Estimation and regression

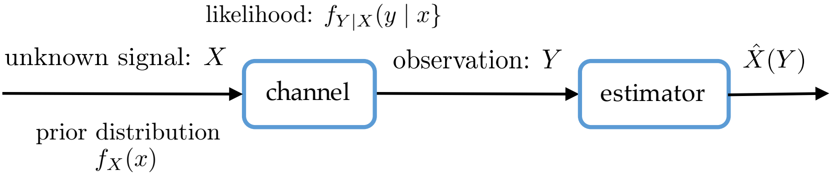
Yuxin Chen

Princeton University, Fall 2018

Outline

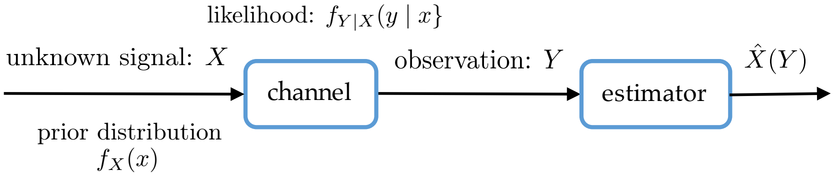
- Minimum mean square error (MMSE) estimation
- Linear minimum mean square error (LMMSE) estimation
- Classical estimation

Estimation



- X is an unknown signal with known prior distribution $f_X(x)$
- X is transmitted over a noisy channel with known likelihood $f_{Y|X}(y \mid x)$
- We observe the output Y and wish to find an estimate $\hat{X}(Y)$ of X

Mean square error (MSE)



- A natural metric to assess the performance of $\hat{X}$ is the mean square error
$$\mathsf{MSE}(\hat{X}) = \mathbb{E}\left[(X - \hat{X}(Y))^2\right]$$
- The estimate that achieves the minimum MSE is called the MMSE estimate of X (given Y)

MMSE estimation

Theorem 6.1

The MMSE estimate of X given the observation Y is

$$\hat{X}(Y) = \mathbb{E}[X|Y] \,.$$

The resulting MSE of $\hat{X}$, i.e. the minimum MSE, is

$$\mathsf{MMSE} = \mathbb{E}[\mathsf{Var}(X|Y)] = \mathsf{Var}(X) - \mathsf{Var}\left(\mathbb{E}[X|Y]\right)$$

Properties of MMSE estimate

- The MMSE estimate is unbiased, since

$$\mathbb{E}\left[\hat{X}\right] = \mathbb{E}\left[\,\mathbb{E}[X \mid Y]\right] = \mathbb{E}[X] \quad \text{(law of iterated expectation)}$$

- If X and Y are independent, then the MMSE estimate is

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X]$$

- For every $Y = y$, the conditional expectation of the estimation error

$$\begin{aligned}\mathbb{E}\left[(X - \hat{X}) \mid Y = y\right] &= \mathbb{E}\left[(X - \mathbb{E}[X \mid Y]) \mid Y = y\right] \\ &= \mathbb{E}\left[X \mid Y = y\right] - \mathbb{E}\left[\,\mathbb{E}[X \mid Y] \mid Y = y\right] = 0\end{aligned}$$

i.e. the error is unbiased for every possible $Y = y$

Properties of MMSE estimate

- The estimation error and the estimate are uncorrelated, i.e. $\mathbb{E}\left[(X-\hat{X})\hat{X}\right]=0$.

  Proof: This follows since

$$\begin{aligned}
\mathbb{E}\left[(X-\hat{X})\hat{X}\right] &= \mathbb{E}\left[\,\mathbb{E}\left[(X-\hat{X})\hat{X} \mid Y\right]\right] \\
&= \mathbb{E}\left[\hat{X}\,\mathbb{E}[(X-\hat{X}) \mid Y]\right] \quad (\hat{X}\text{ is fixed given }Y) \\
&= \mathbb{E}\left[\hat{X}(\underbrace{\mathbb{E}[X \mid Y]-\hat{X}}_{=0})\right] \\
&= 0
\end{aligned}$$

$\square$

In fact, the estimation error is uncorrelated to any function $g(Y)$ of Y (exercise)

