Chapter 1

Estimation theory

In this chapter, an introduction to *estimation theory* is provided. The objective of an estimation problem is to infer the value of an unknown quantity, by using information concerning other quantities (the *data*).
Depending on the type of *a priori* information available on the unknown quantity to be estimated, two different settings can be considered:

- Parametric estimation (the parameter vector to be estimated is deterministic);
- Bayesian estimation (the aim is to estimate a random variable).

In this course, we will focus on the parametric estimation techniques.

1.1 Parametric estimation

The aim of a parametric estimation problem is to estimate a deterministic quantity θ from observations of the random variables $\boldsymbol{y}_1, \dots \boldsymbol{y}_n$.

1.1.1 Problem formulation

Let:

- $\theta \in \Theta \subseteq \mathbb{R}^p$, an unknown vector of *parameters*;
- $\boldsymbol{y} = (\boldsymbol{y}_1, \dots \boldsymbol{y}_n)^T \in \mathcal{Y} \subseteq \mathbb{R}^n$ a vector of random variables, hereafter called *observations* or *measurements*;
- $F_{\boldsymbol{y}}^{\theta}(y), f_{\boldsymbol{y}}^{\theta}(y)$ the cumulative distribution function and the probability density function, respectively, of the observation vector $\boldsymbol{y}$, which depend on the unknown vector θ.

The set Θ, to which the parameter vector θ belongs, is referred to as the *parameter set*. It represents the *a priori* information available on the admissible values of the vector θ. If all values are admissible, $\Theta = \mathbb{R}^p$.

The set $\mathcal{Y}$, containing all the values that the random vector $\boldsymbol{y}$ may take, is known as *observation set* (or *measurement set*). It is assumed that the cdf $F_{\boldsymbol{y}}^{\theta}(y)$ (or equivalently the pdf $f_{\boldsymbol{y}}^{\theta}(y)$) is parameterized by the p parameters $\theta \in \mathbb{R}^p$ (which means that such parameters enter in the expressions of those functions). Hereafter, the word *parameter* will be used to denote the entire unknown vector θ. To emphasize the special case $p = 1$, we will sometimes use the expression *scalar parameter*.

We are now ready to formulate the general version of a parametric estimation problem.

Problem 1.1. *Estimate the unknown parameter $\theta \in \Theta$, by using an observation y of the random vector $\boldsymbol{y} \in \mathcal{Y}$.*

In order to solve Problem 1.1, one has to construct an *estimator*.

Definition 1.1. An *estimator* $T(\cdot)$ of the parameter θ is a function that maps the set of observations to the parameter set:

$$T : \mathcal{Y} \to \Theta.$$

The value $\hat{\theta} = T(y)$, returned by the estimator when applied to the observation y of $\boldsymbol{y}$, is called *estimate* of θ.

An estimator $T(\cdot)$ defines a rule that associates to each realization y of the measurement vector $\boldsymbol{y}$, the quantity $\hat{\theta} = T(y)$ which is an estimate of θ.

Notice that $\hat{\theta}$ can be seen as a realization of the random variable $T(\boldsymbol{y})$; in fact, since $T(\boldsymbol{y})$ is a function of the random variable $\boldsymbol{y}$, the estimate $\hat{\theta}$ is a random variable itself.

1.1.2 Properties of an estimator

According to Definition 1.1, the class of possible estimators is infinite. In order to characterize the quality of an estimator, it is useful to introduce some desired properties.

Unbiasedness

A first desirable property is that the expected value of the estimate $\hat{\theta} = T(y)$ be equal to the actual value of the parameter θ.

Definition 1.2. An estimator $T(\boldsymbol{y})$ of the parameter θ is *unbiased* (or *correct*) if

$$\mathbf{E}^{\theta}[T(\boldsymbol{y})] = \theta, \quad \forall \theta \in \Theta. \tag{1.1}$$

In the above definition we used the notation $\mathbf{E}^{\theta}[\cdot]$, which stresses the dependency on θ of the expected value of $T(\boldsymbol{y})$, due to the fact that the pdf of $\boldsymbol{y}$ is parameterized by θ itself. In fact,

$$\mathbf{E}^{\theta}[T(\boldsymbol{y})] = \int_{-\infty}^{\infty} T(y) f_{\boldsymbol{y}}^{\theta}(y)\, dy$$

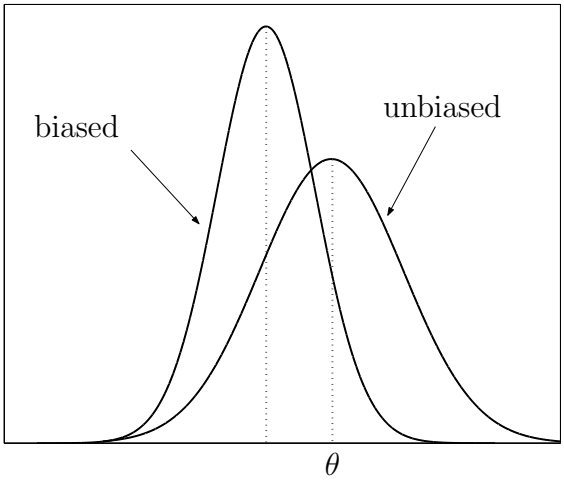

Figure 1.1: Probability density function of an unbiased estimator and of a biased one.

The unbiasedness condition (1.1) guarantees that the estimator $T(\cdot)$ does not introduce systematic errors, i.e., errors that are not averaged out even when considering an infinite amount of observations of $\boldsymbol{y}$. In other words, $T(\cdot)$ does not overestimate neither underestimate θ, on average (see Fig. 1.1).

Example 1.1. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be random variables with mean m. The quantity

$$\overline{\boldsymbol{y}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i \tag{1.2}$$

is the so-called *sample mean*. It is easy to verify that $\overline{\boldsymbol{y}}$ is an unbiased estimator of m. Indeed, due to the linearity of the expected value operator, one has

$$\mathbf{E}[\overline{\boldsymbol{y}}] = \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[\boldsymbol{y}_i] = \frac{1}{n} \sum_{i=1}^{n} m = m.$$

△

Example 1.2. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be scalar random variables, independent and identically distributed (i.i.d.) with mean m and variance σ^2. The quantity

$$\hat{\sigma}_{\boldsymbol{y}}^2 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \overline{\boldsymbol{y}})^2$$

is a biased estimator of the variance σ^2. Indeed, from (1.2) one has

$$\begin{aligned}\mathbf{E}\left[\hat{\sigma}^2_{\boldsymbol{y}}\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left(\boldsymbol{y}_i - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}_j\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}\mathbf{E}\left[\left(n\boldsymbol{y}_i - \sum_{j=1}^{n}\boldsymbol{y}_j\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}\mathbf{E}\left[\left(n(\boldsymbol{y}_i - m) - \sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right)^2\right].\end{aligned}$$

However,

$$\begin{aligned}\mathbf{E}\left[\left(n(\boldsymbol{y}_i - m) - \sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right)^2\right] &= n^2\mathbf{E}\left[(\boldsymbol{y}_i - m)^2\right] \\ &\quad - 2n\mathbf{E}\left[(\boldsymbol{y}_i - m)\sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right] + \mathbf{E}\left[\left(\sum_{j=1}^{n}(\boldsymbol{y}_j - m)\right)^2\right] \\ &= n^2\sigma^2 - 2n\sigma^2 + n\sigma^2 \\ &= n(n-1)\sigma^2\end{aligned}$$

because, for the independency assumption, $\mathbf{E}\left[(\boldsymbol{y}_i - m)(\boldsymbol{y}_j - m)\right] = 0$ for $i \neq j$. Therefore,

$$\mathbf{E}\left[\hat{\sigma}^2_{\boldsymbol{y}}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}n(n-1)\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

△

Example 1.3. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be i.i.d. scalar random variables, with mean m and variance σ^2. The quantity

$$S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^2$$

is called *sample variance*. It is straightforward to verify that S^2 is an unbiased estimator of the variance σ^2. In fact, observing that

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2_{\boldsymbol{y}},$$

one has immeantely

$$\mathbf{E}\left[S^2\right] = \frac{n}{n-1}\mathbf{E}\left[\hat{\sigma}^2_{\boldsymbol{y}}\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

△

Notice that, if $T(\cdot)$ is an unbiased estimator of θ, then $g(T(\cdot))$ is *not* in general an unbiased estimator of $g(\theta)$, unless $g(\cdot)$ is a linear function.

