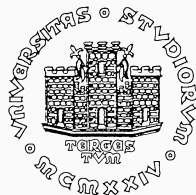


Systems Dynamics

Course ID: 267MI – Fall 2018

Thomas Parisini
Gianfranco Fenu

University of Trieste
Department of Engineering and Architecture



267MI –Fall 2018

Lecture 6

**Definitions and properties of
the estimation and prediction
problems**

The estimation problem

The estimation problem

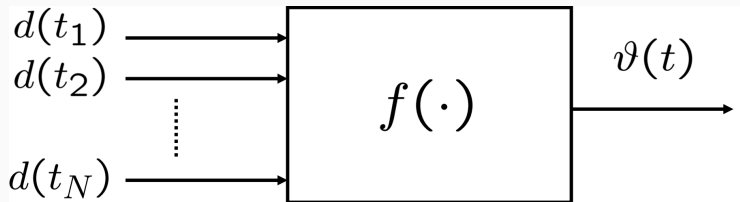
- The estimation problem arises when there is a need of determining one or more unknown quantities **using experimentally observed data**



- In most cases **the unknown parameters are constant**

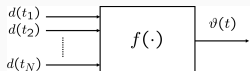
$$\vartheta(t) \equiv \vartheta$$

- $T = \{t_1, t_2, \dots, t_N\}$ set of the observation time-instants
 - In general, there is no need of equally-spaced t_i
 - If there is the possibility of choosing the instants t_i when to get experimental data, it is convenient to have more observations where the experiment is more significant.



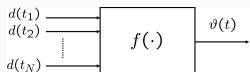
The estimator is a **deterministic function** yielding as output the unknown parameters on the basis of the observed data as inputs

Estimation of constant parameters

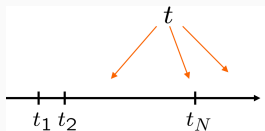


- If $\vartheta(t) \equiv \bar{\vartheta} = \text{const}$ we have a **parametric estimation or identification problem**.
- The estimate given by the estimator is denoted as $\hat{\vartheta}$ or $\hat{\vartheta}_T$ to enhance the set of observation time-instants.
- The “true” value of the parameter is denoted as ϑ° .

Estimation of time-varying parameters



- The estimate generated by the estimator is denoted as $\hat{\vartheta}(t|T)$ or simply as $\hat{\vartheta}(t|N)$ if we can set $T = \{1, 2, \dots, N\}$.
- Typically we have three cases:
 - $t > t_N$: problem of **prediction**
 - $t = t_N$: problem of **filtering**
 - $t < t_N$: problem of **smoothing**



The prediction problem

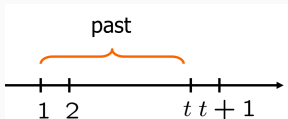
It is a fundamental problem in the context of **dynamical systems identification**

- To set the basics, let us focus on the case of *time-series*
- A sequence of observations $y(1), y(2), \dots, y(t)$ of a variable $y(\cdot)$ is available.
- We want to estimate $y(t + 1)$
- Therefore, we want to design a **predictor**

$$\hat{y}(t + 1 | t) = f [y(t), y(t - 1), \dots, y(1)]$$

The prediction problem (cont.)

- The predictor expresses an estimate $\hat{y}(t+1|t)$ of $y(t+1)$ as a function of t **past** values of $y(\cdot)$



- A predictor is **linear** if

$$\hat{y}(t+1|t) = a_1(t) \cdot y(t) + \dots + a_t(t) \cdot y(1)$$

- A predictor is **finite-memory** (hence uses a limited memory of the past) if

$$\hat{y}(t+1|t) = a_1(t) \cdot y(t) + \dots + a_n(t) \cdot y(t-n+1)$$

The prediction problem (cont.)

- A predictor is linear time-invariant if

$$\hat{y}(t+1|t) = a_1 y(t) + \dots + a_n y(t-n+1)$$

where the parameters a_1, \dots, a_n are **constant**

- We define the vector of parameters $\vartheta^T = [a_1, \dots, a_n]$

Determining a “good” predictor means determining a suitable vector ϑ such that the prediction $\hat{y}(t+1|t)$ is the more accurate possible

The prediction problem (cont.)

