POLITECNICO DI MILANO

Introduction to the theory of estimation

Marco Lovera

Dipartimento di Scienze e Tecnologie Aerospaziali, Politecnico di Milano

The estimation problem

We have an estimation problem every time that

- we want to gather information about an unknown or uncertain parameter (scalar or vector)
- by means of experimental observations.

Parameter

The parameter, denoted as ϑ, can be

- either constant
- or time-varying: $\vartheta(t)$.

Data

Data points are represented as a function of an index *t* belonging to the set of observation instants:

$$d(t), \qquad t \in T = \{t_1, t_2, \ldots, t_N\}$$

so the entire data set is given by

$$d = \{d(t_1), d(t_2), \ldots, d(t_N)\}$$

Estimator

An *estimator* is a function *f*(·) which returns a value for the parameter to be estimated, as a function of data:

$$\widehat{\theta} = f(d).$$

$\hat{\theta}$ is called an *estimate* of θ.

Parametric vs non-parametric estimation

If θ is constant the estimation problem is called

- *parametric* if θ is finite-dimensional, *e.g.*,

$$\theta \in \mathbb{R}^{n_\theta}$$

- non-parametric if θ is infinite-dimensional, *e.g.*,

$$\theta = G(j\omega), \quad 0 \leq \omega < +\infty$$

Prediction, filtering, smoothing

If θ is time-varying, the estimation problem is called

- a *prediction* problem if given *d* we want to estimate θ(t) for $t > t_N$
- a *filtering* problem if given *d* we want to estimate θ(t) for $t = t_N$
- a *smoothing* problem if given *d* we want to estimate θ(t) for $t_1 < t < t_N$.

In all cases, we will denote the estimate as $\widehat{\theta}(t|T)$.

The data-generation mechanism

To formulate problems of estimator analysis and design we need assumptions on the *data-generation mechanism*, *i.e.*, the connection between θ and *d*.

Two viewpoints can be taken:

- deterministic viewpoint
- stochastic viewpoint.

Deterministic data-generation mechanism

- Assume you can change the value of θ and generate different data sets d_1, d_2, ... corresponding to $\theta=\theta_1$, $\theta=\theta_2$, ...

- Then if the data-generation mechanism is deterministic, *repeated experiments corresponding to the same value of θ will yield the same data.*

(think of using a scale to measure your weight)

Deterministic data-generation mechanism

- Is this sensible?

- Most of the time it is not entirely so, for a number of reasons:
  - noise affecting the measurement process
  - other factors besides the value of θ affecting the data-generation mechanism *(e.g.*, disturbances affecting a dynamic system).

- As will be discussed in the following, we refer
  - to the first effect as *measurement noise*
  - to the second effect as *process noise*.

Stochastic data-generation mechanism

The stochastic framework assumes that there is some *randomness* in the data-generation mechanism.

This framework for estimation aims at matching our experience of real processes and measurements, which in the best conditions are repeatable only up to a point.

Stochastic data-generation mechanism

In a stochastic data-generation mechanism:

- data is modelled as a random variable, the probability distribution of which depends on the value of θ.
- θ in turn can be seen in different ways depending on the specific formulation:
  - in Maximum Likelihood estimation θ is treated as an unknown (constant or time-varying) parameter;
  - in Bayesian estimation, on the other hand, θ is also treated as a random variable.

Recap on random variables and probability

Regardless of the specific approach to estimation, some background on random variables and probability is needed to formulate and solve problems.

Back to estimation theory

It is now possible to formalise the estimation problem using the developed theoretical background.

In the following it will be assumed that

- The data generation mechanism is a random experiment
- The random experiment depends on a parameter denoted θ^{o}
- Data is modelled as a random variable
- As a consequence, the estimate is also a random variable.

Back to estimation theory

More formally:

- we denote as d^N a data set composed of N samples
- we denote as f(·) the estimator and
- as $\widehat{\theta}_N = f(d^N)$ the estimate of θ^o computed on the basis of an N samples data set.

First objective: define desirable properties for a generic estimators which can be used to quantify its performance.

Back to estimation theory

Intuitively, as the estimate is a random variable which should approximate reliably a constant parameter we would like it to have:

- an expected value either equal to θ^o or which approaches it for increasing *N*.
- A small variance, possibly decreasing for increasing *N.*
- Is there such a thing as the smallest possible variance?

In the following we will address these questions.