Properties of MMSE estimate

- MMSE estimate is linear:

  Let $X = aU + V$ and $\hat{U}$ and $\hat{V}$ be the MMSE estimates of U and V, respectively. Then, the MMSE estimate of X is

$$\hat{X} = a\hat{U} + \hat{V}$$

Proof: This follows since

$$\hat{X} = \mathbb{E}[aU + V \mid Y] = a\underbrace{\mathbb{E}[U \mid Y]}_{\hat{U}} + \underbrace{\mathbb{E}[V \mid Y]}_{\hat{V}}$$

Proof of Theorem 6.1

To start with, we show that in the absence of any observation, the mean of X is its MMSE estimate.

Lemma 6.2

$\min_a \mathbb{E}\left[(X-a)^2\right] = \mathsf{Var}(X)$ *and the minimum is achieved by* $a = \mathbb{E}[X]$.

Proof: To show this, consider

$$\begin{aligned}
\mathbb{E}\left[(X-a)^2\right] &= \mathbb{E}\left[\left(X - \mathbb{E}[X] \; + \; \mathbb{E}[X] - a\right)^2\right] \\
&= \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] \; + \; \left(\mathbb{E}[X] - a\right)^2 + \\
&\qquad 2\,\mathbb{E}(X - \mathbb{E}[X])(\mathbb{E}[X] - a) \\
&= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \; + \; \left(\mathbb{E}[X] - a\right)^2 \geq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]
\end{aligned}$$

Equality holds iff $a = \mathbb{E}[X]$. $\square$

Proof of Theorem 6.1

We then use this fact to show that $\mathbb{E}[X|Y]$ is the MMSE estimate.

First write

$$\mathbb{E}\left[(X - \hat{X}(Y))^2\right] = \mathbb{E}_Y\left[\mathbb{E}_X[(X - \hat{X}(Y))^2 \mid Y]\right]$$

From the previous fact, we know that for every $Y = y$ the minimum value for $\mathbb{E}_X\left[(X - \hat{X}(y))^2 \mid Y = y\right]$ is obtained when $\hat{X}(y) = \mathbb{E}[X \mid Y = y]$. Therefore the overall MSE is minimized for $\hat{X}(Y) = \mathbb{E}[X \mid Y]$

Remark: $\mathbb{E}[X \mid Y]$ minimizes the MSE conditioned on every $Y = y$ and not just its average over Y

Proof of Theorem 6.1

To find the minimum MSE, consider

$$\begin{aligned}\mathbb{E}\left[(X-\mathbb{E}(X|Y))^2\right] &= \mathbb{E}_Y\left[\mathbb{E}_X\left[(X-\mathbb{E}[X \mid Y])^2|Y\right]\right] \\ &= \mathbb{E}_Y\left[\mathsf{Var}(X|Y)\right]\end{aligned}$$

Finally, by the law of conditional variance,

$$\mathbb{E}\left[\mathsf{Var}(X \mid Y)\right] = \mathsf{Var}(X) - \mathsf{Var}(\mathbb{E}[X \mid Y]),$$

i.e. the minimum MSE is the difference between the variance of the signal and the variance of the MMSE estimate

Example

Let $Y \sim \mathsf{Unif}[-1, 1]$ and $X = Y^2$

The MMSE estimate of X given Y is

$$\mathbb{E}[X \mid Y] = Y^2$$

Example: additive Gaussian noise channel

Consider a communication channel with input $X \sim \mathcal{N}(\mu, P)$, noise $Z \sim \mathcal{N}(0, N)$, and output $Y = X + Z$, where X and Z are independent

Question: find the MMSE estimate of X given Y

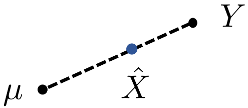
Example: additive Gaussian noise channel

From our previous results on the conditional distribution of jointly Gaussian r.v.s,

$$X \mid \{Y = y\} \sim \mathcal{N}\left(\frac{P}{P+N}y + \frac{N}{P+N}\mu\,,\ \frac{PN}{P+N}\right)$$

