Systems Dynamics

Course ID: 267MI – Fall 2019

Thomas Parisini
Gianfranco Fenu

University of Trieste
Department of Engineering and Architecture

267MI –Fall 2019

Lecture 6
Definitions and Properties of the Estimation and Prediction Problems

Lecture 6: Table of Contents

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

Experimental observations
$d(t)\ , \quad t = t_1\ , \ t_2\ , \ \dots t_N$

→

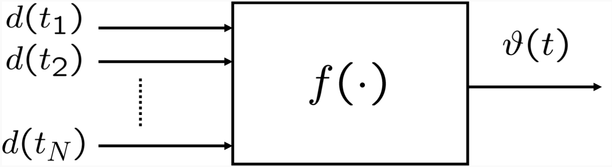
Unknown parameter(s)
$\vartheta(t)$

- In most cases the unknown parameters are constant

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

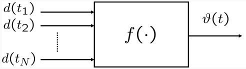
- $T = \{t_1\ , \ t_2\ , \ \dots\ , \ t_N\}$ set of the observation time-instants
  - In general, there is no need of equally-spaced t_i
  - If there is the possibility of choosing the instants t_i when to get experimental data, it is convenient to have more observations where the experiment is more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

Estimator



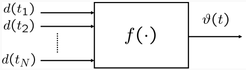
The estimator is a **deterministic function** yielding as output the unknown parameters on the basis of the observed data as inputs

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 $T = \{1\,,\,2\,,\,\ldots\,,\,N\}$.
- Typically we have three cases:
  - $t > t_N$: problem of prediction
  - $t = t_N$: problem of filtering
  - $t < t_N$: problem of smoothing

The estimation problem

Dynamical systems identification: the prediction problem

The prediction problem

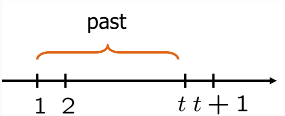
It is a fundamental problem in the context of **dynamical systems identification**

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

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

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

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

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

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

The prediction problem (cont.)

- A predictor is linear time-invariant if

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

  where the parameters $a_1\,,\;\ldots\;,\;a_n$ are constant
- We define the vector of parameters $\vartheta^T = [a_1\,,\;\ldots\;,\;a_n]$

> 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

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

  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: $y(i) \;\; i = 1\,,\;\ldots\;,\;t$
  - we compute

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

  - 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:

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

- Hence

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

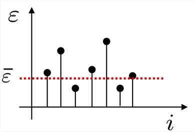
The determination of ϑ° is thus reduced to the solution of an optimization problem.

Remarks

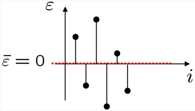
It is very important to clarify the meaning of ε "small"

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

(a)



(b)



- 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”

$$\varepsilon(\cdot) \sim \mathrm{WN}\left(0\,,\ \lambda^2\right)$$

white noise

average

variance

Predictor as a dynamic system

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

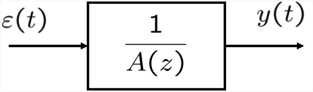
$$\varepsilon(t) = y(t) - \hat{y}(t|t-1) \quad \Rightarrow \quad y(t) = \varepsilon(t) + \hat{y}(t|t-1)$$

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

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

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

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



A Glimpse on Estimation theory & Estimators' characteristics

A Glimpse on Estimation theory & Estimators' characteristics

General concepts and definitions

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 depens on the result s of the random experiment

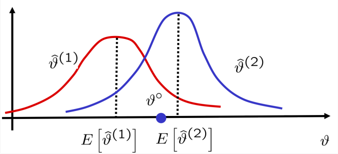
Bias

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

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

- 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



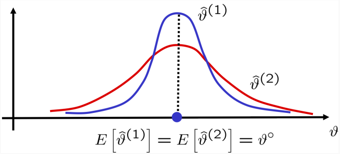
Minimum variance

- 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:

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



- 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

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

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

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

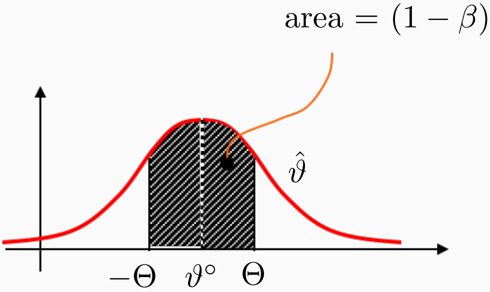
- Recalling that $A \geq 0 \implies \det A \geq 0\,,\ \lambda_i \geq 0\,,\ a_{ii} \geq 0$, we have

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

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 $\hat{\vartheta}$:

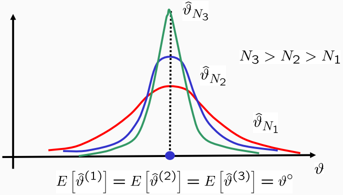


The estimate $\hat{\vartheta}$ belongs to the interval $(-\Theta\,,\ \Theta)$ around ϑ° with confidence $(1-\beta)\cdot 100\%$.

