Introduction to Maximum Likelihood Estimation (MLE)

Example with Normal IID variables

Juan Garbayo

Forecasting & Time Series Analysis

ie UNIVERSITY

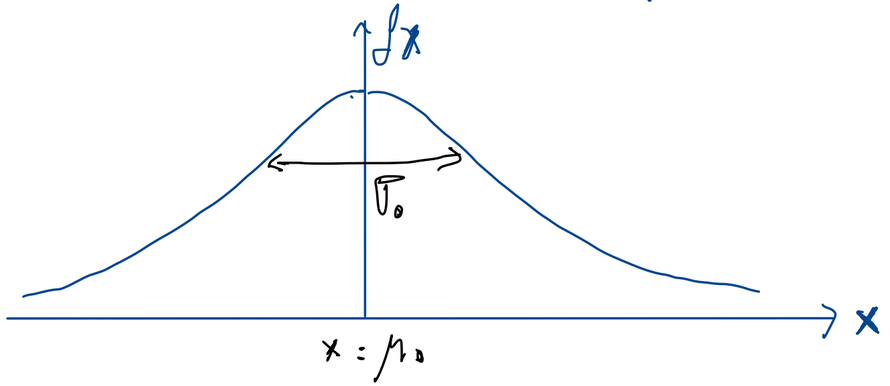
Contents:

1. Example with a normal distribution $N(\mu, \sigma)$
2. Generalization from the example to TS Models.
3. Relationship L and AIC

MLE: Example with Normal Distribution

f_X: probability density function of a normal random variable $X \sim N(\mu, \sigma)$

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



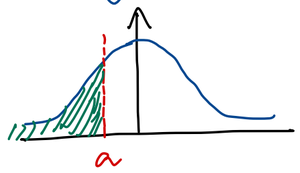
$X \rightarrow$ random variable

$x \rightarrow$ variable of the pdf. Its domain is all the possible outcomes of X.

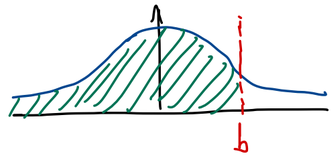
Remember: $f_X(x)$ is a probability DENSITY

Because X is a continuous RANDOM VARIABLE, we do not compute the probability of X taking a specific value, but rather the probability of X being within a specific interval.

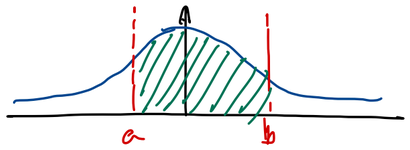
Example ①: $P(X \leq a) = \int_{-\infty}^{a} f_X(x)\,dx$



Example ②: $P(X \leq b) = \int_{-\infty}^{b} f_X(x)\,dx$



Example ③: $P(a \leq X \leq b) = \int_{a}^{b} f_X(x)\,dx$



Of course, because of the axioms of probability:

$$\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$$

↳ NOTE: $f_X(x)$ may be > 1 in a subset.

Joint Probability Density Function of sequence of IID $N(\mu_0, \sigma_0)$ (1/2)

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

Single R.V called X

Known parameters

IID → Independent and Identically Distributed

Experiment: first n elements of an IID sequence of normal random variables

$\overline{X} = \{X_1, X_2, \ldots, X_n\}$ where $X_i \sim N(\mu_0, \sigma_0)$; $\overline{X}$ is a RANDOM VECTOR

The random vector $\overline{X}$ is the process (experiment) of extracting the first n elements of a sequence of independent random variables

The joint probability density function of $\overline{X}$ is the mathematical characterization of that process.

↳ Next page

Joint Probability Density Function of sequence of IID $N(\mu_0, \sigma_0)$ (2/2)

Probability density function of normal R.V $X_i \sim N(\mu, \sigma)$

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

Joint Probability Density Function of $\bar{X} = \{X_1, X_2, \dots, X_n\}$

$$f_{\bar{X}}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = (2\pi\sigma_0^2)^{-n/2} \cdot e^{\frac{-1}{2\sigma_0^2}\left[\sum_{i=1}^{n}(x_i - \mu_0)^2\right]}$$

IID → product of independent events !!