Thus, the MMSE estimate is

$$\hat{X} = \mathbb{E}[X|Y] = \underbrace{\frac{P}{P+N}Y + \frac{N}{P+N}\mu}_{\text{convex combination of } Y \text{ and } \mu}$$



Scalar linear estimation

- In general, the MMSE estimate $\mathbb{E}[X \mid Y]$ is difficult to determine, because the posterior density $f_{X|Y}(x \mid y)$ is not easily determined
- We typically have estimates only of the first and second moments of the signal and the observation, i.e., means, variances, and covariance of X and Y. However, they are in general insufficient for computing the MMSE estimate

Scalar linear estimation

- One useful and widely used compromise is to restrict the estimate to be a linear function of the observation.

- As we shall see, 1st and 2nd moments are sufficient to compute the linear MMSE (LMMSE) estimate of X given Y, i.e. the estimate of the form

$$\hat{X} = aY + b$$

that minimizes the mean square error

$$\mathsf{MSE} = \mathbb{E}\left[(X - \hat{X})^2\right]$$

LMMSE estimate

Theorem 6.3

The LMMSE estimate of X given Y is

$$\hat{X} = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(Y)}(Y - \mathbb{E}[Y]) + \mathbb{E}[X]$$
$$= \rho_{X,Y}\sigma_X \left(\frac{Y - \mathbb{E}[Y]}{\sigma_Y}\right) + \mathbb{E}[X]$$

and its MSE is given by

$$\mathsf{MSE} = \mathsf{Var}(X) - \frac{\mathsf{Cov}^2(X,Y)}{\mathsf{Var}(Y)} = (1 - \rho_{X,Y}^2)\mathsf{Var}(X)$$

- The closer that $\rho_{X,Y}$ is to ± 1, the more that uncertainty about X is reduced

Properties of LMMSE estimate

- $\mathbb{E}[\hat{X}] = \mathbb{E}[X]$, i.e. LMMSE estimate is unbiased (also true for MMSE estimate)

- If $\rho_{X,Y} = 0$, i.e. X and Y are uncorrelated, then $\hat{X} = \mathbb{E}[X]$ (independent of the observation Y)

Properties of LMMSE estimate

- If $\rho_{X,Y} = \pm 1$, i.e. $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ are linearly dependent, then the LMMSE estimate is perfect

- Linearity: Let $X = aU + V$ and $\hat{U}$ and $\hat{V}$ be the LMMSE estimates of U and V, respectively
  Then, the LMMSE estimate of X is

$$\hat{X} = a\hat{U} + \hat{V}$$

Proof of Theorem 6.3

For any given a, we know from Lemma 6.2 that the MMSE estimate of $X - aY$ is its mean $\mathbb{E}[X] - a\,\mathbb{E}[Y]$; hence,

$$b = \mathbb{E}[X] - a\,\mathbb{E}[Y]$$

This reduces the problem to finding the coefficient a that minimizes

$$\mathbb{E}[(X - \mathbb{E}[X]) - a(Y - \mathbb{E}[Y])]^2 = \mathbb{E}[(X - \mathbb{E}[X]) - (\hat{X} - \mathbb{E}[X])]^2,$$

i.e. the problem reduces to finding $\hat{X} - \mathbb{E}[X] = a(Y - \mathbb{E}[Y])$ that minimizes the MSE

The optimal a can be found using calculus (see Chapter 8.3, Oppenheim & Verghese). Here, we will use a geometric argument, which might be more enlightening

Aside: vector space

First we introduce some background needed for the geometric argument

- A vector space $\mathcal{V}$ (e.g. Euclidean space) consists of a set of vectors that are closed under two operations
  - vector addition: if $v_1, v_2 \in \mathcal{V}$ then $v_1 + v_2 \in \mathcal{V}$
  - scalar multiplication: if $a \in \mathbb{R}$ and $v \in \mathcal{V}$, then $av \in \mathcal{V}$

