ESTIMATION THEORY

Michele TARAGNA

Dipartimento di Elettronica e Telecomunicazioni

Politecnico di Torino

michele.taragna@polito.it

Master Course in Mechatronic Engineering

Master Course in Computer Engineering

01RKYQW / 01RKYOV “Estimation, Filtering and System Identification”

Academic Year 2021/2022

Estimation problem

The estimation problem refers to the empirical evaluation of an uncertain variable, like an unknown characteristic parameter or a remote signal, on the basis of observations and experimental measurements of the phenomenon under investigation.

An estimation problem always assumes a suitable mathematical description (*model*) of the phenomenon:

- in the classical statistics, the investigated problems usually involve *static models*, characterized by instantaneous (or algebraic) relationships among variables;
- in this course, estimation methods are introduced also for phenomena that are adequately described by *discrete-time dynamic models*, characterized by relationships among variables that can be represented by means of difference equations (i.e., for simplicity, the time variable is assumed to be discrete).

Estimation problem

$\theta(t)$: real variable to be estimated, scalar or vector, constant or time-varying;

$d(t)$: available data, acquired at N time instants $t_1, t_2, \ldots, t_N$;

$T = \{t_1, t_2, \ldots, t_N\}$: set of time instants used for observations, distributed with regularity (in this case, $T = \{1, 2, \ldots, N\}$) or non-uniformly;

$d = \{d(t_1), d(t_2), \ldots, d(t_N)\}$: observation set.

An **estimator** (or **estimation algorithm**) is a *function* $f(\cdot)$ that, starting from data, associates a value to the variable to be estimated:

$$\theta(t) = f(d)$$

The **estimate** term refers to the particular *value* given by the estimator when applied to the particular observed data.

Estimation problem classification

1) $\theta(t)$ is constant => **parametric identification** problem:

- the estimator is denoted by $\hat{\theta}$ or by $\hat{\theta}_T$;
- the true value of the unknown variable (if makes sense) is denoted by θ_o;

2) $\theta(t)$ is a time-varying function:

- the estimator is denoted by $\hat{\theta}\left(t|T\right)$, or by $\hat{\theta}\left(t|N\right)$ if the time instants for observations are uniformly distributed;
- according to the temporal relationship between t and the last time instant t_N:

2.a) if $t > t_N$ => **prediction** problem;

2.b) if $t = t_N$ => **filtering** problem;

2.c) if $t_1 < t < t_N$ => **regularization** or **interpolation** or **smoothing** problem.

Example of prediction problem: time series analysis

Given a sequence of observations (time series or historical data set) of a variable y:

$$y(1)\,, y(2)\,, \ldots\,, y(t)$$

the goal is to evaluate the next value $y(t+1)$ of this variable

$$\Downarrow$$

it is necessary to find a good **predictor** $\hat{y}(t+1|t)$, i.e., a function of available data that provides the most accurate evaluation of the next value of the variable y:

$$\hat{y}(t+1|t) = f\left(y(t)\,, y(t-1)\,, \ldots\,, y(1)\right) \cong y(t+1)$$

A predictor is said to be *linear* if it is a linear function of data:

$$\hat{y}(t+1|t) = a_1(t)y(t) + a_2(t)y(t-1) + \ldots + a_t(t)y(1) = \sum_{k=1}^{t} a_k(t)y(t-k+1)$$

A linear predictor has *a finite memory* n if it is a linear function of the last n data only:

$$\hat{y}(t+1|t) = a_1(t)y(t) + a_2(t)y(t-1) + \ldots + a_n(t)y(t-n+1) = \sum_{k=1}^{n} a_k(t)y(t-k+1)$$

If all the parameters $a_i(t)$ are constant, the predictor is also *time-invariant*:

$$\hat{y}(t+1|t) = a_1 y(t) + a_2 y(t-1) + \ldots + a_n y(t-n+1) = \sum_{k=1}^{n} a_k y(t-k+1)$$

and it is characterized by the vector of constant parameters

$$\theta = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^T \in \mathbb{R}^n$$

$$\Downarrow$$

The prediction problem becomes a parametric identification problem.

Questions:

- how to measure the predictor quality?
- how to derive the “best” predictor?

