Recitation Five- Time Series

Nathaniel Mark

March 7, 2019

1 Likelihood intuition – simple examples

First, what is a likelihood? The likelihood is a function whose inputs are parameters (e.g μ,σ,θ,...) and its output is the "probability" that we would observe our data if the input parameters actually describe the true world. We call this "probability" the likelihood.[1] We write this as $L(\mu, \sigma, \theta, ...|X)$. Again, we are taking our data as given, and the likelihood changes as our assumed parameter values change.

For example, say that we have a coin with two sides. We do not know what is on the sides of the coin. It could be Tails on both sides ($\theta = TT$), heads on both sides ($\theta = HH$) or heads on one side and tails on the other($\theta = HT$). Now, say we flip the coin 3 times and acquire the data {H,H,H}. What is the probability that we observed this data if the coin is truly $\theta = TT$? 0! What is the probability that we observed this data if the coin is truly $\theta = HH$? 1. What is the probability that we observed this data if the coin is truly $\theta = HT$? $\frac{1}{8}$. Therefore,

$$L(\theta|X = \{H, H, H\}) = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{8} \end{bmatrix} if \begin{bmatrix} \theta = HH \\ \theta = TT \\ \theta = HT \end{bmatrix} \tag{1}$$

What is the maximum likelihood estimate here? Just the theta that maximizes the likelihood. That is, $\hat{\theta}_{MLE} = HH$.

The intuition remains the same when we move to problems where our model has continuous distributions. We still choose the parameter values that maximize the probability distribution function, taking our data values as given. However, now, the value of the probability distribution

[1] It is not quite probability if your data have a continuous probability distribution. That is why I put probability in quotes. But, thinking about it as probability helps to understand, an the intuition is correct.

function is no longer true probability, but for the purposes of understanding, it can be thought of as "probability".[2].

For example, say that we have data on heights of children in a population. We know that in the true world, height of children is distributed normally with variance =1. But, we do not know the mean of the distribution. Say we see a kid who is 4.4 ft tall. What would be our MLE of the mean of the height distribution? That is, given that our data is $X_1 = 4.4$ Of course, it is $\mu = 4.4$, as the peak of the normal pdf is at its mean, so the mean is the most likely place for our data to come from.

Turning to more than one observation complicates things, but not be much. Now, our joint probability density function is:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}) \tag{2}$$

In other words, the joint probability density function of three heights is

$$L(\mu|\mathbf{X}) = f(X_1, X_2, X_3|\mu) = f(X_1|\mu)f(X_2|\mu)f(X_3|\mu)$$

Given our data X, this is simply a function of parameter μ, so we can find the value of μ that maximizes this function.

2 Maximum Likelihood - Definitions

Definition: The Likelihood function $L(\theta|data)$ is equal to the joint probability of observing the data given a parameter θ:

$$L(\theta|data) = f(y_1, y_2, ..., y_T|\theta)$$

where y_i is an observation of data.

Corollary 1: If the data are i.i.d, then the likelihood can be simplified:

$$L(\theta|data) = f(y_1, y_2, ..., y_T|\theta) = \prod_{y=1}^{T} f(y_t|\theta)$$

Corollary 2: If the data are not i.i.d. (e.g. serially correlated) then the likelihood can be simplified:

$$L(\theta|data) = f(y_1, y_2, ..., y_T|\theta) = f(y_1|\theta)f(y_2|y_1, \theta)f(y_3|y_1, y_2, \theta)...$$

[2]It truly is the $lim_{\Delta \to 0} P$(our data falls within ball with radius$\Delta/2$around X , given θ describes the true world)$/\Delta$

by the conditional densities rule.

$$= f(y_1|\theta)\prod_{t=1}^{T} f(y_t|y_{t-1},...y_1,\theta)$$

Definition: The Maximum Likelihood Estimator is the parameter that maximizes the likelihood function:

$$\hat{\theta}^{MLE} = argmax_\theta L(\theta|data)$$

Corollary 3: The maximum likelihood estimator can be attained by finding the maximum of the log-likelihood function. This is often done instead of maximizing the likelihood because it is easier to deal with:

$$\hat{\theta}^{MLE} = argmax_\theta log(L(\theta|data)) = argmax_\theta L(\theta|data)$$

Theorem: The Maximum Likelihood Estimator is:

1. Consistent

$$\hat{\theta}^{MLE} \to \theta^{TRUE}$$

2. Asymptotically Normal

$$\hat{\theta}^{MLE} \to^d N(\theta^{TRUE}, -E[\frac{\partial^2 log(L(\theta|data)}{\partial\theta^{TRUE}\partial\theta^{TRUE\prime}}])$$

3. Efficient
4. *Not* necessarily unbiased

3 Maximum Likelihood - Time Series

Maximum likelihood is very useful for estimating time series models, because it explicitly takes into account conditional probability distributions. I discuss the basic idea of how this is useful.