Aside: inner product

- An inner product is a real-valued operation $\langle u, v \rangle$ satisfying the three conditions:
  - commutativity: $\langle u, v \rangle = \langle v, u \rangle$
  - linearity: $\langle au + v \, , w \rangle = a\langle u, w \rangle + \langle v, w \rangle$
  - nonnegativity: $\langle u, u \rangle \geq 0$ and $\langle u, u \rangle = 0$ iff $u = 0$

Aside: inner product space

- The norm of u is defined as $\|u\| = \sqrt{\langle u, u \rangle}$
- u and v are orthogonal (written $u \perp v$) if $\langle u, v \rangle = 0$
- A vector space with an inner product is called an inner product space

  Example: Euclidean space with dot product

Inner product space for random variables

View $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ as vectors in an inner product space $\mathcal{V}$ that consists of all zero-mean random variables defined over the same probability space, with

- vector addition: $V_1 + V_2 \in \mathcal{V}$
  adding two zero-mean r.v.s yields a zero-mean r.v.

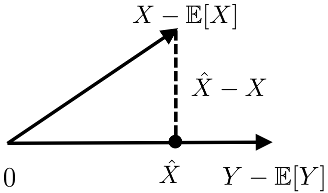
- scalar multiplication: $aV \in \mathcal{V}$
  multiplying a zero-mean r.v. by a constant yields a zero-mean r.v.

- inner product: $\langle V_1, V_2 \rangle = \mathbb{E}[V_1 V_2]$
  exercise: check that this is a legitimate inner product

- norm of V: $\|V\| = \sqrt{\mathbb{E}[V^2]} = \sigma_V$

Proof of Theorem 6.3

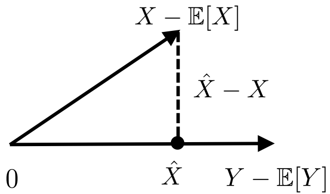
We have the following picture for the r.v.s $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$:



inner product	$\Longleftrightarrow$	Cov(X, Y)
norm of $X - \mathbb{E}[X]$	$\Longleftrightarrow$	σ_X
norm of $Y - \mathbb{E}[Y]$	$\Longleftrightarrow$	σ_Y
$\cos\theta$	$\Longleftrightarrow$	$\rho_{X,Y}$

Orthogonality principle

LMMSE problem can now be recast as a geometry problem



- Find a vector $\hat{X} - \mathbb{E}[X] = a(Y - \mathbb{E}[Y])$ that minimizes $\|X - \hat{X}\|$
- Clearly $(X - \hat{X}) \perp (Y - \mathbb{E}[Y])$ minimizes $\|X - \hat{X}\|$, i.e.,

$$\mathbb{E}\left[(X - \hat{X})(Y - \mathbb{E}[Y])\right] = 0 \implies a = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(Y)}$$

- This argument is called the orthogonality principle.

Example

Let $Y \sim \mathsf{Unif}[-1,1]$ and $X = Y^2$. To find the LMMSE estimate we compute

$$\begin{aligned}\mathbb{E}[Y] &= 0\\ \mathbb{E}[X] &= \int_{-1}^{1} \frac{1}{2} y^2 \,\mathrm{d}y = \frac{1}{3}\\ \mathsf{Cov}(X,Y) &= \mathbb{E}[XY] - 0 = \mathbb{E}[Y^3] = 0\end{aligned}$$

Therefore, LMMSE estimate is $\hat{X} = \mathbb{E}[X] = 1/3$, which completely ignores the observation Y

Vector linear estimation

- Let $X \sim f_X(x)$ be a *scalar* r.v. representing the signal and let $\boldsymbol{Y} = [Y_1, \cdots, Y_n]^\top$ be an n-dimensional RV representing the observations

- The MMSE estimate of X given $\boldsymbol{Y}$ is the conditional expectation $\mathbb{E}[X \mid \boldsymbol{Y}]$. This is often not practical to compute either because the conditional PDF of X given $\boldsymbol{Y}$ is not known or because of high computational cost