More precisely:

- Consider a **finite-memory linear time-invariant** predictor

$$\hat{y}(t+1|t) = a_1 y(t) + \dots + a_n y(t-n+1)$$

where n is “small” with respect to the number of data observed till time-instant t

- The performances of the predictor can be evaluated on the already-available data: $y(i)$ $i = 1, \dots, t$
 - we compute

$$\hat{y}(i+1|i) = a_1 y(i) + \dots + a_n y(i-n+1), \quad \forall i > n$$

- We evaluate the **prediction error**

$$\varepsilon(i+1) = y(i+1) - \hat{y}(i+1|i), \quad \forall i > n$$

The prediction problem (cont.)

The vector $\vartheta^T = [a_1, \dots, a_n]$ is “good” if ε is “small” over the available data.

- Introduce the criterion:

$$J(\vartheta) = \sum_{i=n+1}^t (\varepsilon(i))^2$$

- Hence

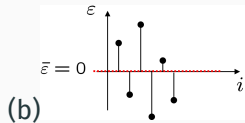
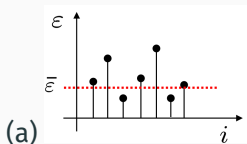
$$\vartheta^\circ = \arg \min_{\vartheta} J(\vartheta)$$

The determination of ϑ° is thus reduced to the solution of an **optimization problem**.

Remarks

It is very important to clarify the meaning of ε “small”

The minimization of $J(\vartheta)$ is not *per se* a fully satisfactory criterion



- CASE (A): not satisfactory because the average error $\bar{\varepsilon}$ is not zero \Rightarrow **systematic error**
- CASE (B): despite the fact that the average error $\bar{\varepsilon}$ is zero, it is not satisfactory because the sequence is alternatively positive and negative; hence, at any time-instant the sign of the next error is known in advance \Rightarrow **The predictor does not embed all the information**

The ideal situation

Prediction error ε with smallest possible average and “as much as **unpredictable** as possible”

$$\varepsilon(\cdot) \sim \text{WN} (0, \lambda^2)$$

The diagram shows the equation $\varepsilon(\cdot) \sim \text{WN} (0, \lambda^2)$ with three red arrows pointing from text labels to parts of the equation: 'white noise' points to 'WN', 'average' points to '0', and 'variance' points to ' λ^2 '.

Predictor as a dynamic system

$$\hat{y}(t|t-1) = a_1 y(t-1) + \dots + a_n y(t-n)$$

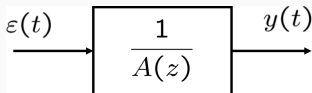
$$\varepsilon(t) = y(t) - \hat{y}(t|t-1) \quad \Rightarrow \quad y(t) = \varepsilon(t) + \hat{y}(t|t-1)$$

$$y(t) = a_1 y(t-1) + \dots + a_n y(t-n) + \varepsilon(t)$$

$$y(t) = (a_1 z^{-1} + \dots + a_n z^{-n}) y(t) + \varepsilon(t)$$

$$A(z)y(t) = \varepsilon(t) \text{ with } A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_n z^{-n}$$

$$y(t) = \frac{1}{A(z)} \varepsilon(t)$$



**A Glimpse on
Estimation theory
&
Estimators' characteristics**

General concepts and definitions

- In general we have:

$$d = d(s, \vartheta^\circ)$$

where

- $d \iff$ observed (measured) data
 - $\vartheta^\circ \iff$ unknown quantity to be estimated
 - $s \iff$ result of the random experiment
- The estimator is a function:

$$\hat{\vartheta} = f[d(s, \vartheta^\circ)]$$

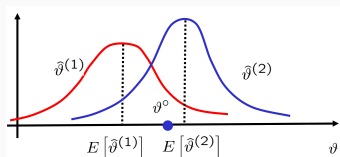
The estimator is a random variable because its value depends on the result s of the random experiment

- In general, the estimator $\hat{\vartheta} = f[d(s, \vartheta^\circ)]$ is **unbiased** if

$$E(\hat{\vartheta}) = \vartheta^\circ$$

- Clearly, it is important to try to ensure that the estimator is unbiased.

In this example, the estimators are both biased but the estimator $\hat{\vartheta}^{(2)}$ is characterized by a lower bias



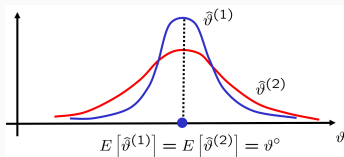
Minimum variance

- The “unbiasedness” (correctness) is not the only criterion to be used to evaluate the quality of an estimator.