Consistency

Another desirable property of an estimator is to provide an estimate that converges to the actual value of θ as the number of measurements grows. Being the estimate a random variable, we need to introduce the notion of convergence in probability.

Definition 1.3. Let $\{\boldsymbol{y}_i\}_{i=1}^{\infty}$ be a sequence of random variables. The sequence of estimators $\hat{\theta}_n = T_n(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ of θ is said to be *consistent* if $\hat{\theta}_n$ converges in probability to θ, for all admissible values of θ, i.e.

$$\lim_{n\to\infty} \mathrm{P}\left(\left\|\hat{\theta}_n - \theta\right\| \geq \varepsilon\right) = 0, \quad \forall \varepsilon > 0, \quad \forall \theta \in \Theta.$$

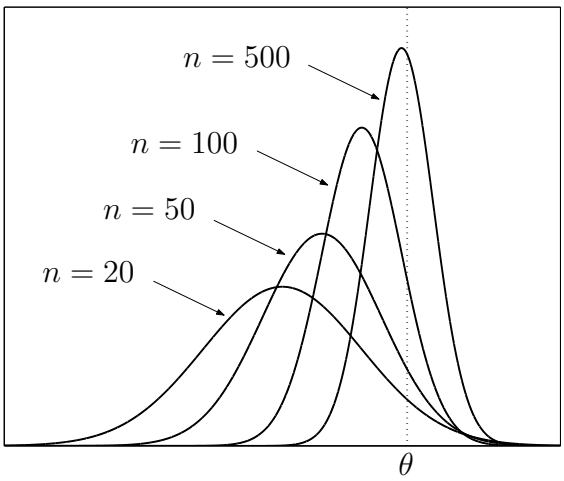

Figure 1.2: Probability density function of a consistent estimator.

Notice that consistency is an asymptotic property of an estimator. It guarantees that, as the number of data goes to infinity, the probability that the estimate differ from the actual value of the parameter goes to zero (see Fig. 1.2).

The next Theorem provides a sufficient condition for consistency of unbiased estimators.

Theorem 1.1. *Let $\hat{\theta}_n$ be a sequence of unbiased estimators of the scalar parameter θ:*

$$\mathbf{E}\left[\hat{\theta}_n\right] = \theta, \quad \forall n, \quad \forall \theta \in \Theta.$$

If

$$\lim_{n\to\infty} \mathbf{E}\left[(\hat{\theta}_n - \theta)^2\right] = 0,$$

then the sequence $\hat{\theta}_n$ is consistent.

Proof

For a random variable $\boldsymbol{x}$, the *Chebishev inequality* holds:

$$\mathrm{P}\left(|\boldsymbol{x} - m_{\boldsymbol{x}}| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbf{E}\left[(\boldsymbol{x} - m_{\boldsymbol{x}})^2\right].$$

Therefore, one has

$$\lim_{n\to\infty} \mathrm{P}\left(\left\|\hat{\theta}_n - \theta\right\| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{1}{\varepsilon^2}\mathbf{E}\left[(\hat{\theta}_n - \theta)^2\right],$$

from which the result follows immediately. □

Therefore, for a sequence of unbiased estimators to be consistent, it is sufficient that the variance of the estimates goes to zero as the number of measurements grows.

Example 1.4. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be i.i.d. random variables with mean m and variance σ^2. In Example 1.1 it has been shown that the sample mean

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$$

is an unbiased estimator of the mean m. Let us now show that it is also a consistent estimator of m. The variance of the estimate is given by

$$\begin{aligned}\mathrm{Var}(\overline{\boldsymbol{y}}) = \mathbf{E}\left[(\overline{\boldsymbol{y}} - m)^2\right] &= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i - m\right)^2\right] \\ &= \frac{1}{n^2}\mathbf{E}\left[\left(\sum_{i=1}^{n}(\boldsymbol{y}_i - m)\right)^2\right] = \frac{\sigma^2}{n}\end{aligned}$$

because the random variables $\boldsymbol{y}_i$ are independent. Therefore,

$$\mathrm{Var}(\overline{\boldsymbol{y}}) = \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty.$$

Hence, due to Theorem 1.1, the sample mean $\overline{\boldsymbol{y}}$ is a consistent estimator of the mean m. △

The result in Example 1.4 is a special case of the following more general celebrated result.

Theorem 1.2. (Law of large numbers)
Let $\{\boldsymbol{y}_i\}_{i=1}^{\infty}$ be a sequence of independent random variables with mean m and finite variance. Then, the sample mean $\overline{\boldsymbol{y}}$ converges to m in probability, i.e., it is a consistent estimator of m.

1.1.3 Performance criteria

Mean square error

A criterion for measuring the quality of the estimate provided by an estimator is the *Mean Square Error*. Let us first consider the case of a scalar parameter ($\theta \in \mathbb{R}$).

Definition 1.4. Let $\theta \in \mathbb{R}$. The *Mean Square Error (MSE)* of an estimator $T(\cdot)$ is defined as

$$\mathrm{MSE}_{T(\cdot)} = \mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - \theta)^2\right]$$

Notice that if an estimator is unbiased, then the MSE is equal to the variance of the estimate $T(\boldsymbol{y})$, and also to the variance of the *estimation error* $T(\boldsymbol{y}) - \theta$. On the other hand, for a biased estimator one has

$$\begin{aligned}
\mathrm{MSE}_{T(\cdot)} &= \mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - m_{\boldsymbol{T(y)}} + m_{\boldsymbol{T(y)}} - \theta)^2\right] \\
&= \mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - m_{\boldsymbol{T(y)}})^2\right] + (m_{\boldsymbol{T(y)}} - \theta)^2 + 2\mathbf{E}^{\theta}\left[\underbrace{(T(\boldsymbol{y}) - m_{\boldsymbol{T(y)}})}_{=0}\underbrace{(m_{\boldsymbol{T(y)}} - \theta)}_{deterministic}\right] \\
&= \mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - m_{\boldsymbol{T(y)}})^2\right] + (m_{\boldsymbol{T(y)}} - \theta)^2
\end{aligned}$$

where $m_{\boldsymbol{T(y)}} = \mathbf{E}\left[T(\boldsymbol{y})\right]$. The above expression shows that the MSE of a biased estimator is the sum of the variance of the estimator and of the square of the deterministic quantity $m_{\boldsymbol{T(y)}} - \theta$, which is called *bias error*. As we will see, the trade off between the variance of the estimator and the bias error is a fundamental limitation in many practical estimation problems.

The MSE can be used to decide which estimator is better within a family of estimators.

Definition 1.5. Let $T_1(\cdot)$ and $T_2(\cdot)$ be two estimators of the parameter θ. Then, $T_1(\cdot)$ is *uniformly preferable* (i.e., "better") to $T_2(\cdot)$ if

$$\mathbf{E}^{\theta}\left[(T_1(\boldsymbol{y}) - \theta)^2\right] \leq \mathbf{E}^{\theta}\left[(T_2(\boldsymbol{y}) - \theta)^2\right], \quad \forall \theta \in \Theta$$

It is worth stressing that in order to be preferable to other estimators, an estimator must provide a smaller MSE for *all* the admissible values of the parameter θ.