Vector linear estimation

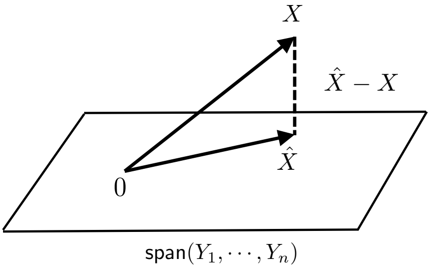
- The LMMSE estimate is often much easier to find since it depends only on the means, variances, and covariances of the r.v.s involved

- To find the LMMSE estimate, first assume that $\mathbb{E}[X] = 0$ and $\mathbb{E}[\boldsymbol{Y}] = \mathbf{0}$. The problem reduces to finding a n-dimensional vector $\boldsymbol{h}$ such that

$$\hat{X} = \boldsymbol{h}^\top \boldsymbol{Y} = \sum_{i=1}^{n} h_i Y_i$$

minimizes the MSE $= \mathbb{E}\left[(X - \hat{X})^2\right]$

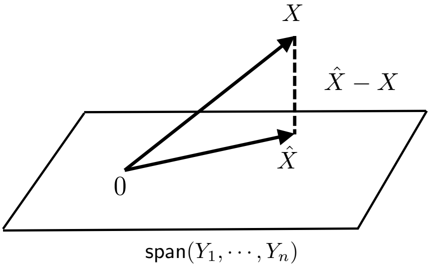
Orthogonality principle



- View the r.v.s $X, Y_1, Y_2, \ldots, Y_n$ as vectors in the inner product space consisting of all zero mean r.v.s

- The linear estimation problem reduces to a geometry problem: find the vector $\hat{X}$ that is closest to X (in norm of error $X - \hat{X}$)

Orthogonality principle



To minimize MSE $= \|X - \hat{X}\|^2$, we choose $\hat{X}$ so that the error vector $X - \hat{X}$ is orthogonal to the subspace spanned by the observations $Y_1, Y_2, \ldots, Y_n$, i.e.,

$$\mathbb{E}\left[(X - \hat{X})Y_i\right] = 0\,, \quad i = 1, 2, \ldots, n\,,$$

$$\Rightarrow \quad \underbrace{\mathbb{E}[Y_i X] = \mathbb{E}[Y_i \hat{X}] = \sum_{j=1}^{n} h_j\, \mathbb{E}[Y_i Y_j]\,, \quad i = 1, \ldots, n}_{\text{a system of } n \text{ linear equations about } n \text{ unknowns } \{h_j\}_{1 \le j \le n}} \qquad (6.1)$$

Orthogonality principle

- Define the cross covariance of $\boldsymbol{Y}$ and X as the n-vector

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}X} = \mathbb{E}\left[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(X - \mathbb{E}[X])\right] = \begin{bmatrix} \sigma_{Y_1 X} \\ \sigma_{Y_2 X} \\ \vdots \\ \sigma_{Y_n X} \end{bmatrix}$$

  For $n = 1$ this is simply the covariance

- The equations (6.1) can be written in vector form as $\boldsymbol{\Sigma}_{\boldsymbol{Y}} \boldsymbol{h} = \boldsymbol{\Sigma}_{\boldsymbol{Y}X}$

- If $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ is nonsingular, we can solve the equations to obtain $\boldsymbol{h} = \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{Y}X}$

LMMSE estimate

- Thus, if $\Sigma_{\boldsymbol{Y}}$ is nonsingular then the LMMSE estimate is:

$$\hat{X} = \boldsymbol{h}^\top \boldsymbol{Y} = \boldsymbol{\Sigma}_{\boldsymbol{Y}X}^\top \Sigma_{\mathbf{Y}}^{-1} \boldsymbol{Y}$$

- Compare this to the scalar case, where $\hat{X} = \frac{\mathsf{Cov}(X,Y)}{\sigma_Y^2} Y$
- Now to find the minimum MSE, consider

$$\begin{aligned}
\mathsf{MSE} &= \mathbb{E}\left[(X - \hat{X})^2\right] \\
&= \mathbb{E}\left[(X - \hat{X})X\right] - \mathbb{E}\left[(X - \hat{X})\hat{X}\right] \\
&= \mathbb{E}\left[(X - \hat{X})X\right], \text{ since by orthogonality } (X - \hat{X}) \perp \hat{X} \\
&= \mathbb{E}[X^2] - \mathbb{E}[\hat{X}X] \\
&= \mathsf{Var}(X) - \mathbb{E}\left[\boldsymbol{\Sigma}_{\boldsymbol{Y}X}^\top \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \boldsymbol{Y} X\right] = \mathsf{Var}(X) - \boldsymbol{\Sigma}_{\boldsymbol{Y}X}^\top \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{Y}X}
\end{aligned}$$