Reminder $e^{a} \cdot e^{b} = e^{a+b}$

Example of Joint Probability Density Function with n = 2

- $f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot e^{\frac{-1}{2\sigma_0^2}(x_1-\mu_0)^2}$
- $f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot e^{\frac{-1}{2\sigma_0^2}(x_2-\mu_0)^2}$

$\bar{X} = \{X_1, X_2\}$

$$f_{\bar{X}}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = \frac{1}{2\pi\sigma_0^2} \cdot e^{\frac{-1}{2\sigma_0^2}\left[(x_1-\mu_0)^2 + (x_2-\mu_0)^2\right]}$$

Example of computation of Joint Probabilities

$$P\left(X_1 \in [a_1, b_1],\ X_2 \in [a_2, b_2]\right) = \int_{x_1=a_1}^{x_1=b_1} \int_{x_2=a_2}^{x_2=b_2} f_{\bar{X}}(x_1, x_2)\, dx_1\, dx_2$$

Estimate μ_0 and σ_0 to best fit a sample

- **Assumption:** data follows a normal distribution $N(\mu, \sigma)$ → μ, σ UNKNOWN
- **Problem:** find μ_0 and σ_0 corresponding to the normal distribution that best fits the sample

Data given: $\{x_1, \dots, x_n\}$ → specific outcome of the random vector $\bar{X} = \{X_1, X_2, \dots, X_n\}$

$\{x_1, \dots, x_n\}$: SAMPLE

- **Approach:** find the values of μ and σ that maximize the probability of the random vector $\bar{X}: \{X_1, X_2, \dots, X_n\}$ taking the values of the sample $\{x_1, x_2, \dots, x_n\}$.

$$\mathcal{L}(\mu, \sigma, x_1, x_2, \dots, x_n) = (2\pi\sigma)^{-n/2} \cdot e^{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \mu)^2\right]}$$

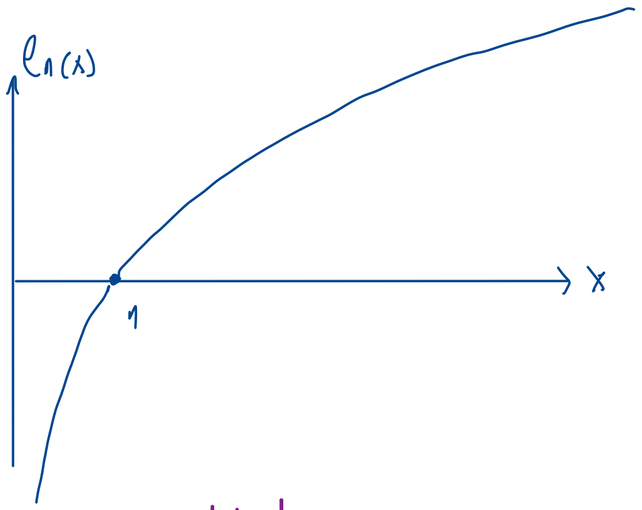
- $\mathcal{L}$: LIKELIHOOD
- μ, σ: model parameters to be estimated
- $x_1, x_2, \dots, x_n$: sample given
- σ: UNKNOWN
- x_i: FROM GIVEN SAMPLE (KNOWN)
- μ: UNKNOWN
- Joint probability density function of the process $\bar{X}$ particularized at the specific outcome (the sample $\{x_1, x_2, \dots, x_n\}$)

In other words: assuming your data was generated by a $N(\mu, \sigma)$, find μ_0 and σ_0 that maximize the probability of occurrence of the observed data.

FORMAL WAY TO WRITE THE OPTIMIZATION PROBLEM:

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

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



Same optimization problem

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

$\ln$: Natural logarithm (base e).