Consistency

We say that an estimator is *consistent* if

$$\text{plim}_{N\to\infty}\widehat{\theta}_N = \theta^o$$

Consistency is a useful property but meaningful only for large samples.

For example, given a consistent estimator $\widehat{\theta}_N$ then for fixed *a* and *b* also

$$\frac{N-a}{N-b}\widehat{\theta}_N$$

is consistent, but behaves very differently for finite *N*.

Bias

The *bias* of an estimator is defined as

$$\text{bias} = E[\widehat{\theta}_N] - \theta^o$$

so it represents a measure of the *average* estimation error, for finite *N*.

Can we quantify for a given estimator whether it will have a bias?

Unbiasdness

We say that an estimator is *unbiased* if

$$E[\widehat{\theta}_N] = \theta^o$$

We say that it is asymptotically unbiased if

$$\lim_{N \to \infty} E[\widehat{\theta}_N] = \theta^o$$

Consistency vs unbiasedness

- Consistency does not imply unbiasdness for finite *N*.
- However, a consistent estimator having an asymptotic distribution with finite expected value will be unbiased.
- Unbiasedness, in turn, does not in general imply consistency.

Sufficient conditions for consistency

An estimator is consistent if it is asymptotically unbiased, *i.e.*,

$$\lim_{N\to\infty} E[\widehat{\theta}_N] = \theta^o$$

and

$$\lim_{N\to\infty} \mathrm{Var}[\widehat{\theta}_N] = 0$$

Example: bias analysis for the sample mean

Data x_i, *i*=1, ..., *N* drawn independently from $x \sim G(\mu, \sigma^2)$

We consider the sample mean

$$\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

as an estimator for μ. We get

$$E[\widehat{\mu}_N] = E[\frac{1}{N}\sum_{i=1}^{N} x_i] = \frac{1}{N}E[\sum_{i=1}^{N} x_i] =$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{N}N\mu = \mu$$

so the sample mean is an unbiased estimator.

Example: bias analysis for the sample variance

Data x_i, i=1, ..., N drawn independently from

We consider the sample variance

$$\hat{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_N)^2$$

as an estimator for σ^2. We get

$$E[\hat{\sigma}_N^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_N)^2 =$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j)^2 = \ldots = \frac{N-1}{N}\sigma^2$$

so in this case the estimator is biased.

Estimator variance

Within the class of unbiased estimators for a given problem, we anticipate that different estimators will lead to estimates with different variances.

We are interested in making the variance of the estimate as small as possible.

Questions:

- can we make the estimator variance *arbitrarily small* by picking a suitable estimator?
- If not, can we compute the smallest achievable variance?

Example: variance analysis for linear estimators

Consider the following problem:

- data d_i, i=1, ..., N is provided
- the d_is are independent but NOT identically distributed:

$$d_i \sim D(\theta^o, \mathsf{Var}[d_i]), \qquad i = 1, \ldots, N$$

Problem: design an estimator for θ^o.

Example: variance analysis for linear estimators

We compare three estimators, defined as follows:

1. Sample mean: $\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} d_i$

2. First sample: $\widehat{\theta}_N = d_1$

3. General linear estimator: $\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} \alpha_i^{(N)} d_i$

Example:
variance analysis for linear estimators

Bias analysis:

1. Sample mean: $E[\widehat{\theta}_N] = \frac{1}{N}\sum_{i=1}^{N} E[d_i] = \theta^o$

2. First sample: $E[\widehat{\theta}_N] = E[d_1] = \theta^o$

3. General linear estimator:

$$E[\widehat{\theta}_N] = \frac{1}{N}\sum_{i=1}^{N} \alpha_i^{(N)} E[d_i] = [\sum_{i=1}^{N} \alpha_i^{(N)}]\theta^o$$

So 1 and 2 are unbiased; 3 is unbiased if $\sum_{i=1}^{N} \alpha_i^{(N)} = 1$

Example:
variance analysis for linear estimators

Variance analysis:

1. Sample mean: $\mathrm{Var}[\widehat{\theta}_N] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[d_i] \underset{N\to\infty}{\longrightarrow} 0$

2. First sample: $\mathrm{Var}[\widehat{\theta}_N] = \mathrm{Var}[d_1]$

3. General linear estimator: $\mathrm{Var}[\widehat{\theta}_N] = \sum_{i=1}^{N}\alpha_i^{(N)2}\mathrm{Var}[d_i]$

   we now have to determine the set of α_is leading to the minimum variance subject to

$$\sum_{i=1}^{N}\alpha_i^{(N)} = 1$$

Example:
variance analysis for linear estimators

Consider the cost function

$$J = \mathsf{Var}[\widehat{\theta}_N] + \lambda(1 - [\sum_{i=1}^{N} \alpha_i^{(N)}])$$

where λ is a Lagrange multiplier.

