FSAN/ELEG815: Statistical Learning

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

6. Maximum Likelihood and Bayes Estimation

Maximum Likelihood and Bayes Estimation

Estimation

Estimation is the inference of unknown quantities. Two cases are considered:

1. Quantity is fixed, but unknown – parameter estimation
2. Quantity is random and unknown – random variable estimator

Parameter Estimation

Consider a set of observations forming a vector

$$\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$$

Assumption: The x_i *RV*s come from a known density governed by unknown (but fixed) parameter θ

Objective: Estimate θ. What optimality criteria should be used?

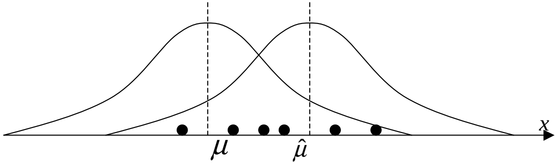
Definition (Maximum Likelihood Estimation)

The maximum likelihood estimate of θ is the value $\hat{\theta}_{\text{ML}}(\mathbf{x})$ which makes the $\mathbf{x}$ observations most likely

$$\hat{\theta}_{\text{ML}}(\mathbf{x}) = \underset{\theta}{\operatorname{argmax}} \, f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$$

Example

Let $x_i \sim N(\mu, \sigma^2)$. Given N observations, find the ML estimate of μ.



For i.i.d. samples

$$\begin{aligned} f_{\mathbf{x}|\mu}(\mathbf{x}|\mu) &= \prod_{i=1}^{N} f_{x_i|\mu}(x_i|\mu) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad \text{[Gaussian case]} \\ &\triangleq \text{likelihood function} \end{aligned}$$

Thus the estimate of the mean it is set as

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}} \, f_{\mathbf{x}|\mu}(\mathbf{x}|\mu)$$

Interpretation: Set the distribution mean to the value that makes obtaining the observed samples most likely.

Note: Maximizing $f_{\mathbf{x}|\mu}(\mathbf{x}|\mu)$ is equivalent to maximizing any monotonic function of $f_{\mathbf{x}|\mu}(\mathbf{x}|\mu)$. Choosing $\ln(\cdot)$

$$\begin{aligned}
\ln(f_{\mathbf{x}|\mu}(\mathbf{x}|\mu)) &= \ln\left(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)\\
&= -N\ln(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\\
&= -N\ln(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N}\frac{x_i^2}{2\sigma^2} + \mu\sum_{i=1}^{N}\frac{x_i}{\sigma^2} - \sum_{i=1}^{N}\frac{\mu^2}{2\sigma^2}
\end{aligned}$$

Taking the derivative and equating to 0,

$$\frac{\partial\ln(f_{\mathbf{x}|\mu}(\mathbf{x}|\mu))}{\partial\mu} = \sum_{i=1}^{N}\frac{x_i}{\sigma^2} - \frac{N\mu}{\sigma^2} = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}x_i \triangleq \text{sample mean}$$

General Maximum Likelihood Result

General Statement: The ML estimate of θ is

$$\hat{\theta}_{\text{ML}}(\mathbf{x}) = \underset{\theta}{\operatorname{argmax}}\, f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$$

Solution: The ML estimate of θ is obtained as the solution to

$$\left.\frac{\partial}{\partial\theta} f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)\right|_{\theta=\theta_{\text{ML}}} = 0$$

or

$$\left.\frac{\partial}{\partial\theta} \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]\right|_{\theta=\theta_{\text{ML}}} = 0$$

- $f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$ is the likelihood function of θ.
- $\hat{\theta}_{\text{ML}}$ is a *RV* since it is a function of the *RVs* $x_1, x_2, \cdots, x_N$

Historical Note: ML estimation was pioneered by geneticist and statistician Sir R. A. Fisher between 1912 and 1922

Example

The time between customer arrivals at a bar is a *RV* with distribution

$$f_T(T) = \alpha e^{-\alpha T} U(T)$$

Objective: Estimate the arrival rate α based on N measured arrival intervals $T_1, T_2, \cdots, T_N$.