In this case, both estimators are unbiased.

However:

$$\text{var} [\hat{\vartheta}^{(1)}] \ll \text{var} [\hat{\vartheta}^{(2)}]$$



- Hence, the estimator $\hat{\vartheta}^{(1)}$ has a higher probability of yielding estimates closer to the true value ϑ° as compared with the estimator $\hat{\vartheta}^{(2)}$
- Therefore, the goal is to reduce the variance of the estimator as much as possible.

Minimum variance (cont.)

- In general, under the same bias characteristics, we say that the estimator $\hat{\vartheta}^{(1)}$ is better than the estimator $\hat{\vartheta}^{(2)}$ if

$$\text{var} \left[\hat{\vartheta}^{(1)} \right] \leq \text{var} \left[\hat{\vartheta}^{(2)} \right]$$

that is, if the matrix (ϑ may be a vector)

$$\text{var} \left[\hat{\vartheta}^{(2)} \right] - \text{var} \left[\hat{\vartheta}^{(1)} \right] \geq 0$$

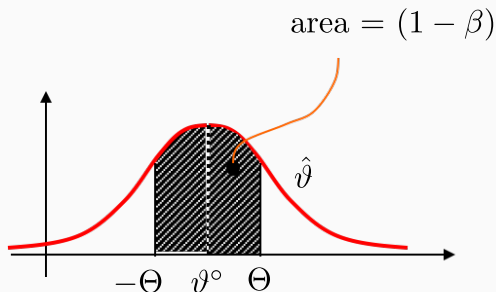
- Recalling that $A \geq 0 \implies \det A \geq 0, \lambda_i \geq 0, a_{ii} \geq 0$, we have

$$\text{var} \left[\hat{\vartheta}^{(2)} \right] - \text{var} \left[\hat{\vartheta}^{(1)} \right] \geq 0 \implies \text{var} \left[\hat{\vartheta}_i^{(2)} \right] \geq \text{var} \left[\hat{\vartheta}_i^{(1)} \right]$$

where $\hat{\vartheta}_i^{(1)}, \hat{\vartheta}_i^{(2)}$ denote the i -th components of the vectors $\hat{\vartheta}^{(1)}, \hat{\vartheta}^{(2)}$.

Estimate's confidence

Consider an estimator $\hat{\vartheta}$:

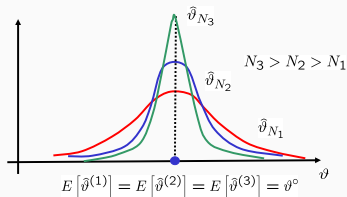


The estimate $\hat{\vartheta}$ belongs to the interval $(-\Theta, \Theta)$ around ϑ° with confidence $(1 - \beta) \cdot 100\%$.

Asymptotic characteristics

- If the number N of available data increases over time
 - the available information to compute the estimate increases
 - the uncertainty decreases
- From this perspective the estimator $\hat{\vartheta}_N$ is “good” if

$$\lim_{N \rightarrow \infty} \text{var} [\hat{\vartheta}_N] = 0$$



Convergence in “quadratic mean”

- When the estimate $\hat{\vartheta}_N$ is computed on the basis of a time-increasing amount of data N , another estimate’s quality criterion is

$$\lim_{N \rightarrow \infty} \text{E} \left[\left\| \hat{\vartheta}_N - \vartheta^\circ \right\|^2 \right] = 0 \quad (*)$$

If (*) holds we say that the estimate $\hat{\vartheta}_N$ **converges to ϑ° in “quadratic mean”**

- Notice that $\hat{\vartheta}_N$ is a random vector, ϑ° is a constant vector and $\left\| \hat{\vartheta}_N - \vartheta^\circ \right\|^2$ is a scalar random variable with a well-defined expected value.