The above definitions can be extended quite naturally to the case of a parameter vector $\theta \in \mathbb{R}^p$.

Definition 1.6. Let $\theta \in \mathbb{R}^p$. The *Mean Square Error (MSE)* of an estimator $T(\cdot)$ is defined as

$$\begin{aligned}
\mathrm{MSE}_{T(\cdot)} &= \mathbf{E}^{\theta}\left[\|T(\boldsymbol{y}) - \theta)\|^2\right] \\
&= \mathbf{E}^{\theta}\left[\mathrm{tr}\{(T(\boldsymbol{y}) - \theta)(T(\boldsymbol{y}) - \theta)^T\}\right] \\
&= \mathrm{tr}\{\mathbf{E}^{\theta}\left[(T(\boldsymbol{y}) - \theta)(T(\boldsymbol{y}) - \theta)^T\right]\}
\end{aligned}$$

where $\mathrm{tr}(M)$ denotes the trace of the matrix M.[1]

[1]Remember that for any vector $v \in \mathbb{R}^n$, it holds $\|v\|^2 = v^T v = \mathrm{tr}(vv^T)$.

Notice that, if $T(\boldsymbol{y})$ is unbiased, $\mathrm{MSE}_{T(\cdot)}$ corresponds to the trace of the covariance matrix of the estimation error.

The concept of uniformly preferable estimator is analogous to that in Definition 1.5. It can be also defined in terms of inequality between the corresponding covariance matrices, i.e., $T_1(\cdot)$ is uniformly preferable to $T_2(\cdot)$ if

$$\mathbf{E}^{\theta}\left[(T_1(\boldsymbol{y})-\theta)(T_1(\boldsymbol{y})-\theta)^T\right] \leq \mathbf{E}^{\theta}\left[(T_2(\boldsymbol{y})-\theta)(T_2(\boldsymbol{y})-\theta)^T\right]$$

where the matrix inequality $A \leq B$ means that $B-A$ is a positive semidefinite matrix.

1.1.4 Minimum variance unbiased estimator

Let us restrict our attention to unbiased estimators. Since we have introduced the concept of mean square error, it is natural to look for the estimator which minimizes this performance index.

Definition 1.7. An unbiased estimator $T^*(\cdot)$ of the scalar parameter θ is a *Uniformly Minimum Variance Unbiased Estimator (UMVUE)* if

$$\mathbf{E}^{\theta}\left[(T^*(\boldsymbol{y})-\theta)^2\right] \leq \mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)^2\right], \quad \forall \theta \in \Theta \tag{1.3}$$

for all unbiased estimators $T(\cdot)$ of θ.

Notice that for an estimator to be UMVUE, it has to satisfy the following conditions:

- be unbiased;
- have minimum variance among all unbiased estimators;
- the previous conditions must hold for every admissible value of the parameter θ.

Unfortunately, there are many problems for which there does not exist any $UMVUE$ estimator. For this reason, we often restrict the class of estimators, in order to find the best one within the considered class. A popular choice is that of *linear* estimators, i.e., taking the form

$$T(\boldsymbol{y}) = \sum_{i=1}^{n} a_i \boldsymbol{y}_i, \tag{1.4}$$

with $a_i \in \mathbb{R}$.

Definition 1.8. A linear unbiased estimator $T^*(\cdot)$ of the scalar parameter θ is the *Best Linear Unbiased Estimator (BLUE)* if

$$\mathbf{E}^\theta\left[(T^*(\boldsymbol{y})-\theta)^2\right] \leq \mathbf{E}^\theta\left[(T(\boldsymbol{y})-\theta)^2\right], \quad \forall \theta \in \Theta$$

for every linear unbiased estimator $T(\cdot)$ of θ.

Differently from the UMVUE estimator, the BLUE estimator takes on a simple form and can be easily computed (one has just to find the optimal values of the coefficients a_i).

Example 1.5. Let $\boldsymbol{y}_i$ be independent random variables with mean m and variance $\sigma_{\boldsymbol{i}}^2$, $i = 1, \ldots, n$. Assume the variances $\sigma_{\boldsymbol{i}}^2$ are known. Let us compute the BLUE estimator of m. Being the estimator linear, it takes on the form (1.4). In order to be unbiased, $T(\cdot)$ must satisfy

$$\mathbf{E}^\theta\left[T(\boldsymbol{y})\right] = \mathbf{E}^\theta\left[\sum_{i=1}^n a_i \boldsymbol{y}_i\right] = \sum_{i=1}^n a_i \mathbf{E}^\theta\left[\boldsymbol{y}_i\right] = m\sum_{i=1}^n a_i = m$$

Therefore, we must enforce the constraint

$$\sum_{i=1}^n a_i = 1 \tag{1.5}$$

Now, among all the estimators of form (1.4), with the coefficients a_i satisfying (1.5), we need to find the minimum variance one. Being the observations $\boldsymbol{y}_i$ independent, the variance of $T(\boldsymbol{y})$ is given by

$$\begin{aligned}\mathbf{E}^\theta\left[(T(\boldsymbol{y})-m)^2\right] &= \mathbf{E}^\theta\left[\left(\sum_{i=1}^n a_i\boldsymbol{y}_i - m\right)^2\right] = \mathbf{E}^\theta\left[\left(\sum_{i=1}^n a_i\boldsymbol{y}_i - \sum_{i=1}^n a_i m\right)^2\right] \\ &= \mathbf{E}^\theta\left[\left(\sum_{i=1}^n a_i(\boldsymbol{y}_i - m)\right)^2\right] = \sum_{i=1}^n a_i^2\,\mathbf{E}^\theta\left[(\boldsymbol{y}_i-m)^2\right] = \sum_{i=1}^n a_i^2\sigma_{\boldsymbol{i}}^2.\end{aligned}$$

Summing up, in order to determine the BLUE estimator, we have to solve the following constrained optimization problem:

$$\min_{a_i} \sum_{i=1}^n a_i^2\sigma_{\boldsymbol{i}}^2$$

s.t.

$$\sum_{i=1}^n a_i = 1$$

Let us write the Lagrangian function

$$\mathcal{L}(a_1, \ldots, a_n, \lambda) = \sum_{i=1}^n a_i^2\sigma_{\boldsymbol{i}}^2 + \lambda\left(\sum_{i=1}^n a_i - 1\right)$$

and compute the stationary points by imposing

$$\frac{\partial \mathcal{L}(a_1, \ldots, a_n, \lambda)}{\partial a_i} = 0, \quad i = 1, \ldots, n \tag{1.6}$$

$$\frac{\partial \mathcal{L}(a_1, \ldots, a_n, \lambda)}{\partial \lambda} = 0. \tag{1.7}$$

From (1.7) we obtain the constraint (1.5), while (1.6) implies that

$$2a_i\sigma_{\boldsymbol{i}}^2 + \lambda = 0, \quad i = 1, \ldots, n$$

from which $a_i = -\dfrac{\lambda}{2\sigma_{\boldsymbol{i}}^2},\ i = 1, \ldots, n$. Substituting a_i into (1.5), one gets

$$\lambda = -\frac{1}{\displaystyle\sum_{i=1}^{n} \frac{1}{2\sigma_{\boldsymbol{i}}^2}}$$

and then

$$a_i = \frac{\dfrac{1}{\sigma_{\boldsymbol{i}}^2}}{\displaystyle\sum_{j=1}^{n} \frac{1}{\sigma_{\boldsymbol{j}}^2}}, \quad i = 1, \ldots, n.$$