Assuming that the arrivals are independent,

$$\begin{aligned} f(T_1, T_2, \cdots, T_N) &= \prod_{i=1}^{N} f_T(T_i) \\ &= \prod_{i=1}^{N} \alpha e^{-\alpha T_i} = \alpha^N e^{-\alpha \sum_{i=1}^{N} T_i} \\ \Rightarrow \ln[f(T_1, T_2, \cdots, T_N)] &= [N \ln(\alpha) - \alpha \sum_{i=1}^{N} T_i] \end{aligned}$$

Taking the derivative and equating to 0,

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln[f(T_1,T_2,\cdots,T_N)] &= \frac{\partial}{\partial\alpha}[N\ln(\alpha)-\alpha\sum_{i=1}^{N}T_i] \\ &= \frac{N}{\alpha}-\sum_{i=1}^{N}T_i = 0\end{aligned}$$

Solving for α gives the ML estimate

$$\Rightarrow \hat{\alpha}_{\mathrm{ML}} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}T_i} = \frac{1}{\overline{\overline{T}}}$$

Result: The ML estimate of arival rate for exponentially distributed samples is the reciprocal of the sample mean arrival

Properties of Estimates

Since $\hat{\theta}_N$ is a function of *RVs* $x_1, x_2, \cdots, x_N$, estimates are*RVs* and we can state the following properties:

- An estimate $\hat{\theta}_N$ is unbiased if

$$E\{\hat{\theta}_N\} = \theta \qquad \text{bias} \triangleq E\{\hat{\theta}_N\} - \theta$$

- $\hat{\theta}_N$ is consistent (converges in probability) if

$$\lim_{N\to\infty} \Pr\{|\hat{\theta}_N - \theta| < \epsilon\} = 1 \quad \text{for arbitrary } \epsilon$$

- $\hat{\theta}_N$ is efficient in comparison to other estimators if

$$\text{var}(\hat{\theta}_N) < \text{var}(\hat{\theta}_{\text{other}})$$

Note: If $\hat{\theta}_N$ is unbiased and efficient with respect to $\hat{\theta}_{N-1}$ for all N (i.e., $\text{var}(\hat{\theta}_N)$ converges to 0), then $\hat{\theta}_N$ is a consistent estimate

To prove the consistent estimate result, note that by the Tchebycheff inequality

$$\Pr\{|\hat{\theta}_N - \theta| > \epsilon\} \leq \frac{\text{var}(\hat{\theta}_N)}{\epsilon^2}$$

If $\text{var}(\hat{\theta}_N) < \text{var}(\hat{\theta}_{N-1})$, the above gives

$$\lim_{N\to\infty} \Pr\{|\hat{\theta}_N - \theta| > \epsilon\} = 0$$

or

$$\lim_{N\to\infty} \Pr\{|\hat{\theta}_N - \theta| < \epsilon\} = 1$$

That is, it converges in probability, or is consistent

QED

Example

Let $\{x_i\}$ be WSS with uncorrelated samples. Is the sample mean a consistent estimator for this sequence?

Step 1: Consider the bias

$$\begin{aligned} E\{\hat{\mu}_N\} &= E\left\{\frac{1}{N}\sum_{i=1}^{N} x_i\right\} \\ &= \frac{1}{N}(N\mu) = \mu \end{aligned}$$

Result: $\hat{\mu}_N$ is unbiased

Step 2: Consider the variance

$$\text{var}(\hat{\mu}_N) = E\left\{(\hat{\mu} - \mu)^2\right\}$$

$$
\begin{aligned}
\text{var}(\hat{\mu}_N) &= E\left\{(\hat{\mu}-\mu)^2\right\} \\
&= E\left\{\left(\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)-\mu\right)^2\right\} \\
&= \frac{1}{N^2}E\left\{\left(\sum_{i=1}^{N}(x_i-\mu)\right)^2\right\} \qquad \text{[assume uncorrelated]} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}E\{(x_i-\mu)^2\} + \frac{1}{N^2}\underbrace{E(\text{cross terms})}_{=0} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}E\{(x_i-\mu)^2\} = \frac{1}{N^2}(N\sigma^2) = \frac{\sigma^2}{N}
\end{aligned}
$$

Result: $\hat{\mu}_N$ is unbiased and $\text{var}(\hat{\mu}_N) < \text{var}(\hat{\theta}_{N-1}) \Rightarrow \hat{\mu}_N$ is consistent

Theorem (Cramer-Rao Bound (1945, 1946))

If $\hat{\theta}$ is an unbiased estimate of θ, then

$$var(\hat{\theta}) \geq \left(E\left\{ \left(\frac{\partial}{\partial \theta} \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)] \right)^2 \right\} \right)^{-1}$$

or equivalently

$$var(\hat{\theta}) \geq \left(-E\left\{ \frac{\partial^2}{\partial \theta^2} \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)] \right\} \right)^{-1}$$

where it is assumed

$$\frac{\partial}{\partial \theta} f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) \quad \textit{and} \quad \frac{\partial^2}{\partial \theta^2} f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) \quad \textit{exist}$$

Note: If any estimate satisfies the bound with equality, it is an efficient (minimum variance) estimate

Proof:

Since $\hat{\theta}$ is unbiased

$$E\{\hat{\theta}-\theta\} = \int_{-\infty}^{\infty} (\hat{\theta}-\theta) f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) d\mathbf{x} = 0$$

Taking the derivative

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}(\hat{\theta}-\theta) f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) d\mathbf{x} &= 0 \\
\Rightarrow \underbrace{-\int_{-\infty}^{\infty} f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) d\mathbf{x}}_{-1} + \int_{-\infty}^{\infty} \frac{\partial f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)}{\partial\theta}(\hat{\theta}-\theta) d\mathbf{x} &= 0 \\
\Rightarrow \int_{-\infty}^{\infty} \frac{\partial f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)}{\partial\theta}(\hat{\theta}-\theta) d\mathbf{x} &= 1 \qquad (*)
\end{aligned}$$

Note the following equality

$$\frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial \theta} f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) = \frac{\partial f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)}{\partial \theta}$$

Using this in $(*)$

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{\partial f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)}{\partial \theta}(\hat{\theta}-\theta) d\mathbf{x} &= 1 \\ \Rightarrow \int_{-\infty}^{\infty} \frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial \theta} f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)(\hat{\theta}-\theta) d\mathbf{x} &= 1 \end{aligned}$$

This can be equivalently expressed as

$$\left(\int_{-\infty}^{\infty} \left(\frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial \theta} \sqrt{f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)} \right) \left(\sqrt{f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)}(\hat{\theta}-\theta) \right) d\mathbf{x} \right)^2 = 1$$

Definition (Cauchy–Schwarz Inequality (1821 disc.; 1859 cont.))

Cauchy-Schwarz's inequality states (for square–integrable complex–valued functions),

$$\left| \int f(x)g(x)\, dx \right|^2 \leq \int |f(x)|^2\, dx \cdot \int |g(x)|^2\, dx$$

with equality only if $f(x) = k \cdot g(x)$, where k is a constant

Thus

$$\left(\int_{-\infty}^{\infty} \left(\frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial \theta} \sqrt{f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)} \right) \left(\sqrt{f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)} (\hat{\theta} - \theta) \right) d\mathbf{x} \right)^2 = 1$$

$$\Rightarrow \left(\int_{-\infty}^{\infty} \left(\frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial \theta} \right)^2 f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) d\mathbf{x} \right) \left(\int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) d\mathbf{x} \right) \geq 1$$

Note

$$\int_{-\infty}^{\infty}(\hat{\theta}-\theta)^2 f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)d\mathbf{x} = \text{var}(\hat{\theta}) \qquad (*)$$

and

$$\int_{-\infty}^{\infty}\left(\frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial\theta}\right)^2 f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)d\mathbf{x} = E\left\{\left(\frac{\partial \ln(f_{\mathbf{x}|\theta}(\mathbf{x}|\theta))}{\partial\theta}\right)^2\right\} \qquad (**)$$

Thus using $(*)$ and $(**)$ in

$$\left(\int_{-\infty}^{\infty}\left(\frac{\partial \ln[f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)]}{\partial\theta}\right)^2 f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)d\mathbf{x}\right)\left(\int_{-\infty}^{\infty}(\hat{\theta}-\theta)^2 f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)d\mathbf{x}\right) \geq 1$$

$$\Rightarrow \text{var}(\hat{\theta}) \geq \left[E\left\{\left(\frac{\partial \ln(f_{\mathbf{x}|\theta}(\mathbf{x}|\theta))}{\partial\theta}\right)^2\right\}\right]^{-1}$$

with equality iff

$$\frac{\partial}{\partial\theta}\ln(f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)) = k(\hat{\theta}-\theta)$$

QED

Thus the bound in met iff

$$\frac{\partial}{\partial\theta}\ln(f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)) = k(\hat{\theta}-\theta)$$

Let $\theta = \hat{\theta}_{\text{ML}}$ in the above

$$\underbrace{\frac{\partial}{\partial\theta}\ln(f_{\mathbf{x}|\theta}(\mathbf{x}|\theta))\Big|_{\theta=\hat{\theta}_{\text{ML}}}}_{=0 \text{ by ML criteria}} = k(\hat{\theta}-\theta)\Big|_{\theta=\hat{\theta}_{\text{ML}}}$$

Therefore, the RHS must equal zero, or

$$\hat{\theta} = \hat{\theta}_{\text{ML}}$$

Result: If an efficient estimate (one that satisfies the bound with equality) exists, then it is the ML estimate

Note: If an efficient estimator doesn't exist, then we don't know how good $\hat{\theta}_{\text{ML}}$ is