Introduction to Maximum Likelihood Estimation (MLE)

Example with Normal IID variables

Juan Garbayo

Forecasting & Time Series Analysis

ie UNIVERSITY

Contents:

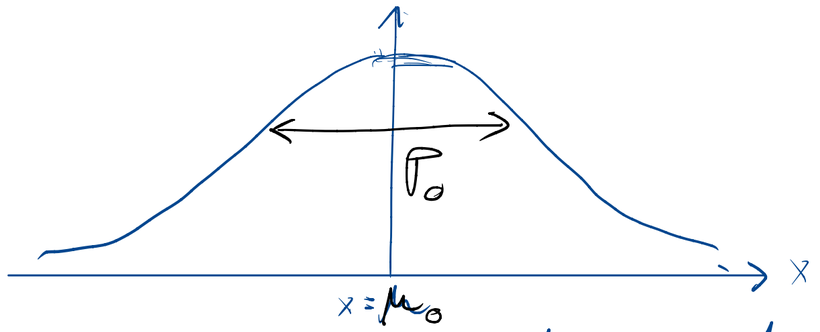
1. Example with a normal distribution
2. Generalization from the example
3. Relationship L and AIC

MLE → Example Normal Distribution

- f_X → Probability density function of a normal random variable: $X \sim N(\mu_0, \sigma_0)$

  the sample:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot e^{\frac{-1}{2}\left[\frac{x-\mu_0}{\sigma_0}\right]^2}$$



- X_n → first n elements of an IID sequence of normal random variables with μ_0 and σ_0^2 → $\{X_1, X_2, \ldots, X_n\}$ with X

- Joint Probability Density Function of $\{X_1 \ldots X_n\}$

- $f_{X_n}(x_1, \ldots, x_n) \overset{(=)}{\downarrow} \prod_{i=0}^{n} f_X(x_i) = (2\pi\sigma_0^2)^{-n/2} \cdot e^{\frac{-1}{2\sigma_0^2}\left[\sum_{i=0}^{n}(x_i - \mu_0)^2\right]}$

IID = product of ind. events!!

PROBLEM: μ_0 and σ_0 unknown (parameters of our model)

MAXIMUM LIKELIHOOD ESTIMATION: given a sample $\{x_1, \dots x_n\}$ (the sample is now a given, not a variable) obtain the values of σ and μ that maximize the the value of the joint probability density function particularized at the given sample:

Realization of the RVs $\{x_1, \dots, x_n\}$

$$\mathcal{L}(\mu, \sigma) = (2\pi\sigma^2)^{-n/2} \cdot e^{\frac{-1}{2\sigma}\left[\sum_{i=0}^{n}(x_i - \mu)\right]}$$

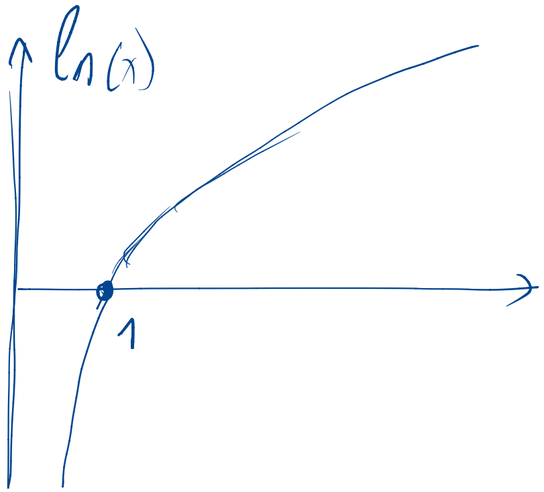
x_i: fixed, the given sample

In other words: assuming the sample was generated by a $N(\mu, \sigma^2)$, find the values of μ and σ that maximize the probability of occurrence of the sample.

FORMALLY:

$$\{\mu_0, \sigma_0\} = \underset{\{\mu,\sigma\} \in \mathbb{R} \times \mathbb{R}^+}{\operatorname{argmax}} \; \mathcal{L}(\mu, \sigma)$$

Because the function $\ln(x)$ is monotonically increasing:



$$\underset{\{\mu,\sigma\} \in \mathbb{R} \times \mathbb{R}^+}{\operatorname{argmax}} \; \mathcal{L}(\mu, \sigma) = \underset{\{\mu,\sigma\} \in \mathbb{R} \times \mathbb{R}^+}{\operatorname{argmax}} \; \ln\left(\mathcal{L}(\mu, \sigma)\right)$$

$\ln\left(\mathcal{L}(\mu, \sigma)\right) \rightarrow$ log-likelihood.