Therefore, the BLUE estimator of the mean m is given by

$$\hat{\boldsymbol{m}}_{BLUE} = \frac{1}{\displaystyle\sum_{i=1}^{n} \frac{1}{\sigma_{\boldsymbol{i}}^2}} \sum_{i=1}^{n} \frac{1}{\sigma_{\boldsymbol{i}}^2} \boldsymbol{y}_i \tag{1.8}$$

Notice that if all the measurements have the same variance $\sigma_{\boldsymbol{i}}^2 = \sigma^2$, the estimator $\hat{\boldsymbol{m}}_{BLUE}$ boils down to the sample mean

$$\hat{\boldsymbol{m}}_{BLUE} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i = \overline{\boldsymbol{y}}.$$

This means that the BLUE estimator can be seen as a generalization of the sample mean, in the case when the measurements $\boldsymbol{y}_i$ have different accuracy (i.e., different variance $\sigma_{\boldsymbol{i}}^2$). In fact, the BLUE estimator is a weighted average of the observations, in which the weights are inversely proportional to the variance of the measurements or, seen another way, directly proportional to the precision of each observation. Let us assume that for a certain i, $\sigma_{\boldsymbol{i}}^2 \to \infty$. This means that the measurement $\boldsymbol{y}_i$ is completely unreliable. Then, the weight $\frac{1}{\sigma_{\boldsymbol{i}}^2}$ of $\boldsymbol{y}_i$ within $\hat{\boldsymbol{m}}_{BLUE}$ will tend to zero. On the other hand, for an infinitely precise measurement $\boldsymbol{y}_j$ ($\sigma_{\boldsymbol{j}}^2 \to 0$), the corresponding weight $\frac{1}{\sigma_{\boldsymbol{j}}^2}$ will be predominant over all the other weights and the BLUE estimate will approach that measurement, i.e., $\hat{m}_{BLUE} \simeq \boldsymbol{y}_j$. △

1.2 Cramér-Rao bound

This paragraph introduces a fundamental result which establishes a lower bound to the variance of every unbiased estimator of the parameter θ.

Theorem 1.3. (Cramér-Rao bound) *Let $T(\cdot)$ be an unbiased estimator of the scalar parameter θ based on the observations y of the random variables $\boldsymbol{y} \in \mathcal{Y} \in \mathbb{R}^n$, and let that the observation set $\mathcal{Y}$ be independent from θ. Then, under some technical regularity assumptions[2], it holds*

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)^2\right] \geq \left[I_n(\theta)\right]^{-1}, \tag{1.9}$$

where

$$I_n(\theta) = \mathbf{E}^{\theta}\left[\left(\frac{\partial \ln f_{\boldsymbol{y}}^{\theta}(y)}{\partial \theta}\right)^2\right] \tag{1.10}$$

is called Fisher information. *Moreover, if the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent and identically distributed with the same pdf $f_{\boldsymbol{y}_1}^{\theta}(y_1)$, one has*

$$I_n(\theta) = n\, I_1(\theta).$$

When θ is a p-dimensional vector, the Cramér-Rao bound (1.9) becomes

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)\,(T(\boldsymbol{y})-\theta)^T\right] \geq \left[I_n(\theta)\right]^{-1},$$

where the inequality must be intended in matricial sense and the matrix $I_n(\theta) \in \mathbb{R}^{p\times p}$ is the so-called *Fisher information matrix*

$$I_n(\theta) = \mathbf{E}^{\theta}\left[\left(\frac{\partial \ln f_{\boldsymbol{y}}^{\theta}(y)}{\partial \theta}\right)\left(\frac{\partial \ln f_{\boldsymbol{y}}^{\theta}(y)}{\partial \theta}\right)^T\right].$$

Notice that the matrix $\mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)\,(T(\boldsymbol{y})-\theta)^T\right]$ is the covariance matrix of the unbiased estimator $T(\cdot)$.

Theorem 1.3 states that there does not exist any estimator with variance smaller than $[I_n(\theta)]^{-1}$. Notice that $I_n(\theta)$ depends, in general, on the actual value of the parameter θ (because the partial derivatives must be evaluated in θ) which is unknown. For this reason, an approximation of the lower bound is usually computed in practice, by replacing θ with an estimate $\hat{\theta}$. Nevertheless, the Cramér-Rao is also important because it allows to define the key concept of *efficiency* of an estimator.

Definition 1.9. An unbiased estimator $T(\cdot)$ is *efficient* if its variance achieves the Cramér-Rao bound, i.e.

$$\mathbf{E}^{\theta}\left[(T(\boldsymbol{y})-\theta)^2\right] = \left[I_n(\theta)\right]^{-1}.$$

[2]For details, see V.K. Rohatgi, A.K.Md.E. Saleh, "An introduction to probability and statistics", 2nd edition, Wiley Interscience, 2001.

An efficient estimator has the least possible variance among all unbiased estimators (therefore, it is also a UMVUE).

In the special case of i.i.d. observations $\boldsymbol{y}_i$, Theorem 1.3 states that $I_n(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information of a single observation. Therefore, for a fixed θ, the Cramér-Rao bound decreases as $\frac{1}{n}$, as the number of observations n grows.

Example 1.6. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be i.i.d. random variables with mean $m_{\boldsymbol{y}}$ and variance $\sigma^2_{\boldsymbol{y}}$. In Examples 1.1 and 1.4, we have seen that the sample mean

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$$

is a consistent unbiased estimator of the mean $m_{\boldsymbol{y}}$. Being the observations i.i.d., from Theorem 1.3 one has

$$\mathbf{E}^\theta\left[(\overline{\boldsymbol{y}} - m_{\boldsymbol{y}})^2\right] = \frac{\sigma^2_{\boldsymbol{y}}}{n} \geq \left[I_n(\theta)\right]^{-1} = \frac{\left[I_1(\theta)\right]^{-1}}{n}.$$

Let us now assume that the $\boldsymbol{y}_i$ are distributed according to the Gaussian pdf

$$f_{\boldsymbol{y_i}}(y_i) = \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} e^{-\frac{(y_i - m_{\boldsymbol{y}})^2}{2\sigma^2_{\boldsymbol{y}}}}.$$

Let us compute the Fisher information of a single measurement

$$I_1(\theta) = \mathbf{E}^\theta\left[\left(\frac{\partial \ln f^\theta_{\boldsymbol{y_1}}(y_1)}{\partial\theta}\right)^2\right].$$

In this example, the unknown parameter to be estimated is the mean $\theta = m$. Therefore,

$$\frac{\partial \ln f^\theta_{\boldsymbol{y_1}}(y_1)}{\partial\theta} = \frac{\partial}{\partial m}\left(\ln\frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \frac{(y_1 - m)^2}{2\sigma^2_{\boldsymbol{y}}}\right)\Bigg|_{m=m_{\boldsymbol{y}}} = \frac{y - m_{\boldsymbol{y}}}{\sigma^2_{\boldsymbol{y}}},$$

and hence,

$$I_1(\theta) = \mathbf{E}^\theta\left[\frac{(y - m_{\boldsymbol{y}})^2}{\sigma^4_{\boldsymbol{y}}}\right] = \frac{1}{\sigma^2_{\boldsymbol{y}}}.$$

The Cramér-Rao bound takes on the value

$$\left[I_n(\theta)\right]^{-1} = \frac{\left[I_1(\theta)\right]^{-1}}{n} = \frac{\sigma^2_{\boldsymbol{y}}}{n},$$

which is equal to the variance of the estimator $\overline{\boldsymbol{y}}$. Therefore, we can conclude that: *in the case of i.i.d. Gaussian observations, the sample mean is an efficient estimator of the mean.* △

1.3 Maximum Likelihood Estimator

In general, for a given parametric estimation problem, an efficient estimator may not exist. In Example 1.6, it has been shown that the Cramér-Rao bound allows one to check if an estimator is efficient. However, it remains unclear how to find suitable candidates for efficient estimators and, in the case that such candidates turn out to be not efficient, whether it is possible to conclude that for the problem at hand there are no efficient estimators. An answer to these questions is provided by the class of *Maximum Likelihood* estimators.