LMMSE estimate

- Compare this to the scalar case, where minimum MSE is $\text{Var}(X) - \frac{\text{Cov}(X,Y)^2}{\sigma_Y^2}$
- If X or $\boldsymbol{Y}$ have nonzero mean, the LMMSE estimate $\hat{X} = h_0 + \boldsymbol{h}^\top \boldsymbol{Y}$ is determined by first finding the MMSE linear estimate of $X - \mathbb{E}[X]$ given $\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]$ (minimum MSE for $\hat{X}'$ and $\hat{X}$ are the same), which is $\hat{X}' = \boldsymbol{\Sigma}_{\boldsymbol{Y}X}^\top \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])$, and then setting $\hat{X} = \hat{X}' + \mathbb{E}[X]$ (since $\mathbb{E}[\hat{X}] = \mathbb{E}[X]$ is necessary)

Example

Let X be the r.v. representing a signal with mean μ and variance P. The observations are $Y_i = X + Z_i$, for $i = 1, 2, \ldots, n$, where the Z_i are zero mean uncorrelated noise with variance N, and X and Z_i are also uncorrelated

Find the LMMSE estimate of X given $\boldsymbol{Y}$ and its MSE

Example

- To find the LMMSE estimate for general n, first let $X' = X - \mu$ and $Y_i' = Y_i - \mu$. Thus X' and $\boldsymbol{Y}'$ are zero mean

- The LMMSE estimate of X' given $\boldsymbol{Y}'$ is given by $\hat{X}_n' = \boldsymbol{h}^\top \boldsymbol{Y}'$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} \boldsymbol{h} = \boldsymbol{\Sigma}_{\boldsymbol{Y}X}, \quad \text{thus}$$

$$\begin{bmatrix} P+N & P & \cdots & P \\ P & P+N & \cdots & P \\ \vdots & \vdots & \ddots & \vdots \\ P & P & \cdots & P+N \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \begin{bmatrix} P \\ P \\ \vdots \\ P \end{bmatrix}$$

Example

By symmetry, $h_1 = h_2 = \cdots = h_n = \frac{P}{nP+N}$. Thus

$$\hat{X}'_n = \frac{P}{nP+N} \sum_{i=1}^{n} Y'_i$$

Therefore

$$\begin{aligned}\hat{X}_n &= \frac{P}{nP+N}\left(\sum_{i=1}^{n}(Y_i - \mu)\right) + \mu \\ &= \frac{P}{nP+N}\left(\sum_{i=1}^{n} Y_i\right) + \frac{N}{nP+N}\mu\end{aligned}$$

Example

If $\mu = 0$, then

$$\hat{X}_n = \frac{nP}{nP + N}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$

- $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is sample mean, which is a sufficient statistic for this case
- The estimate is obtained by "shrinking" the sample mean towards zero (this is an instance of the so-called "shrinkage estimator")

Classical estimation

This is the scenario where the parameter (or transmitted signal) X is not random, but is rather viewed as an unknown constant

Given observations $\boldsymbol{Y} = [Y_1, \cdots, Y_n]^\top$, an estimator is a random variable of the form $\hat{X}_n = g(\boldsymbol{Y})$.