Assume that $\{X_1,...,X_T\}$ is an AR or MA process, where $\epsilon_t \sim N(0,\sigma^2)$.

Then, $\{X_1,...,X_T\}$ is distributed as follows:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix} = \mathcal{N}(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(1) & \dots & \gamma(T-1) & \\ \vdots & & & \\ \gamma(T-1) & & & \end{bmatrix}) \tag{3}$$

Therefore, we can estimate AR or MA models in the following way:

Step One:

Parametrize the model. That is, find the $\gamma(0)$, $\gamma(1)$, $\gamma(2)$ in terms of the parameters of the model.

Step Two:

Define the likelihood function as:

$$L(\theta|X_1, ...X_2) = f(X_1|\theta)\prod_{t=1}^{T} f(X_t|X_{t-1}, ...X_1, \theta)$$

Step Three: Take the log of the likelihood:

$$\mathcal{L}(\theta|X_1, ...X_2) = log(f(X_1|\theta)) + \sum_{t=1}^{T} log(f(X_t|X_{t-1}, ...X_1, \theta))$$

Step Four: Find the values of θ that maximize the log-likelihood function. This value is the maximum likelihood estimator.

4 Maximum Likelihood - Special Case of MA(1) Example

Assume that we have an MA(1) model where the variance is know to be equal to 1:

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1}$$

Where $\epsilon_t \sim N(0, 1)$ The model: That is, assume that our time series $\{X_t\}_{t=1,...,T}$ are distributed multivariate normal such that

$$Cov(X_t, X_t) = 1 + \theta_1$$

$$Cov(X_t, X_{t-1}) = \theta_1$$

$$Cov(X_t, X_{t-h}) = 0$$

That is,

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix} = \mathcal{N}(\begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} 1+\theta_1^2 & \theta_1 & \dots & 0 \\ \theta_1 & \dots & 0 & \\ \vdots & & & \\ 0 & & & \end{bmatrix}) \tag{4}$$

The question is: given data X, what is the μ and θ_1 that maximize the log-likelihood?

One way to do this is by brute force: try many different possible values of μ and θ and see which

one maximizes the log-likelihood function.

We are going to ask you to do the first step of this for an AR model in your homework: That is, given one value of μ and θ, what is the log-likelihood?

There are generally two ways of computing this in Python:

1. Directly asking python the probability density associated with your X and parameter values using the joint distribution above (that is, directly finding the value of $L(\mu, \theta|\mathbf{X}) = f(X_1, X_2, ...X_T)$), then taking the log to get $\mathcal{L}(\mu, \theta|\mathbf{X}) = ln(f(X_1, X_2, ...X_T))$
2. Noting that we can rewrite the joint probability density functions of time series in terms of their conditional probability density functions. That is, we can write:

$$L(\mu, \theta|\mathbf{X}) = f(X_1, X_2, ...X_T) = f(X_1)f(X_2|X_1)f(X_3|X_1, X_2)...f(X_T|X_1, X_2, ..., X_{T-1})$$

and, using log rules:

$$\mathcal{L}(\mathbf{X}|\mu, \theta) =$$

$$ln(f(X_1, X_2, ...X_T) = ln(f(X_1))+ln(f(X_2|X_1))+ln(f(X_3|X_1, X_2))...+ln(f(X_T|X_1, X_2, ..., X_{T-1}))$$

then asking python for the values of $f(X_1)$, $f(X_2|X_1)$, etc., and plugging them into the equation

In our example above, we can further simplify (because correlation of X_t and X_{t-h} is zero for h more than 1) to

$$f(X_1, X_2, ...X_T) = f(X_1)f(X_2|X_1)f(X_3|X_2)...f(X_T|X_{T-1})$$

Finally, we can use the rules about joint densities of normals to figure out these conditional distributions:

$$(X_1|X_2 = x_2) \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2)$$

, from wikipedia, and plugging in our values:

$$(X_t|X_{t-1} = x_{t-1}) \sim N(\mu + \frac{\theta_1}{1 + \theta_1^2}(x_{t-1} - \mu), 1 + \frac{\theta_1^4}{1 + \theta^2})$$

Theoretically,