Definition 1.10. Let $\boldsymbol{y}$ be a vector of observations with pdf $f_{\boldsymbol{y}}^{\theta}(y)$, depending on the unknown parameter $\theta \in \Theta$. The *likelihood function* is defined as

$$L(\theta|y) = f_{\boldsymbol{y}}^{\theta}(y)\,.$$

It is worth remarking that, once the realization y of the random variable $\boldsymbol{y}$ has been observed (i.e., after the data have been collected), the likelihood function depends only on the unknown parameter θ (indeed, we refer to $L(\theta|y)$ as the likelihood of θ "given" y).

A meaningful way to estimate θ is to choose the value that maximizes the probability of the observed data. In fact, by exploiting the meaning of the probability density function, maximizing $f_{\boldsymbol{y}}^{\theta}(y)$ with respect to θ corresponds to choose θ in such a way that the measurement y has the highest possible probability of having been observed, among all feasible scenarios $\theta \in \Theta$.

Definition 1.11. The *Maximum Likelihood (ML) estimator* of the unknown parameter θ is given by

$$T_{ML}(y) = \arg\max_{\theta\in\Theta} L(\theta|y).$$

In several problems, in order to ease the computation, it may be convenient to maximize the so-called *log-likelihood* function:

$$\ln L(\theta|y).$$

Being the natural logarithm a monotonically increasing function, $L(\theta|y)$ and $\ln L(\theta|y)$ achieve their maxima in the same values.

Remark 1.1. Assuming that the pdf $f_{\boldsymbol{y}}^{\theta}(y)$ be a differentiable function of $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$, with Θ an open set, if $\hat{\theta}$ is a maximum for $L(\theta|y)$, it has to be a solution of the equations

$$\left.\frac{\partial L(\theta|y)}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0, \quad i = 1, \ldots, p \tag{1.11}$$

or equivalently of

$$\left.\frac{\partial \ln L(\theta|y)}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0, \quad i = 1, \ldots, p. \tag{1.12}$$

It is worth observing that in many problems, even for a scalar parameter ($p = 1$), equation (1.11) may admit more than one solution. It may also happen that the likelihood function is not differentiable everywhere in Θ or that Θ is not an open set, in which case the maximum can be achieved on the boundary of Θ. For all these reasons, the computation of the maximum likelihood estimator requires to study the function $L(\theta|y)$ over the entire domain Θ (see Exercise 1.5). Clearly, this may be a formidable task for high dimensional parameter vectors.

Example 1.7. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be independent Gaussian random variables, with unknown mean $m_{\boldsymbol{y}}$ and known variance $\sigma^2_{\boldsymbol{y}}$. Let us compute the ML estimator of the mean $m_{\boldsymbol{y}}$.

Being the measurements independent, the likelihood is given by

$$L(\theta|y) = f^{\theta}_{\boldsymbol{y}}(y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} e^{-\frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}}}.$$

In this case, it is convenient to maximize the log-likelihood, which takes on the form

$$\begin{aligned} \ln L(\theta|y) &= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}} \right) \\ &= n \ln \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \sum_{i=1}^{n} \frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}}. \end{aligned}$$

By imposing the condition (1.12), one gets

$$\frac{\partial \ln L(\theta|y)}{\partial \theta} = \frac{\partial}{\partial m} \left(n \ln \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{y}}} - \sum_{i=1}^{n} \frac{(y_i-m)^2}{2\sigma^2_{\boldsymbol{y}}} \right) \Bigg|_{m=\hat{m}_{ML}} = 0,$$

from which

$$\sum_{i=1}^{n} \frac{y_i - \hat{m}_{ML}}{\sigma^2_{\boldsymbol{y}}} = 0,$$

and hence

$$\hat{m}_{ML} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i.$$

Therefore, in this case the ML estimator coincides with the sample mean. Since the observations are i.i.d. Gaussian variables, this estimator is also efficient (see Example 1.6). △

The result in Example 1.7 is not restricted to the specific setting or pdf considered. The following general theorem illustrates the importance of maximum likelihood estimators, in the context of parametric estimation.

Theorem 1.4. *Under the same assumptions for which the Cramér-Rao bound holds, if there exists an efficient estimator $T^*(\cdot)$, then $T^*(\cdot)$ is a maximum likelihood estimator.*

Therefore, if we are looking for an efficient estimator, the only candidates are maximum likelihood estimators.

Example 1.8. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be independent Gaussian random variables, with mean $m_{\boldsymbol{y}}$ and variance $\sigma^2_{\boldsymbol{y}}$, both unknown. Let us compute the Maximum Likelihood estimator of the mean and the variance.

Similarly to what observed in Example 1.7, the log-likelihood turns out to be

$$\ln L(\theta|y) = n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(y_i - m)^2}{2\sigma^2}.$$

The unknown parameter vector to be estimated is $\theta = (m, \sigma^2)^T$, for which condition (1.12) becomes

$$\frac{\partial \ln L(\theta|y)}{\partial \theta_1} = \frac{\partial}{\partial m} \left(n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(y_i - m)^2}{2\sigma^2} \right) \Bigg|_{(m=\hat{m}_{ML}, \sigma^2=\hat{\sigma}^2_{ML})} = 0,$$

$$\frac{\partial \ln L(\theta|y)}{\partial \theta_2} = \frac{\partial}{\partial \sigma^2} \left(n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(y_i - m)^2}{2\sigma^2_{\boldsymbol{y}}} \right) \Bigg|_{(m=\hat{m}_{ML}, \sigma^2=\hat{\sigma}^2_{ML})} = 0.$$

By differentiating with respect m and σ^2, one gets

$$\sum_{i=1}^{n} \frac{y_i - \hat{m}_{ML}}{\hat{\sigma}^2_{ML}} = 0$$

$$-\frac{n}{2\sigma^2_{ML}} + \frac{1}{2\sigma^4_{ML}} \sum_{i=1}^{n} (y_i - \hat{m}_{ML})^2 = 0,$$

from which

$$\hat{m}_{ML} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i$$

$$\sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{m}_{ML})^2.$$

Although $\mathbf{E}^{\theta}[\hat{m}_{ML}] = m_{\boldsymbol{y}}$ (see Example 1.1), one has $\mathbf{E}^{\theta}[\sigma^2_{ML}] = \frac{n-1}{n}\sigma^2_{\boldsymbol{y}}$ (see Example 1.2). Therefore, in this case, the Maximum Likelihood estimator is biased and hence it is not efficient. Due to Theorem 1.4, we can conclude that there does not exist any efficient estimator for the parameter $\theta = (m, \sigma^2)^T$. △

The previous example shows that Maximum Likelihood estimators can be biased. However, besides the motivations provided by Theorem 1.4, there exist other reasons that make such estimators attractive.

Theorem 1.5. *If the random variables $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are i.i.d., then (under suitable technical assumptions)*

$$\lim_{n \to +\infty} \sqrt{I_n(\theta)}\,(T_{ML}(\boldsymbol{y}) - \theta)$$

is a random variable with standard normal distribution $N(0,1)$.