MLE → Normal Distr. (Cont'd)

$$L(\mu, \sigma, x_1, \dots, x_n) \overset{\text{IID}}{=} \prod_{i=0}^{n} f_X(x_i) = (2\pi\sigma^2)^{-n/2} \cdot e^{-\frac{1}{2\sigma^2}\left[\sum_{i=0}^{n}(x_i-\mu)^2\right]}$$

LIKELIHOOD FUNCTION

Reminder:

- $Ln(a \cdot b) = Ln(a) + Ln(b)$
- $Ln(a^b) = b \cdot Ln(a)$
- $e^{Ln(a)} = a$

Take logarithms:

$$\ell = Ln(L)$$

Ln Ln

constant !!

LOG-LIKELIHOOD FUNCTION

$$\ell(\mu, \sigma^2, x_1 \dots x_n) = \frac{-n}{2} Ln(2\pi) - \frac{n}{2} Ln(\sigma^2) - \frac{1}{2\sigma^2} \tilde{\sum_{i=0}} (x_i - \mu)^2$$

MLE – Normal Distribution → μ_0, σ_0

- $\frac{\partial}{\partial \mu} \ell(\mu, \sigma, x_1 \dots x_n) = 0 \Rightarrow \frac{-1}{\sigma^2} \sum_{i=0}^{n} (x_i - \hat{\mu}) \cdot (-1) = 0 \Rightarrow \sum_{i=0}^{n} (x_i - \hat{\mu}) = 0$

$$\rightarrow \sum_{i=0}^{n} x_i - n\hat{\mu} = 0 \rightarrow \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=0}^{n} x_i}$$

Mean sample.

- $\frac{\partial}{\partial \sigma} \ell(\mu, \sigma, x_1 \dots x_n) = 0 \rightarrow -\frac{n}{2} \frac{1}{\hat{\sigma}^2} \cdot 2\hat{\sigma} - \frac{1}{2} \cdot (-2) \hat{\sigma}^{-3} \sum_{i=0}^{n} (x_i - \hat{\mu})^2 = 0$

$$\rightarrow \frac{1}{\hat{\sigma}^2} \left[-n + \frac{1}{\hat{\sigma}^2} \sum_{i=0}^{n} (x_i - \hat{\mu})^2 \right] = 0$$

$\neq 0 \ (\hat{\sigma} \in \mathbb{R}^+)$

$$= 0 \rightarrow n\hat{\sigma}^2 = \sum_{i=0}^{n} (x_i - \mu)^2 \rightarrow \boxed{\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=0}^{n} (x_i - \hat{\mu})^2}$$

Unadjusted sample variance.

Conclusions MLE

- General method to estimate the parameters of a statistical model to maximize its fit to a given sample.
- In the example we have:
  - model → $N(\mu, \sigma)$
  - sample → $\{x_1, \dots, x_n\}$
  - parameters → μ, σ
- The $\mathcal{L}$ can be computed for other statistical models and maximized to estimate the parameters that result in a best fit to a given sample.
- Example:
  - model → ETS, ARIMA... → obtain the likelihood of y_T under the given model
  - sample → y_T
  - parameters of ETS or ARIMA → $\{\theta_1, \dots \theta_p\}$

$\mathcal{L}(\theta_1, \dots, \theta_p, y_T)$ → Values of $\{\theta_1 \dots \theta_p\}$ that maximize $\mathcal{L}$

ESTIMATES $\{\hat{\theta}_1, \dots \hat{\theta}_p\}$

Relationship with AIC → Rule: the smaller the AIC, the better (defined to work in this manner).

- The likelihood (and log likelihood) L is a metric of how well the model fits the sample. → $L\uparrow$ → fit improves.

- $AIC = f_1(L) + f_2(p)$ — L: Likelihood; p: Number of parameters (complexity)
  - $f_1(L)$ such that if $L\uparrow \rightarrow f_1(L)\downarrow \rightarrow AIC\downarrow$
  - $f_2(p)$ such that if $p\uparrow \rightarrow f_2(p)\uparrow \rightarrow AIC\uparrow$

- Case (1): L increases without increasing too much p → $L\uparrow\uparrow\uparrow$ and $p\uparrow$ → $AIC\downarrow$ ✓
- Case (2): L increases at the cost of increasing p too much → $L\uparrow$ and $p\uparrow\uparrow\uparrow$ → $AIC\uparrow$ ✗
  - Prevents overfitting

Example of AIC

$$AIC_{\text{ETS Models}} = \underbrace{-2\log(L)}_{\sim f_1(L)} + \underbrace{2p}_{\sim f_2(p)}$$

- $-2\log(L)$: Likelihood of a TS assuming it was generated by an ETS model
  - more complex that the example of the normal dist
- $2p$: Number of parameters in the model

Check behavior

- $\frac{df_1}{dL} = -\frac{2}{L} < 0 \Rightarrow L \uparrow \Rightarrow AIC \downarrow$ ($L > 0$)
- $\frac{df_2}{dp} = 2 > 0 \rightarrow p \uparrow \Rightarrow AIC \uparrow$