Almost-sure convergence

- Recall that the estimator based on N data is

$$\hat{\vartheta}_N(s, \vartheta^\circ) = f[d(s, \vartheta^\circ)]$$

- For a given $\bar{s} \in S$, we have a sequence

$$\hat{\vartheta}_1(\bar{s}, \vartheta^\circ), \hat{\vartheta}_2(\bar{s}, \vartheta^\circ), \dots, \hat{\vartheta}_N(\bar{s}, \vartheta^\circ), \dots$$

- It may happen that:

$$\bar{s} \in S \longrightarrow \lim_{N \rightarrow \infty} \hat{\vartheta}_N(\bar{s}, \vartheta^\circ) = \vartheta^\circ$$

$$\tilde{s} \in S \longrightarrow \lim_{N \rightarrow \infty} \hat{\vartheta}_N(\tilde{s}, \vartheta^\circ) \neq \vartheta^\circ$$

Almost-sure convergence (cont.)

- Introduce the set of random experiment results

$$A \subset S, A = \left\{ s \in S : \lim_{N \rightarrow \infty} \hat{\vartheta}_N(s, \vartheta^\circ) = \vartheta^\circ \right\}$$

- If $A = S$  **Sure convergence**
- If $A \subset S$ and $P(A) = 1$  **Almost-sure convergence**

Note that, if the measure of the set $S \setminus A$ is zero, this implies $P(A) = 1$ and hence *almost-sure convergence*.

- Clearly $A = S \implies P(A) = 1$

Sure convergence  **Almost-sure convergence**

- An estimator characterized by almost-sure convergence properties is called **consistent**.

Example 1

- Consider N scalar data $d(1), d(2), \dots, d(N)$ such that

$$\mathbb{E}[d(i)] = \vartheta^\circ, \quad i = 1, 2, \dots, N$$

- Assume that data are mutually un-correlated, that is

$$\mathbb{E}\{[d(i) - \vartheta^\circ][d(j) - \vartheta^\circ]\} = 0, \quad \forall i \neq j$$

- Consider the estimator

$$\hat{\vartheta}_N = \frac{1}{N} \sum_{i=1}^N d(i)$$

Sampled-average estimator

Example 1 (cont.)

- Bias:

$$E \left[\hat{\vartheta}_N \right] = E \left\{ \frac{1}{N} \sum_{i=1}^N [d(i)] \right\} = \frac{1}{N} \sum_{i=1}^N E [d(i)] = \frac{1}{N} \sum_{i=1}^N \vartheta^\circ = \vartheta^\circ$$

the estimator is unbiased

- Variance:

$$\begin{aligned} \text{var} \left(\hat{\vartheta}_N \right) &= E \left\{ \left[\hat{\vartheta}_N - E \left(\hat{\vartheta}_N \right) \right]^2 \right\} = E \left\{ \left[\frac{1}{N} \sum_{i=1}^N d(i) - \frac{1}{N} \sum_{i=1}^N \vartheta^\circ \right]^2 \right\} \\ &= E \left\{ \frac{1}{N^2} \left[\sum_{i=1}^N d(i) - \sum_{i=1}^N \vartheta^\circ \right]^2 \right\} = \frac{1}{N^2} \sum_{i=1}^N E \left\{ [d(i) - \vartheta^\circ]^2 \right\} \\ &= \frac{1}{N^2} \sum_{i=1}^N \text{var} [d(i)] \end{aligned}$$

the "cross-terms" are zero because of the assumption on un-correlated data

Example 1 (cont.)

- If $\text{var}[d(i)] \leq \bar{\sigma}$, $i = 1, 2, \dots, N$

$$\lim_{N \rightarrow \infty} \text{var}(\hat{\vartheta}_N) \leq \lim_{N \rightarrow \infty} \frac{\bar{\sigma}}{N} = 0$$

the estimator converges in quadratic mean

Example 2

- Consider N scalar data $d(1), d(2), \dots, d(N)$ such that

$$E[d(i)] = \vartheta^\circ, \quad i = 1, 2, \dots, N$$

- Assume that the data are mutually un-correlated, that is

$$E\{[d(i) - \vartheta^\circ][d(j) - \vartheta^\circ]\} = 0, \quad \forall i \neq j$$

- Consider the estimator

$$\hat{\vartheta}_N = \sum_{i=1}^N \alpha(i) d(i)$$

Example 2 (cont.)

- Bias:

$$\mathbb{E} \left[\hat{\vartheta}_N \right] = \mathbb{E} \left\{ \sum_{i=1}^N \alpha(i) d(i) \right\} = \sum_{i=1}^N \alpha(i) \mathbb{E} [d(i)] = \vartheta^\circ \sum_{i=1}^N \alpha(i)$$

The estimator is unbiased $\longleftrightarrow \sum_{i=1}^N \alpha(i) = 1 \quad (\star)$

N.B. in the previous case $\alpha(i) = \frac{1}{N}$ and hence (\star) holds

Condition (\star) is a constraint to be satisfied so that the estimator is unbiased.