In case of i.i.d. random variables, Theorem 1.5 states that the maximum likelihood estimator is:

- asymptotically unbiased $\quad \leftrightarrow \quad \mathbf{E}^\theta\left[T_{ML}(\boldsymbol{y})\right] = \theta$
- asymptotically efficient $\quad \leftrightarrow \quad \mathbf{E}^\theta\left[(T_{ML}(\boldsymbol{y}) - \theta)^2\right] = \dfrac{1}{I_n(\theta)}$
- consistent $\quad \leftrightarrow \quad \lim\limits_{n \to +\infty} \dfrac{1}{I_n(\theta)} = \lim\limits_{n \to +\infty} \dfrac{1}{n\, I_1(\theta)} = 0$
- asymptotically normal

1.4 Nonlinear estimation with additive noise

A popular class of estimation problems is the one in which the aim is to estimate a parameter θ, by using n measurements $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ corrupted by *additive noise*. Formally, let

$$h : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}^n$$

be a deterministic function of θ. The aim is to estimate θ by using the observations

$$\boldsymbol{y} = h(\theta) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ represents the measurement noise, modeled as a vector of random variables with pdf $f_{\boldsymbol{\varepsilon}}(\varepsilon)$.

Under this assumptions, the likelihood function is given by

$$L(\theta|y) = f^\theta_{\boldsymbol{y}}(y) = f_{\boldsymbol{\varepsilon}}(y - h(\theta))\,.$$

In the case in which the measurement noise $\boldsymbol{\varepsilon}$ is distributed according to the Gaussian pdf

$$f_{\boldsymbol{\varepsilon}}(\varepsilon) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_{\boldsymbol{\varepsilon}})^{1/2}} e^{-\frac{1}{2}\varepsilon^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \varepsilon}$$

with zero mean and known covariance matrix $\Sigma_{\boldsymbol{\varepsilon}}$, the log-likelihood function takes on the form

$$\ln L(\theta|y) = K - \frac{1}{2}(y - h(\theta))^T \Sigma_{\boldsymbol{\varepsilon}}^{-1}(y - h(\theta)),$$

where K is a constant that does not depend on θ. The computation of the maximum likelihood estimator boils down to the following optimization problem

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\max_\theta \ln L(\theta|y) \\ &= \arg\min_\theta (y - h(\theta))^T \Sigma_{\boldsymbol{\varepsilon}}^{-1}(y - h(\theta)). \end{aligned} \tag{1.13}$$

Being $h(\cdot)$, in general, a nonlinear function of θ, the solution of (1.13) can be computed by resorting to numerical methods. Clearly, the computational complexity depends not only on the number p of parameters to be estimated and on the size n of the data set, but also on the structure of $h(\cdot)$. For example, if $h(\cdot)$ is convex there are efficient algorithms that allow to solve problems with very large n and p, while if $h(\cdot)$ is nonconvex the problem may become intractable even for relatively small values of p.

1.5 Linear estimation problems

An interesting scenario is the one in which the relationship between the unknown parameters and the data is linear, i.e. $h(\theta) = U\,\theta$, where $U \in \mathbb{R}^{n\times p}$. In this case, the measurement equation takes on the form

$$\boldsymbol{y} = U\theta + \boldsymbol{\varepsilon}. \tag{1.14}$$

In the following, we will assume that $\text{rank}(U) = p$, which means that the number of linearly independent measurements is not smaller than the number of parameters to be estimated (otherwise, the problem is ill posed).

We now introduce two popular estimators that can be used to estimate θ in the setting (1.14). We will discuss their properties, depending on the assumptions we make on the measurement noise $\boldsymbol{\varepsilon}$. Let us start with the *Least Squares* estimator.

Definition 1.12. Let $\boldsymbol{y}$ be a vector of random variables related to θ according to (1.14). The estimator

$$T_{LS}(\boldsymbol{y}) = (U^TU)^{-1}U^T\boldsymbol{y} \tag{1.15}$$

is called *Least Squares (LS)* estimator of the parameter θ.

The name of this estimator comes from the fact that it minimizes the sum of the squared differences between the data realization y and the model $U\theta$, i.e.

$$\hat{\theta}_{LS} = \arg\min_{\theta} \|y - U\theta\|^2 .$$

Indeed,

$$\|y - U\theta\|^2 = (y - U\theta)^T(y - U\theta) = y^Ty + \theta^TU^TU\theta - 2y^TU\theta.$$

By differentiating with respect to θ, on gets

$$\frac{\partial}{\partial\theta}\|y - U\theta\|^2\Big|_{\theta=\hat{\theta}_{LS}} = 2\hat{\theta}_{LS}^TU^TU - 2y^TU = 0,$$

where the properties $\frac{\partial x^T A x}{\partial x} = 2x^T A$ and $\frac{\partial Ax}{\partial x} = A$ have been exploited. By solving with respect to $\hat{\theta}_{LS}^T$, one gets

$$\hat{\theta}_{LS}^T = y^T U (U^T U)^{-1}.$$

Finally, by transposing the above expression and taking into account that the matrix $(U^T U)$ is symmetric, one obtains the equation (1.15).

It is worth stressing that the LS estimator does not require any *a priori* information about the noise $\boldsymbol{\varepsilon}$ to be computed. As we will see in the sequel, however, the properties of $\boldsymbol{\varepsilon}$ will influence those of the LS estimator.

Definition 1.13. Let $\boldsymbol{y}$ be a vector of random variables related to θ according to (1.14). Let $\Sigma_{\boldsymbol{\varepsilon}}$ be the covariance matrix of $\boldsymbol{\varepsilon}$. The estimator:

$$T_{GM}(\boldsymbol{y}) = (U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1} U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{y} \tag{1.16}$$

is called *Gauss-Markov (GM)* estimator (or Weighted Least Squares Estimator) of the parameter θ.

Similarly to what has been shown for the LS estimator, it is easy to verify that the GM estimator minimizes the *weighted* sum of squared errors between y and $U\theta$, i.e.

$$\hat{\theta}_{GM} = \arg\min_{\theta} (y - U\theta)^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} (y - U\theta).$$

Notice that the Gauss-Markov estimator requires the knowledge of the covariance matrix $\Sigma_{\boldsymbol{\varepsilon}}$ of the measurement noise. By using this information, the measurements are weighted with a matricial weight that is inversely proportional to their uncertainty.

Under the assumption that the noise has zero mean, $\mathbf{E}[\boldsymbol{\varepsilon}] = 0$, it is easy to show that both the LS and the GM estimator are unbiased. For the LS estimator one has

$$\begin{aligned} \mathbf{E}^{\theta}\left[\hat{\theta}_{LS}\right] &= \mathbf{E}^{\theta}\left[(U^T U)^{-1} U^T \boldsymbol{y}\right] = \mathbf{E}^{\theta}\left[(U^T U)^{-1} U^T (U\theta + \boldsymbol{\varepsilon})\right] \\ &= \mathbf{E}^{\theta}\left[\theta + (U^T U)^{-1} U^T \boldsymbol{\varepsilon}\right] = \theta. \end{aligned}$$

For the GM estimator,

$$\begin{aligned} \mathbf{E}^{\theta}\left[\hat{\theta}_{GM}\right] &= \mathbf{E}^{\theta}\left[(U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1} U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{y}\right] = \mathbf{E}^{\theta}\left[(U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1} U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} (U\theta + \boldsymbol{\varepsilon})\right] \\ &= \mathbf{E}^{\theta}\left[\theta + (U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1} U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{\varepsilon}\right] = \theta. \end{aligned}$$

If the noise vector $\boldsymbol{\varepsilon}$ has non-zero mean, $m_{\boldsymbol{\varepsilon}} = \mathbf{E}[\boldsymbol{\varepsilon}]$, but the mean $m_{\boldsymbol{\varepsilon}}$ is known, the LS and GM estimators can be easily amended to remove the bias. In fact, if we define the new vector of random variables $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} - m_{\boldsymbol{\varepsilon}}$, the equation (1.14) can be rewritten as

$$\boldsymbol{y} - m_{\boldsymbol{\varepsilon}} = U\theta + \tilde{\boldsymbol{\varepsilon}}, \tag{1.17}$$

and being clearly $\mathbf{E}[\tilde{\boldsymbol{\varepsilon}}] = 0$, $\mathbf{E}\left[\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^T\right] = \Sigma_{\boldsymbol{\varepsilon}}$, all the treatment can be repeated by replacing $\boldsymbol{y}$ with $\boldsymbol{y} - m_{\boldsymbol{\varepsilon}}$. Therefore, the expressions of the LS and GM estimators remain those in (1.15) and (1.16), with $\boldsymbol{y}$ replaced by $\boldsymbol{y} - m_{\boldsymbol{\varepsilon}}$.