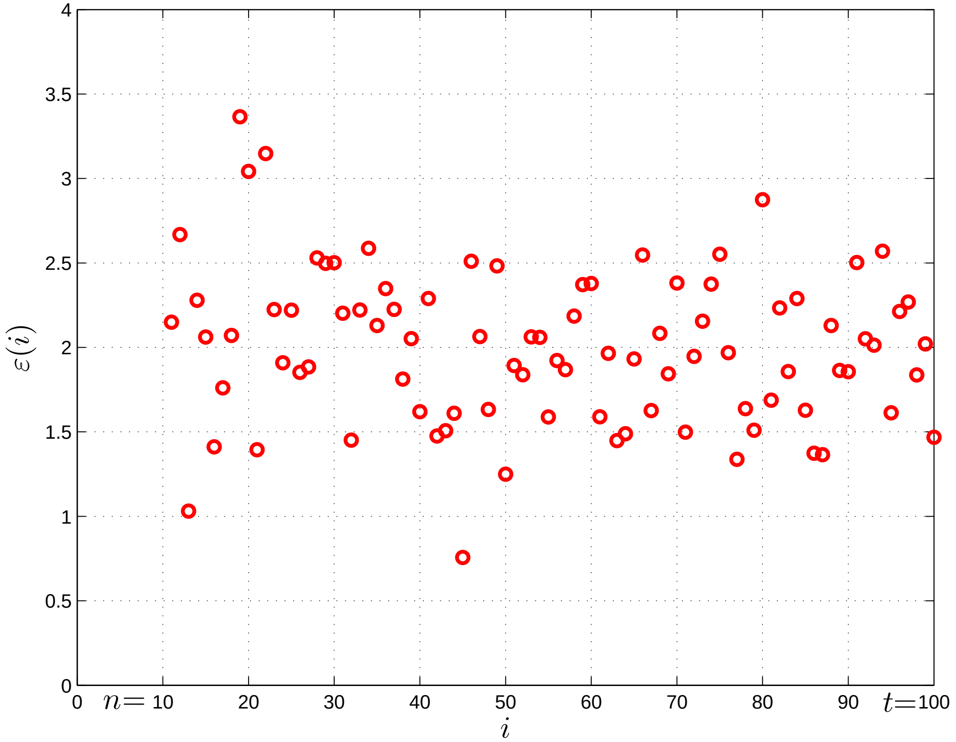
If the predictive model is linear, time-invariant, with finite memory n much shorter than the total number of data measured up to time instant t, its predictive capability over the available data $y(i)$, $i = 1, 2, \ldots, t$, can be evaluated in the following way:

- at each instant $i \geq n$, the prediction $\hat{y}(i+1|i)$ of the next value is computed:
  $\hat{y}(i+1|i) = a_1 y(i) + a_2 y(i-1) + \ldots + a_n y(i-n+1) = \sum_{k=1}^{n} a_k y(i-k+1)$
  and its *prediction error* $\varepsilon(i+1)$ with respect to $y(i+1)$ is evaluated:
  $$\varepsilon(i+1) = y(i+1) - \hat{y}(i+1|i)$$
- the model described by θ is a good predictive model if the error ε is "small" over all the available data $\Rightarrow$ the following figure of merit is introduced:
  $$J(\theta) = \sum_{k=n+1}^{t} \varepsilon(k)^2 \qquad \text{(sum of squares of prediction errors)}$$
- the best predictor is the one that minimizes J and the value of its parameters is:
  $$\theta^* = \arg\min_{\theta \in \mathbb{R}^n} J(\theta)$$

For example, if $t = 100$ and $n = 10 \ll t$, for a given $\theta = [a_1 \cdots a_{10}]^T$ it results:

$$\begin{cases} \hat{y}(11|10) = a_1 y(10) + \ldots + a_{10} y(1) & \Rightarrow \quad \varepsilon(11) = y(11) - \hat{y}(11|10) \\ \hat{y}(12|11) = a_1 y(11) + \ldots + a_{10} y(2) & \Rightarrow \quad \varepsilon(12) = y(12) - \hat{y}(12|11) \\ \quad \vdots & \quad \vdots \\ \hat{y}(100|99) = a_1 y(99) + \ldots + a_{10} y(90) & \Rightarrow \quad \varepsilon(100) = y(100) - \hat{y}(100|99) \end{cases}$$

and then the behaviour of the prediction error sequence $\varepsilon(\cdot)$ is plotted:



Fundamental question: is the predictor minimizing J necessarily a “good” model?