- We call $\hat{X}_n$ unbiased if $\mathbb{E}[\hat{X}_n] = X$ for every possible value of X
- We call $\hat{X}_n$ asymptotically unbiased if $\lim_{n\to\infty} \mathbb{E}[\hat{X}_n] = X$ for every possible value of X
- We call $\hat{X}_n$ consistent if for every possible value of X , $\hat{X}_n$ converges to X with probability approaching 1

Maximum likelihood estimation (MLE)

The maximum likelihood (ML) estimate is a value of the parameter that maximizes the likelihood, namely,

$$\hat{X}_n^{\mathsf{mle}} = \arg\max_x p_{\mathbf{Y}|X}(y_1, \cdots, y_n \mid x)$$

If the n observations are independent, then

$$\begin{aligned}\hat{X}_n^{\mathsf{mle}} &= \arg\max_x \prod_{i=1}^n p_{Y|X}(y_i \mid x) \\ &= \underbrace{\arg\max_x \sum_{i=1}^n \log p_{Y|X}(y_i \mid x)}_{\text{often analytically or computationally more convenient}}\end{aligned}$$

Example: biased coin

Suppose we wish to estimate the probability of heads, denoted by $X \in [0,1]$, of a biased coin. We consider n independent tosses $\{Y_1, \cdots, Y_n\}$ and let k be the number of heads observed.

To find the MLE, we note that the likelihood function is given by

$$f_{\boldsymbol{Y}|X}(y_1, \cdots, y_n \mid x) = x^k (1-x)^{n-k}$$

To find the MLE, differentiating $x^k(1-x)^{n-k}$ w.r.t. x and setting it to zero, we obtain

$$kx^{k-1}(1-x)^{n-k} - (n-k)x^k(1-x)^{n-k-1} = 0,$$

$$\Longrightarrow \qquad \hat{X}^{\mathsf{mle}} = \frac{k}{n} = \frac{Y_1 + \cdots + Y_n}{n}$$

Example: biased coin

$$\hat{X}^{\mathsf{mle}} = \frac{Y_1 + \cdots + Y_n}{n}$$

We can thus see that

- $\hat{X}^{\mathsf{mle}}$ is unbiased, namely, $\mathbb{E}[\hat{X}^{\mathsf{mle}}] = \mathbb{E}\left[\frac{Y_1+\cdots+Y_n}{n}\right] = X$

We can also see that under the uniform prior $X \sim \mathsf{Unif}(0,1)$, the MMSE estimate of X given k (the number of heads observed) is (exercise)

$$\hat{X}^{\mathsf{mmse}} = \mathbb{E}[X \mid k] = \frac{k+1}{n+2}$$

When $n \to \infty$, MMSE estimate and MLE coincide

Example: estimating mean and variance

Consider estimating the mean μ and variance v of a normal distribution using n i.i.d. samples $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, v)$. The corresponding likelihood function is

$$\begin{aligned} f_{\boldsymbol{Y}|\mu,v}(y_1, \cdots, y_n \mid \mu, v) &= \prod_{i=1}^{n} f_{Y_i|\mu,v}(y_i \mid \mu, v) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(y_i-\mu)^2}{2v}} \\ &= \frac{1}{(2\pi v)^{\frac{n}{2}}} \exp\left(-\frac{n\overline{s}_n^2}{2v}\right) \exp\left(-\frac{n(m_n-\mu)^2}{2v}\right), \end{aligned}$$

where m_n and $\overline{s}_n^2$ are respectively the realized values of

$$M_n = \frac{1}{n}\sum_{i=1}^{n} Y_n \qquad \text{and} \qquad \overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_n - M_n)^2.$$

Example: estimating mean and variance

The log-likelihood function is

$$\log f_{\boldsymbol{Y}|\mu,v} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log v - \frac{n\overline{s}_n^2}{2v} - \frac{n(m_n - \mu)^2}{2v}.$$

Setting to zero the derivatives of this function w.r.t. μ and v, we have

$$\hat{\mu} = m_n \qquad \text{and} \qquad \hat{v} = \overline{s}_n^2.$$

Remark: note that M_n is the *sample mean* (which is unbiased), while $\overline{S}_n^2$ can be viewed as a *sample variance*. One can check that $\overline{S}_n^2$ is asymptotically unbiased.