This constraint characterizes a class of unbiased estimators

Example 2 (cont.)

- Let us now determine the best estimator among the unbiased ones (hence satisfying the constraint (\star)) choosing the **minimum variance** one

$$\begin{cases} \min \text{var} \left(\hat{\vartheta}_N \right) = \min \sum_{i=1}^N [\alpha(i)]^2 \text{var} [d(i)] \\ 1 - \sum_{i=1}^N \alpha(i) = 0 \end{cases}$$

un-correlated data

By using the Lagrange multipliers technique we have:

$$J \left(\hat{\vartheta} \right) = \sum_{i=1}^N [\alpha(i)]^2 \cdot \text{var} [d(i)] + \lambda \left(1 - \sum_{i=1}^N \alpha(i) \right)$$

Example 2 (cont.)

$$\frac{\partial J}{\partial \alpha(i)} = 0 \iff 2\alpha(i) \text{var}[d(i)] - \lambda = 0 \iff \alpha(i) = \frac{\lambda}{2 \text{var}[d(i)]}$$

- Now, imposing the constraint (★) for unbiasedness

$$\sum_{i=1}^N \alpha(i) = 1 \iff \frac{\lambda}{2} \sum_{i=1}^N \frac{1}{\text{var}[d(i)]} = 1 \iff \lambda = \frac{2}{\sum_{i=1}^N \frac{1}{\text{var}[d(i)]}}$$

$$\alpha(i) = \frac{1}{\text{var}[d(i)]} \alpha \quad \text{with} \quad \alpha = \frac{1}{\sum_{i=1}^N \frac{1}{\text{var}[d(i)]}}$$

Hence, $\alpha(i)$ is chosen to be inversely proportional to the data variance $\text{var}[d(i)]$: the bigger the data variance, the smaller the associated weight (consistent with intuition).

Example 2 (cont.)

- Let us compute the estimator's variance:

$$\begin{aligned}\text{var}(\hat{\vartheta}_N) &= \text{E} \left\{ \left[\hat{\vartheta}_N - \text{E}(\hat{\vartheta}_N) \right]^2 \right\} = \text{E} \left\{ \left[\sum_{i=1}^N \alpha(i)d(i) - \vartheta^\circ \sum_{i=1}^N \alpha(i) \right]^2 \right\} \\ &= \text{E} \left\{ \left[\sum_{i=1}^N \alpha(i) [d(i) - \vartheta^\circ] \right]^2 \right\} = \sum_{i=1}^N [\alpha(i)]^2 \text{E} \left\{ [d(i) - \vartheta^\circ]^2 \right\} \\ &= \sum_{i=1}^N (\alpha(i))^2 \text{var} [d(i)] = \alpha^2 \sum_{i=1}^N \frac{1}{\text{var} [d(i)]} = \frac{1}{\sum_{i=1}^N \frac{1}{\text{var} [d(i)]}}\end{aligned}$$

Example 2 (cont.)

- If $\text{var}[d(i)] \leq \bar{\sigma}$, $i = 1, 2, \dots, N$

$$\lim_{N \rightarrow \infty} \text{var}(\hat{\vartheta}_N) \leq \lim_{N \rightarrow \infty} \frac{\bar{\sigma}}{N} = 0$$

the estimator converges in quadratic mean

Generalization

- When the quantities to be estimated are **time-varying**, it is necessary to modify the estimators' quality indexes.
- Denote with $\hat{\vartheta}(t|t-1)$ the estimate of $\vartheta^\circ(t)$ exploiting data collected till time-instant $t-1$
- Clearly, as $\vartheta^\circ(t)$ varies over time, it does not make sense to talk about asymptotic convergence in terms of data in the past that may turn up not to be meaningful any more.
- A typical criterion is

$$\mathbb{E} \left[\left\| \hat{\vartheta}(t|t-1) - \vartheta^\circ(t) \right\|^2 \right] \leq c$$

where c is a suitably small positive scalar

- **In this time-varying case what matters is not “convergence” but “boundedness”**

267MI –Fall 2018

Lecture 6

**Definitions and properties of the
estimation and prediction prob-
lems**

END