The case in which the mean of $\boldsymbol{\varepsilon}$ is unknown is more intriguing. In some cases, one may try to estimate it from the data, along with the parameter θ. Assume for example that $\mathbf{E}[\boldsymbol{\varepsilon}_i] = \bar{m}_\varepsilon$, $\forall i$. This means that $\mathbf{E}[\boldsymbol{\varepsilon}] = \bar{m}_\varepsilon \cdot \mathbb{1}$, where $\mathbb{1} = [1 \;\; 1 \;\; \dots \;\; 1]^T$. Now, one can define the extended parameter vector $\bar{\theta} = [\theta^T \;\; \bar{m}_\varepsilon]^T \in \mathbb{R}^{p+1}$, and use the same decomposition as in (1.17) to obtain

$$\boldsymbol{y} = [U \;\; \mathbb{1}]\bar{\theta} + \tilde{\boldsymbol{\varepsilon}}$$

Then, one can apply the LS or GM estimator, by replacing U with $[U \;\; \mathbb{1}]$, to obtain a simultaneous estimate of the p parameters θ and of the scalar mean $\bar{m}_\varepsilon$.

An important property of the Gauss-Markov estimator is that of being the minimum variance estimator among all linear unbiased estimators, i.e., the BLUE (see Definition 1.8). In fact, the following result holds.

Theorem 1.6. *Let $\boldsymbol{y}$ be a vector of random variables related to the parameter θ according to (1.14). Let $\Sigma_{\boldsymbol{\varepsilon}}$ be the covariance matrix of $\boldsymbol{\varepsilon}$. Then, the BLUE estimator of θ is the Gauss-Markov estimator* (1.16)*. The corresponding variance of the estimation error is given by*

$$\mathbf{E}\left[(\hat{\theta}_{GM} - \theta)(\hat{\theta}_{GM} - \theta)^T\right] = (U^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} U)^{-1}. \tag{1.18}$$

In the special case $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2 I_n$ (with I_n identity matrix of dimension n), i.e., when the variables $\boldsymbol{\varepsilon}$ are uncorrelated and have the same variance $\sigma_{\boldsymbol{\varepsilon}}^2$, the BLUE estimator is the Least Squares estimator (1.15)*.*

Proof

Since we consider the class of linear unbiased estimators, we have $T(\boldsymbol{y}) = A\boldsymbol{y}$, and $\mathbf{E}[A\boldsymbol{y}] = A\mathbf{E}[\boldsymbol{y}] = AU\theta$. Therefore, one must impose the constraint $AU = I_p$ to guarantee that the estimator is unbiased.

In order to find the minimum variance estimator, it is necessary to minimize (in matricial sense) the covariance of the estimation error

$$\begin{aligned}
\mathbf{E}\left[(A\boldsymbol{y} - \theta)(A\boldsymbol{y} - \theta)^T\right] &= \mathbf{E}\left[(AU\theta + A\boldsymbol{\varepsilon} - \theta)(AU\theta + A\boldsymbol{\varepsilon} - \theta)^T\right] \\
&= \mathbf{E}\left[A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T A^T\right] \\
&= A\Sigma_{\boldsymbol{\varepsilon}} A^T
\end{aligned}$$

where we have enforced the constraint $AU = I_p$ in the second equality. Then, the BLUE estimator is obtained by solving the constrained optimization problem

$$\begin{aligned}
A_{BLUE} \;=\; & \arg \min_A A\Sigma_{\boldsymbol{\varepsilon}} A^T \\
& \text{s.t.} \\
& AU = I_p
\end{aligned} \tag{1.19}$$

and then setting $T(\boldsymbol{y}) = A_{BLUE}\,\boldsymbol{y}$.
Being the constraint $AU = I_p$ linear in the matrix A, it is possible to parameterize all the admissible solutions A as

$$A = (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1} + M \tag{1.20}$$

with $M \in \mathbb{R}^{p\times n}$ such that $MU = 0$. It is easy to check that all matrices A defined by (1.20) satisfy the constraint $AU = I_p$. It is therefore sufficient to find the one that minimizes the quantity $A\Sigma_{\boldsymbol{\varepsilon}}A^T$. By substituting A with the expression (1.20), one gets

$$\begin{aligned} A\Sigma_{\boldsymbol{\varepsilon}}A^T &= (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\Sigma_{\boldsymbol{\varepsilon}}\Sigma_{\boldsymbol{\varepsilon}}^{-1}U(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1} \\ &\quad +(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\Sigma_{\boldsymbol{\varepsilon}}M^T \\ &\quad +M\Sigma_{\boldsymbol{\varepsilon}}\Sigma_{\boldsymbol{\varepsilon}}^{-1}U(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1} + M\Sigma_{\boldsymbol{\varepsilon}}M^T \\ &= (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1} + M\Sigma_{\boldsymbol{\varepsilon}}M^T \\ &\geq (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1} \end{aligned}$$

where the second equality is due to $MU = 0$, while the final inequality exploits the fact that $\Sigma_{\boldsymbol{\varepsilon}}$ is positive definite and hence $M\Sigma_{\boldsymbol{\varepsilon}}M^T$ is positive semidefinite. Since the expression $(U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}$ does not dipend on M, we can conclude that the solution of problem (1.19) is obtained by setting $M = 0$ in (1.20), which amounts to choosing $A_{BLUE} = (U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}U)^{-1}U^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}$. Therefore, the BLUE estimator coincides with the Gauss-Markov one. The expression of the covariance of the estimation error (1.18) is obtained from $A\Sigma_{\boldsymbol{\varepsilon}}A^T$ when $M = 0$.
Finally, if $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2 I_n$ one has $A_{BLUE} = (U^TU)^{-1}U^T$ (whatever is the value of $\sigma_{\boldsymbol{\varepsilon}}^2$) and hence the GM estimator boils down to the LS one. □

In Section 1.4 it has been observed that, if the measurement noise $\boldsymbol{\varepsilon}$ is Gaussian, the Maximum Likelihood estimator can be computed by solving the optimization problem (1.13). If the observations depend linearly on θ, as in (1.14), such a problem becomes

$$\hat{\theta}_{ML} = \arg\min_{\theta}(y - U\theta)^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}(y - U\theta). \tag{1.21}$$

As it has been noticed after Definition 1.13, the solution of (1.21) is actually the Gauss-Markov estimator. Therefore, we can state that: *in the case of linear observations corrupted by additive Gaussian noise, the Maximum Likelihood estimator coincides with the Gauss-Markov estimator.* Moreover, it is possible to show that in this setting

$$\mathbf{E}^{\theta}\left[\left(\frac{\partial \ln f_{\boldsymbol{y}}^{\theta}(y)}{\partial\theta}\right)\left(\frac{\partial \ln f_{\boldsymbol{y}}^{\theta}(y)}{\partial\theta}\right)^T\right] = U^T\Sigma^{-1}U$$

and hence the Gauss-Markov estimator is *efficient* (and UMVUE).

Finally, if the measurements are also independent and have the same variance $\sigma_{\boldsymbol{\varepsilon}}^2$, i.e., being the noise Gaussian, $\boldsymbol{\varepsilon} \sim N(0, \sigma_{\boldsymbol{\varepsilon}}^2 I_n)$, then, the GM estimator boils down to the LS one. Therefore: *in the case of linear observations, corrupted by independent and identically distributed Gaussian noise, the Maximum Likelihood estimator coincides with the Least Squares estimator.*

The following table summarizes the properties of the GM and LS estimators, depending on the assumptions made on the noise $\boldsymbol{\varepsilon}$.