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

$$\lim_{N\to\infty} \mathrm{var}\left[\hat{\vartheta}_N\right] = 0$$



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

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

  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 $\left\|\hat{\vartheta}_N - \vartheta^\circ\right\|$ is a scalar random variable with a well-defined expected value.

Almost-sure convergence

- Recall that the estimator based on N data is

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

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

$$\hat{\vartheta}_1\left(s\,,\ \vartheta^\circ\right)\ ,\ \hat{\vartheta}_2\left(s\,,\ \vartheta^\circ\right)\ ,\ \ldots\ ,\ \hat{\vartheta}_N\left(s\,,\ \vartheta^\circ\right)\ ,\ \ldots$$

- It may happen that:

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

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

Almost-sure convergence (cont.)

- Introduce the set of random experiment results

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

- 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**.

A Glimpse on Estimation theory & Estimators' characteristics

Examples

Example 1

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

$$\mathrm{E}\left[d(i)\right] = \vartheta^{\circ}\,, \quad i = 1\,,\ 2\,,\ \ldots\,,\ N$$

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

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

- Consider the estimator

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

Sampled-average estimator

Example 1 (cont.)

- Bias:

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

the estimator is unbiased

- Variance:

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

$$= \mathrm{E}\left\{\frac{1}{N^2}\left[\sum_{i=1}^{N}d(i) - \sum_{i=1}^{N}\vartheta^{\circ}\right]^2\right\} = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{E}\left\{\left[d(i) - \vartheta^{\circ}\right]^2\right\}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{var}\left[d(i)\right]$$

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

Example 1 (cont.)

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

$$\lim_{N\to\infty} \mathrm{var}\left(\hat{\vartheta}_N\right) \leq \lim_{N\to\infty} \frac{\bar{\sigma}}{N} = 0$$

the estimator converges in quadratic mean

Example 2

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

$$\mathrm{E}\left[d(i)\right] = \vartheta^{\circ}\ , \quad i = 1\,,\ 2\,,\ \ldots\,,\ N$$

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

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

- Consider the estimator

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

Example 2 (cont.)

- Bias:

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

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

N.B. in the previous case $\alpha(i) = \dfrac{1}{N}$ 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

un-correlated data

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

By using the Lagrange multipliers technique we have:

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

Example 2 (cont.)

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

- Now, imposing the constraint $(\star)$ for unbiasedness

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

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

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

Example 2 (cont.)

- Let us compute the estimator's variance:

$$\operatorname{var}\left(\hat{\vartheta}_N\right) = \mathrm{E}\left\{\left[\hat{\vartheta}_N - \mathrm{E}\left(\hat{\vartheta}_N\right)\right]^2\right\} = \mathrm{E}\left\{\left[\sum_{i=1}^{N} \alpha(i) d(i) - \vartheta^\circ \sum_{i=1}^{N} \alpha(i)\right]^2\right\}$$

$$= \mathrm{E}\left\{\left[\sum_{i=1}^{N} \alpha(i)\left[d(i) - \vartheta^\circ\right]\right]^2\right\} = \sum_{i=1}^{N} [\alpha(i)]^2\, \mathrm{E}\left\{[d(i) - \vartheta^\circ]^2\right\}$$

$$= \sum_{i=1}^{N} (\alpha(i))^2 \operatorname{var}[d(i)] = \alpha^2 \sum_{i=1}^{N} \frac{1}{\operatorname{var}[d(i)]} = \frac{1}{\displaystyle\sum_{i=1}^{N} \frac{1}{\operatorname{var}[d(i)]}}$$

Example 2 (cont.)

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

$$\lim_{N\to\infty} \mathrm{var}\left(\hat{\vartheta}_N\right) \leq \lim_{N\to\infty} \frac{\bar{\sigma}}{N} = 0$$

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

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

  where c is a suitably small positive scalar
- In this time-varying case what matters is not "convergence" but "boundedness"

267MI –Fall 2019

Lecture 6
Definitions and Properties of the Estimation and Prediction Problems

END