We seek optimal α_is by imposing stationarity

$$\frac{\partial J}{\partial \alpha_i^{(N)}} = 0 \quad \Rightarrow \quad \alpha_i^{(N)} = \frac{\lambda}{2\mathsf{Var}[d_i]}$$

Example: variance analysis for linear estimators

Substituting in the constraint

$$\sum_{i=1}^{N} \alpha_i^{(N)} = 1$$

one gets

$$\lambda = \frac{2}{\sum_i \frac{1}{\mathrm{Var}[d_i]}}$$

and finally

$$\alpha_i^{(N)} = \frac{1}{\mathrm{Var}[d_i]}\alpha, \qquad \alpha = \frac{1}{\sum_i \frac{1}{\mathrm{Var}[d_i]}}$$

Example:
variance analysis for linear estimators

- We have constructed the MV linear estimator.
- Is it possible to find an estimator with smaller variance by enlarging the class of estimators?
- And if so, is it possible to reduce the variance arbitrarily in this way?

The Cramer-Rao bound

- The answer to the question is no: there is an intrinsic level of uncertainty in estimation problems which cannot be removed completely.

- The optimal performance of an estimator, in terms of variance, is quantified by the Cramer-Rao bound.

The Cramer-Rao bound

Consider the following problem:

- data d_i, i=1, ..., N is provided

- the d_is are independent, and $d_i \sim f(q_i|\theta^o), \quad i = 1, \dots, N$

so that the joint density for $d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}$

is given by

$$f(q_1, q_2, \dots, q_N|\theta^o) = f_1(q_1|\theta^o) f_2(q_2|\theta^o) \dots f_N(q_N|\theta^o)$$

Problem: design an estimator for θ^o.

The Cramer-Rao bound

Theorem (Cramer-Rao inequality):

For any unbiased estimator $\widehat{\theta}_N$ of θ we have that

$$Var[\widehat{\theta}_N] \geq -\frac{1}{E[\frac{\partial^2 \log f}{\partial \theta^2}]}.$$

The Cramer-Rao bound

Proof.

By definition of pdf we have that

$$\int\int\ldots\int f(q_1,q_2,\ldots,q_N)dq_1dq_2\ldots dq_N = 1$$

So differentiating with respect to θ we have

$$\int\int\ldots\int\frac{\partial f(q_1,q_2,\ldots,q_N)}{\partial\theta}dq_1dq_2\ldots dq_N = 0.$$

which is equivalent to

$$\int\int\ldots\int\frac{1}{f}\frac{\partial f}{\partial\theta}f dq_1dq_2\ldots dq_N = 0.$$

$$\int\int\ldots\int\frac{\partial\log f}{\partial\theta}f dq_1dq_2\ldots dq_N = 0 \quad\Rightarrow\quad E[\frac{\partial\log f}{\partial\theta}] = 0.$$

The Cramer-Rao bound

We now use the unbiasedness assumption:

$$E[\widehat{\theta}_N] = \int\int \ldots \int \widehat{\theta}_N f dq_1 dq_2 \ldots dq_N = \theta$$

and again differentiating we get

$$\int\int \ldots \int \widehat{\theta}_N \frac{\partial f}{\partial \theta} dq_1 dq_2 \ldots dq_N = 1$$

$$\int\int \ldots \int \widehat{\theta}_N \frac{1}{f}\frac{\partial f}{\partial \theta} f dq_1 dq_2 \ldots dq_N = 1$$

$$\int\int \ldots \int \widehat{\theta}_N \frac{\partial \log f}{\partial \theta} f dq_1 dq_2 \ldots dq_N = 1$$

The Cramer-Rao bound

Now recall that

$$\int\int\dots\int\frac{\partial\log f}{\partial\theta}f\,dq_1dq_2\dots dq_N=0\qquad\Rightarrow\qquad E[\frac{\partial\log f}{\partial\theta}]=0.$$

and use the fact that θ is constant to write:

$$\int\int\dots\int\theta\frac{\partial\log f}{\partial\theta}f\,dq_1dq_2\dots dq_N=0.$$