Assumptions on ε	**Properties GM**	**Properties LS**
none	$\arg\min_\theta (y - U\theta)^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} (y - U\theta)$ with known $\Sigma_{\boldsymbol{\varepsilon}}$	$\arg\min_\theta \|y - U\theta\|^2$
$\mathbf{E}[\boldsymbol{\varepsilon}]$ known	unbiased	unbiased
$\mathbf{E}[\boldsymbol{\varepsilon}] = m_{\boldsymbol{\varepsilon}}$ $\mathbf{E}\left[(\boldsymbol{\varepsilon} - m_{\boldsymbol{\varepsilon}})(\boldsymbol{\varepsilon} - m_{\boldsymbol{\varepsilon}})^T\right] = \Sigma_{\boldsymbol{\varepsilon}}$	BLUE	BLUE if $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2 I_n$
$\boldsymbol{\varepsilon} \sim N(m_{\boldsymbol{\varepsilon}}, \Sigma_{\boldsymbol{\varepsilon}})$	ML estimator efficient, UMVUE	ML estimator if $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2 I_n$

Table 1.1: Properties of GM and LS estimators.

Example 1.9. On an unknown parameter θ, we collect n measurements

$$y_i = \theta + v_i \ , \qquad i = 1, \ldots, n$$

where the v_i are realizations of n random variables $\boldsymbol{v}_i$, independent, with zero mean and variance σ_i^2, $i = 1, \ldots, n$.
It is immediate to verify that the measurements y_i are realizations of random variables $\boldsymbol{y}_i$, with mean θ and variance σ_i^2. Therefore, the estimate of θ can be cast in terms of the estimate of the mean of n random variables (see Examples 1.1 and 1.5, and Exercise 1.1).

1.6 Exercises

1.1. Verify that in the problem of Example 1.9, the LS and GM estimators of θ coincide respectively with $\overline{\boldsymbol{y}}$ in (1.2) and $\hat{\boldsymbol{m}}_{BLUE}$ in (1.8).

1.2. Let $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ be two i.i.d. random variables, with pdf given by

$$f(\delta) = \begin{cases} \theta e^{-\theta\delta} & \text{if } \delta \geq 0 \\ 0 & \text{if } \delta < 0 \end{cases}$$

Let δ_1, δ_2 be the available observations of $\boldsymbol{d}_1$, $\boldsymbol{d}_2$. Find the Maximum Likelihood estimator of θ.

1.3. Let $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ be independent Gaussian random variables such that

$$\mathbf{E}\left[\boldsymbol{d}_1\right] = m, \quad \mathbf{E}\left[\boldsymbol{d}_2\right] = 3m, \quad \mathbf{E}\left[(\boldsymbol{d}_1 - m)^2\right] = 2, \quad \mathbf{E}\left[(\boldsymbol{d}_2 - 3m)^2\right] = 4$$

Let δ_1, δ_2 be the available observations of $\boldsymbol{d}_1$, $\boldsymbol{d}_2$. Find:

a) the minimum variance estimator of m among all linear unbiased estimators;

b) the variance of such an estimator;

c) the Maximum Likelihood estimator (is it different from the estimator in item a)?).

1.4. Two measurements are available on the unknown quantity x:

$$\boldsymbol{y}_1 = x + \boldsymbol{d}_1$$
$$\boldsymbol{y}_2 = 2x + \boldsymbol{d}_2$$

where $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ are independent disturbances modeled as random variables with pdf

$$f(\delta) = \begin{cases} \lambda e^{-\lambda\delta} & \text{if } \delta \geq 0 \\ 0 & \text{if } \delta < 0 \end{cases}$$

a) Find the Maximum Likelihood estimator of x.

b) Determine if the ML estimator is unbiased.

1.5. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two random variables with joint pdf

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \begin{cases} \dfrac{1}{2\theta^3}(3x+y) & 0 \leq x \leq \theta, \quad 0 \leq y \leq \theta \\ 0 & \text{elsewhere} \end{cases}$$

where θ is a real parameter.

a) Assume $\theta = 1$ and suppose that an observation y of the random variable $\boldsymbol{y}$ is available. Compute the minimum MSE estimator $\hat{\boldsymbol{x}}_{MSE}$ of $\boldsymbol{x}$, based on the observation y.

b) Assume θ is unknown and suppose that an observation y of the random variable $\boldsymbol{y}$ is available. Compute the ML estimator $\hat{\theta}_{ML}$ of the parameter θ, based on the measurement y. Establish if such an estimator is unbiased.

c) Assume θ is unknown and suppose that two observations x and y of the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are available. Compute the ML estimator $\hat{\theta}_{ML}$ of the parameter θ, based on the measurements x and y.

1.6. Let $\theta \in [-2, 2]$ and consider the function

$$f^{\theta}(x) = \begin{cases} \theta x + 1 - \dfrac{\theta}{2} & \text{if } x \in [0, 1] \\ 0 & \text{elsewhere} \end{cases}$$

a) Show that for all $\theta \in [-2, 2]$, f^{θ} is a probability density function.

b) Let $\boldsymbol{y}$ be a random variable with pdf f^{θ}. Compute mean and variance of $\boldsymbol{y}$ as functions of θ.

c) Compute the Maximum Likelihood estimator of θ, based on an observation y of the random variable $\boldsymbol{y}$.

d) Let $\boldsymbol{y}_1$, ..., $\boldsymbol{y}_n$ be n random variables, each one distributed according to the pdf f^{θ}, and consider the estimator

$$T(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = 12 \left(\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{y}_k - \frac{1}{2} \right)$$

Show that $T(\cdot)$ is an unbiased estimator of θ.

e) Find the variance of the estimation error for the estimator $T(\cdot)$ defined in item d), in the case $n = 1$. Compute the Fisher information $I_1(\theta)$ and show that the inequality (1.9) holds.

1.7. Let a and b be two unknown quantities, for which we have three different measurements:

$$\begin{aligned} \boldsymbol{y}_1 &= a + \boldsymbol{v}_1 \\ \boldsymbol{y}_2 &= b + \boldsymbol{v}_2 \\ \boldsymbol{y}_3 &= a + b + \boldsymbol{v}_3 \end{aligned}$$

where $\boldsymbol{v}_i$, $i = 1, 2, 3$, are independent random variables, with zero mean.
Let $\mathbf{E}[\boldsymbol{v}_1^2] = \mathbf{E}[\boldsymbol{v}_3^2] = 1$ and $\mathbf{E}[\boldsymbol{v}_2^2] = \frac{1}{2}$. Find:

a) The LS estimator of a and b;

b) The GM estimator of a and b;

c) The variance of the estimation error $\mathbf{E}\left[(a-\hat{a})^2+(b-\hat{b})^2\right]$, for the estimators computed in items a) and b).

Compare the obtained estimates with those one would have if the observation $\boldsymbol{y}_3$ were not available. How does the variance of the estimation error change?

1.8. Let X be an unknown quantity and assume the following measurement is available

$$\boldsymbol{y} = \ln\left(\frac{1}{X}\right) + \boldsymbol{v}$$

where $\boldsymbol{v}$ is a random variable, whose pdf is given by $f_{\boldsymbol{v}}(v) = \begin{cases} e^{-v} & v \geq 0 \\ 0 & v < 0 \end{cases}$.

a) Find the Maximum Likelihood estimator of X. Establish if it is biased or not. Is it possible to find an unbiased estimator of X?

b) Plot the estimate obtained in items a) as functions of y.