The predictor quality depends on the fact that the temporal behaviour of the prediction error sequence $\varepsilon(\cdot)$ has the following characteristics:

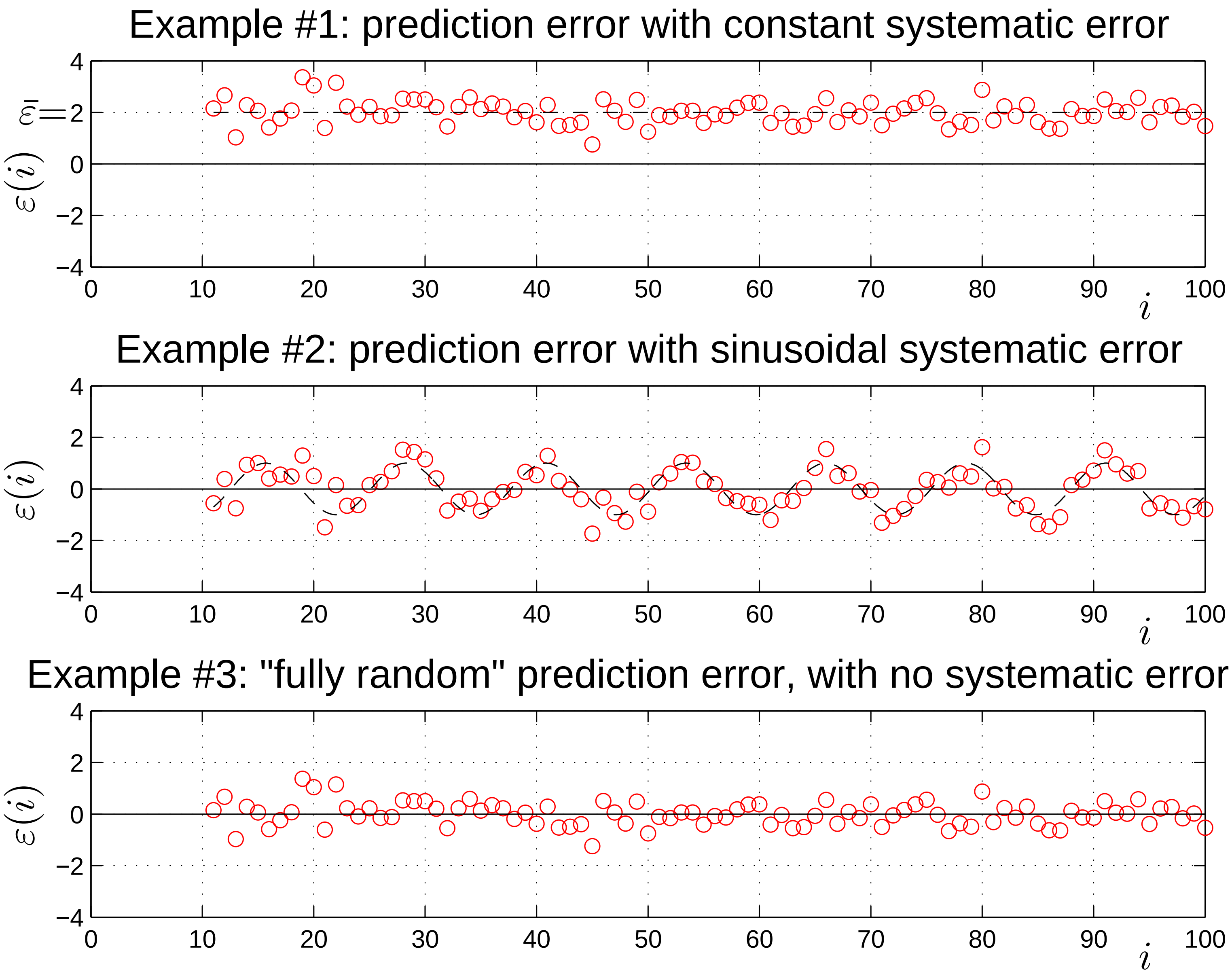
- its mean value is zero, i.e., it does not show a systematic error;
- it is “fully random”, i.e., it does not contain any regularity element.