We have also proved that

$$\int\int\dots\int\widehat{\theta}_N\frac{\partial\log f}{\partial\theta}f\,dq_1dq_2\dots dq_N=1$$

So by subtraction we have

$$\int\int\dots\int(\widehat{\theta}_N-\theta)\frac{\partial\log f}{\partial\theta}f\,dq_1dq_2\dots dq_N=1.$$

Cauchy-Schwartz inequality

For any pair of functions g(x) and h(x) for which the involved integrals exist, we have that

$$\int_{-\infty}^{+\infty} g(x)h(x)dx \leq \int_{-\infty}^{+\infty} g^2(x)dx \int_{-\infty}^{+\infty} h^2(x)dx.$$

The Cramer-Rao bound

In our case letting

$$g(x) = (\hat{\theta}_N - \theta)\sqrt{f} \quad h(x) = \frac{\partial \log f}{\partial \theta}\sqrt{f}$$

we get

$$1 = \int\int \dots \int (\hat{\theta}_N - \theta)\frac{\partial \log f}{\partial \theta} f dq_1 dq_2 \dots dq_N$$

$$\leq \int\int \dots \int (\hat{\theta}_N - \theta)^2 f dq_1 dq_2 \dots dq_N \int\int \dots \int (\frac{\partial \log f}{\partial \theta})^2 f dq_1 dq_2 \dots dq_N.$$

$$Var[\hat{\theta}_N] \qquad\qquad E[(\frac{\partial \log f}{\partial \theta})^2]$$

So finally we get $Var[\hat{\theta}_N] \geq \frac{1}{E[(\frac{\partial \log f}{\partial \theta})^2]}.$

The Cramer-Rao bound

As a final step note that

$$E[\frac{\partial \log f}{\partial \theta}] = \int\int \ldots \int \frac{\partial \log f}{\partial \theta} f dq_1 dq_2 \ldots dq_N = 0$$

and differentiating

$$\int\int \ldots \int [(\frac{1}{f}\frac{\partial f}{\partial \theta})^2 + \frac{\partial^2 \log f}{\partial \theta^2}] f dq_1 dq_2 \ldots dq_N = 0$$

therefore

$$E[\frac{\partial^2 \log f}{\partial \theta^2}] = -E[(\frac{\partial \log f}{\partial \theta})^2]$$

So finally we get $Var[\widehat{\theta}_N] \geq -\frac{1}{E[\frac{\partial^2 \log f}{\partial \theta^2}]}.$

The Cramer-Rao bound: example

Consider the following problem:

- data d_i, i=1, ..., N is provided
- the d_is are independent, and $d_i \sim G(\theta, \sigma^2), \quad i = 1, \ldots, N$

so that the joint density for $d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}$

is given by

$$f(q_1, q_2, \ldots, q_N|\theta^o) = f_1(q_1|\theta^o) f_2(q_2|\theta^o) \ldots f_N(q_N|\theta^o)$$

Problem: compute the C-R bound for estimates of θ.

The Cramer-Rao bound: example

The joint density can be written as

$$f(q_1, q_2, \ldots, q_N|\theta) = f_1(q_1|\theta) f_2(q_2|\theta) \ldots f_N(q_N|\theta)$$

$$f(q_1, q_2, \ldots, q_N|\theta) = \frac{1}{(\sigma\sqrt{2\pi})^N} e^{-\frac{(q_1-\theta)^2}{2\sigma^2}} \ldots e^{-\frac{(q_N-\theta)^2}{2\sigma^2}}$$

$$f(q_1, q_2, \ldots, q_N|\theta) = \frac{1}{(\sigma\sqrt{2\pi})^N} e^{-\frac{\sum_{i=1}^N (q_i-\theta)^2}{2\sigma^2}}$$

so the logarithm of the joint density is

$$\log f(q_1, q_2, \ldots, q_N|\theta) = -(\sigma\sqrt{2\pi})^N - \frac{\sum_{i=1}^N (q_i - \theta)^2}{2\sigma^2}$$

The Cramer-Rao bound: example

Taking the first and second derivatives we get

$$\frac{\partial \log f(q_1, q_2, \ldots, q_N|\theta)}{\partial\theta} = \frac{\sum_{i=1}^{N}(q_i - \theta)}{\sigma^2}$$

$$\frac{\partial^2 \log f(q_1, q_2, \ldots, q_N|\theta)}{\partial\theta^2} = -\frac{N}{\sigma^2}$$

and therefore

$$Var[\widehat{\theta}_N] \geq -\frac{1}{E[\frac{\partial^2 \log f}{\partial\theta^2}]} = \frac{\sigma^2}{N}.$$

Note that $\frac{\sigma^2}{N} \to 0 \quad \text{for} \quad N \to \infty$ but is always non-zero for finite N.

