicator term above (it will be one as C will be large) and the additional constants ($\sqrt{n}/64\sqrt{2}$) to get

$$\text{Evidence} \propto \left[\frac{\Gamma(p/2)}{\|X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^p}\right]\left[\frac{\Gamma\left(\frac{n-p-k-1}{2}\right)}{\|Y-X(\hat{\omega})\hat{\beta}(\hat{\omega})\|^{n-p-k}}\right]\left[\frac{\Gamma(k/2)}{\|\hat{\omega}\|^k}\left|\frac{1}{2}HS(\hat{\omega})\right|^{-1/2}\right]. \tag{5}$$

The right hand side above is a product of three terms. It is interesting to compare these terms with the following formula that we derived in the last class for linear regression:

$$\text{Evidence(linear model)} \propto \left[\frac{\Gamma\left(\frac{p}{2}\right)}{\|X\hat{\beta}\|^p}\right]\left[\frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y-X\hat{\beta}\|^{n-p}}\right]. \tag{6}$$

The linear regression evidence formula has two terms while the Evidence for the nonlinear model has three terms. The first term is the same in both the formulae. The second term is very similar except that in the case of nonlinear regression, $n-p$ is replaced by $n-p-k$ and this makes sense because the degrees of freedom is now further reduced by k because of the k parameters in ω. Finally the third term in the Evidence formula (5) has no analogue in the formula (6).

2 Recommended List of Readings for Today

The Evidence formula (5) can be found in Chapter 5 of the book *Bayesian spectrum analysis and parameter estimation* by Larry Bretthorst.