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

PROBLEM (1) | PROBLEM (2)

Log - Likelihood for normal distribution case

Likelihood

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

Reminder:
- $\ln(a \cdot b) = \ln(a) + \ln(b)$
- $\ln(a^b) = b \ln(a)$
- $e^{\ln(a)} = a$

$\ln$ $\qquad$ $\ln$

Log - Likelihood

$$\ell(\mu, \sigma, x_1, \dots, x_n) = \underbrace{\frac{-n}{2}\ln(2\pi)}_{\text{constant}} \underbrace{- \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2}_{f(\mu,\sigma)}$$

MLE – Normal Distribution → μ_0, σ_0

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

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

Mean sample.

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

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

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

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

↙ 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:

$$\begin{cases} \underline{\text{model}} \rightarrow N(\mu, \sigma) \\ \underline{\text{sample}} \rightarrow \{x_1, \dots, x_n\} \; ; \; \text{Process} \rightarrow \{X_1, \dots, X_n\} \\ \underline{\text{parameters}} \rightarrow \mu, \sigma \end{cases}$$

- The L can be computed for other statistical models and maximized to estimate the parameters that result in a best fit of the model to the given sample.

Maximum Likelihood Estimation - More general notation

Problem: given a sample of n elements $\{x_1, \dots, x_n\} = \bar{x}$ and a statistical model dependant on p parameters $\{\theta_1, \dots \theta_p\} = \bar{\theta}$ find the values of the parameters that maximize the joint probability of occurrence of the sample under the assumption that the model generated the sample.

$$L(\bar{\theta}, \bar{x}) = f_{(H)}(\bar{\theta}, \bar{X}) \Big|_{\bar{X} = \bar{x}}$$

- $L(\bar{\theta}, \bar{x})$: Likelihood
  - $\bar{\theta}$: Model params
  - $\bar{x}$: Sample $\{x_1, x_2, \dots, x_n\}$
- $f_{(H)}(\bar{\theta}, \bar{X})$: joint probability density function
  - H: set of possible values for the parameters $\bar{\theta} = \{\theta_1, \dots, \theta_p\}$
  - $\bar{X}$: Random vector $\{X_1, X_2, \dots, X_n\}$

$\bar{\theta} = \{\theta_1, \theta_2, \dots, \theta_p\}$

$$\hat{\bar{\theta}} = \underset{\bar{\theta} \in H}{\operatorname{argmax}} \; L(\bar{\theta}, \bar{x})$$

↝ Optimization problem to be solved.

- H: set of possible values taken by the parameters $\{\theta_1, \dots, \theta_p\}$
- $\hat{\bar{\theta}}$: Values of the params that maximize the likelihood

MLE on Time Series Models

- model: ETS, ARIMA ...
- process/model $\bar{Y}_T = \{Y_1, Y_2, \dots, Y_T\}$ → the collection of R.V that describe the stochastic process
- sample: $\bar{y}_T = \{y_1, y_2, \dots, y_T\}$ → the TS data given → the specific realization of the TS.
- parameters $\bar{\theta} = \{\theta_1, \theta_2, \dots \theta_p\}$
- Likelihood function $\mathcal{L}(\bar{\theta}, \bar{y}_T) = f_{\Theta}(\bar{\theta}, \bar{Y}_T)\big|_{\bar{Y}_T = \bar{y}_T}$

→ Joint probability density function of the TS (of the collection of R. Variables) assuming the TS process obeys the equations of the model.

→ particularized at the sample.

- Optimization problem:

$$\hat{\bar{\theta}} = \underset{\bar{\theta} \in H}{\operatorname{argmax}}\ \mathcal{L}(\bar{\theta}, \bar{y}_T)$$

→ Set of possible values for the vector $\bar{\theta}$ of parameters

$\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p\}$ values of parameters that maximize $\mathcal{L}$

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 than the example of the normal distr.
- $2p$: Number of parameters in the model

Check behavior

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