Properties of MLE

MLE has several appealing properties:

- **Invariance principle:** if $\hat{X}_n^{\mathsf{mle}}$ is the MLE of X, then for any one-to-one function h of X, the MLE of the parameter $\zeta = h(X)$ is simply $h(\hat{X}_n)$
- **Consistency:** under very mild technical assumptions, MLE is consistent
- **Asymptotic normality:** the distribution of $\frac{\hat{X}_n^{\mathsf{mle}} - x}{\sigma(\hat{X}_n^{\mathsf{mle}})}$ approaches a standard normal distribution, where $\sigma^2(\hat{X}_n^{\mathsf{mle}})$ is the variance of $\hat{X}_n^{\mathsf{mle}}$

Optimal unbiased estimator?

One might often want to find the "best" *unbiased* estimator. To this end, we can adopt the following approaches

1. Find a fundamental lower bound, say $B(x)$, on the variance of *any* unbiased estimator of X

2. Find an unbiased estimator $\hat{X}$ of X that satisfies

$$\mathrm{Var}_{X=x}(\hat{X}) = B(x)$$

Cramér-Rao lower bound (optional)

Theorem 6.4

Let $Y_1, \cdots, Y_n$ be n i.i.d. samples with conditional density $f_{Y|X}$. Let $W(\boldsymbol{Y}) = W(Y_1, \cdots, Y_n)$ be any unbiased *estimator. Then under mild technical conditions, we have*

$$\mathsf{Var}_{X=x}\left(W(\boldsymbol{Y})\right) \geq \frac{1}{\underbrace{n\mathbb{E}_{X=x}\left[\left(\frac{\partial}{\partial x}\log f_{Y|X}(y \mid x)\right)^2\right]}_{:=\mathcal{I}\ (\textit{Fisher information of a sample})}}$$

As the Fisher information of a sample gets larger, we have "more information" about the unknown parameter X, and hence a smaller bound on the variance of the best unbiased estimator

Optimality of MLE (optional)

When the number n of samples grows (i.e. $n \to \infty$), one has

$$\sqrt{n}(\hat{X}^{\mathsf{mle}} - X) \;\sim\; \mathcal{N}(0, \mathcal{I}^{-1})$$

under mild technical conditions.

In other words, the MLE is asymptotically efficient, in the sense that it achieves the Cramér-Rao lower bound when $n \to \infty$

Example: estimating variance

Consider estimating the variance v of a normal distribution using n i.i.d. samples $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, v)$, where μ is known. The corresponding likelihood function is

$$f_{\boldsymbol{Y}|v}(y_1, \cdots, y_n \mid \mu, v) = \prod_{i=1}^{n} f_{Y_i|\mu,v}(y_i \mid \mu, v) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(y_i-\mu)^2}{2v}}$$

The log-likelihood function is

$$\log f_{\boldsymbol{Y}|v} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log v - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2v}.$$

Setting to zero the derivatives of this function w.r.t. v, we have

$$\hat{v}^{\mathsf{mle}} = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n}.$$

which obeys (exercise!)

$$\mathsf{Var}(\hat{v}^{\mathsf{mle}}) = \frac{2v^2}{n}$$

Example: estimating variance

We then compute the CR lower bound.

$$\frac{\partial^2}{\partial v^2} \log f_{Y_i|v}(y) = \frac{1}{2v^2} - \frac{(y-\mu)^2}{v^3}$$

and

$$\mathcal{I} = -\mathbb{E}\left[\frac{\partial^2}{\partial v^2} \log f_{Y_i|v}(y_i)\right] = -\frac{1}{2v^2} + \mathbb{E}\left[\frac{(y-\mu)^2}{v^3}\right] = \frac{1}{2v^2}.$$

Thus, for any unbiased estimator $\hat{v}$, the CF bound says

$$\mathsf{Var}\left(\hat{v}\right) \geq \frac{1}{n\mathcal{I}} = \frac{2v^2}{n}.$$

Clearly, the MLE $\hat{v}^{\mathsf{mle}}$ attains this bound

Reference

[1] "*Lecture notes for Statistical Signal Processing*," A. El Gamal.

[2] "*Signals, Systems, and Inference*," A. Oppenheim, G. Verghese.

[3] "*Introduction to probability (2nd Edition)*," D. Bertsekas, J. Tsitsiklis, *Athena Scientific*, 2008.