When is the minimum variance attained?

- Theorem: a necessary and sufficient condition for the minimum variance to be attained is that there exists a function $A(\theta)$ such that

$$\frac{\partial \log f}{\partial \theta} = A(\theta)(\widehat{\theta}_N - \theta)$$

for all sets of observations.

Example

Consider the following problem:

- data d_i, i=1, ..., N is provided

- the d_is are independent, and $d_i \sim G(\theta, \sigma^2), \quad i = 1, \dots, N$

so that the joint density for $d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}$

is given by

$$f(q_1, q_2, \dots, q_N|\theta^o) = f_1(q_1|\theta^o) f_2(q_2|\theta^o) \dots f_N(q_N|\theta^o)$$

Problem: estimate θ using $\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} d_i.$

Example

Computing the derivative of log f we get

$$\frac{\partial \log f(q_1, q_2, \ldots, q_N|\theta)}{\partial\theta} = \frac{\sum_{i=1}^{N}(q_i - \theta)}{\sigma^2}$$

and plugging the observations in place of the running variables:

$$\frac{\partial \log f(d_1, d_2, \ldots, d_N|\theta)}{\partial\theta} = \frac{\sum_{i=1}^{N}(d_i - \theta)}{\sigma^2}$$

which can be written as

$$\frac{\partial \log f(d_1, d_2, \ldots, d_N|\theta)}{\partial\theta} = \frac{\sum_{i=1}^{N} d_i}{\sigma^2} - \frac{N}{\sigma^2}\theta = \frac{N}{\sigma^2}\widehat{\theta}_N - \frac{N}{\sigma^2}\theta = \frac{N}{\sigma^2}(\widehat{\theta}_N - \theta)$$

The Cramer-Rao bound for vectors of parameters

In most problems of practical interest we have to estimate a vector of parameters:

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{n_\theta} \end{bmatrix}.$$

The Cramer-Rao bounds generalises to

$$\begin{aligned} Var[\widehat{\theta}_N] &= E[(\widehat{\theta}_N - \theta)(\widehat{\theta}_N - \theta)^T] \\ &\geq -E[\frac{\partial^2 \log f}{\partial\theta\partial\theta^T}]^{-1} = E[(\frac{\partial \log f}{\partial\theta})(\frac{\partial \log f}{\partial\theta})^T]^{-1}. \end{aligned}$$

The Fischer information matrix

The quantity

$$M = -E[\frac{\partial^2 \log f}{\partial \theta \partial \theta^T}] = E[(\frac{\partial \log f}{\partial \theta})(\frac{\partial \log f}{\partial \theta})^T]$$

is known as the Fischer information matrix.

In terms of M the bound can be written as

$$Var[\widehat{\theta}_N] \geq M^{-1}.$$

Clearly, letting G the inverse of M we have that

$$Var[\widehat{\theta}_{N,i}] \geq \frac{1}{g_{ii}}.$$

Efficiency

An estimator is called efficient if it is unbiased and it reaches the Cramer-Rao bound for all values of the parameter.

An estimator is called asymptotically efficient if it is unbiased and it reaches the Cramer-Rao bound for all values of the parameter, at least when N goes to infinity.

Variance, bias and mean-square error

Consider a generic estimator $\widehat{\theta}_N$ not necessarily unbiased:

$$E[\widehat{\theta}_N] \neq \theta.$$

Then the mean square error

$$E[(\widehat{\theta}_N - \theta)^2]$$

will be different from the variance

$$E[(\widehat{\theta}_N - E[\widehat{\theta}_N])^2].$$

In particular, we have that

$$E[(\widehat{\theta}_N - \theta)^2] = E[(\widehat{\theta}_N - E[\widehat{\theta}_N])^2] + (E[\widehat{\theta}_N] - \theta)^2$$

Variance, bias and mean-square error

In other words, the MSE equals the sum of variance and squared bias.

For unbiased estimators, MSE and variance coincide.

In some applications, (biased) minimum MSE estimators are used, rather than minimum variance ones.