In probabilistic terms, this corresponds to require that the behaviour of the error $\varepsilon(\cdot)$ is that of a **white noise** (WN) process, i.e., a sequence of independent random variables with zero mean value and constant variance σ^2:

$$\varepsilon(\cdot) = WN\left(0, \sigma^2\right)$$

$$\Downarrow$$

A predictor is a “good” model if $\varepsilon(\cdot)$ has the white noise probabilistic characteristics.


Example #1: prediction error with constant systematic error
$\bar{\varepsilon}=2$
$\varepsilon(i)$
4
0
−2
−4
0 10 20 30 40 50 60 70 80 90 100
i
Example #2: prediction error with sinusoidal systematic error
$\varepsilon(i)$
4
2
0
−2
−4
0 10 20 30 40 50 60 70 80 90 100
i
Example #3: "fully random" prediction error, with no systematic error
$\varepsilon(i)$
4
2
0
−2
−4
0 10 20 30 40 50 60 70 80 90 100
i

Then, the prediction problem can be recast as the study of a **stochastic system**, i.e., a dynamic system whose inputs are probabilistic signals; in fact:

$$\begin{cases} \hat{y}(t|t-1) = a_1 y(t-1) + a_2 y(t-2) + \ldots + a_n y(t-n) \\ \varepsilon(t) = y(t) - \hat{y}(t|t-1) \end{cases} \Rightarrow$$

$$y(t) = \hat{y}(t|t-1) + \varepsilon(t) = a_1 y(t-1) + a_2 y(t-2) + \ldots + a_n y(t-n) + \varepsilon(t)$$

represents a discrete-time LTI dynamic system with output $y(t)$ and input $\varepsilon(t)$

$$\Downarrow$$

$\mathcal{Z}$-transforming, with $\mathcal{Z}[y(t-k)] = z^{-k}Y(z)$ and z^{-1} the unitary delay operator:

$$Y(z) = a_1 z^{-1} Y(z) + a_2 z^{-2} Y(z) + \ldots + a_n z^{-n} Y(z) + \varepsilon(z)$$

$$\Downarrow$$

$$H(z) = \frac{Y(z)}{\varepsilon(z)} = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_n z^{-n}} = \frac{z^n}{z^n - a_1 z^{n-1} - a_2 z^{n-2} - \ldots - a_n}$$

represents the transfer function of a LTI dynamic system $\Rightarrow$ in order to be a “good” model, its input $\varepsilon(\cdot)$ shall have the white noise probabilistic characteristics.

Classification of data descriptions

- The actually available information is always:
  - bounded $\Rightarrow$ the measurement number N is necessarily finite;
  - corrupted by different kinds of uncertainty (e.g., measurement noise).

- The uncertainty affecting the data can be described:
  - in probabilistic terms $\Rightarrow$ we talk about **statistical** or **classical estimation**;
  - in terms of set theory, as a member of some bounded set $\Rightarrow$
    we talk about **Set Membership** or **Unknown But Bounded** (UBB) **estimation**.

Random experiment and random source of data

S : **outcome space**, i.e., the set of possible outcomes s of the random experiment;

$\mathcal{F}$: **space of results of interest**, i.e., the set of the combinations of interest where the outcomes in S can be clustered;

$P(\cdot)$: **probability** function defined in $\mathcal{F}$ that associates to any event in $\mathcal{F}$ a real number between 0 and 1.

$\mathcal{E} = (S, \mathcal{F}, P(\cdot))$: **random experiment**

Example: throw a dice with six sides to see if an odd or even number is drawn $\Rightarrow$

- $S = \{1, 2, 3, 4, 5, 6\}$ is the set of 6 sides of the dice;
- $\mathcal{F} = \{A, B, S, \emptyset\}$, with $A = \{2, 4, 6\}$ and $B = \{1, 3, 5\}$ the results of interest, i.e., the even and odd number sets;
- $P(A) = P(B) = 1/2$ (if the dice is not fixed), $P(S) = 1$, $P(\emptyset) = 0$.

A **random variable** of the experiment $\mathcal{E}$ is a variable v whose values depend on the outcome s of $\mathcal{E}$ through of a suitable function $\varphi(\cdot) : S \to V$, where V is the set of possible values of v:

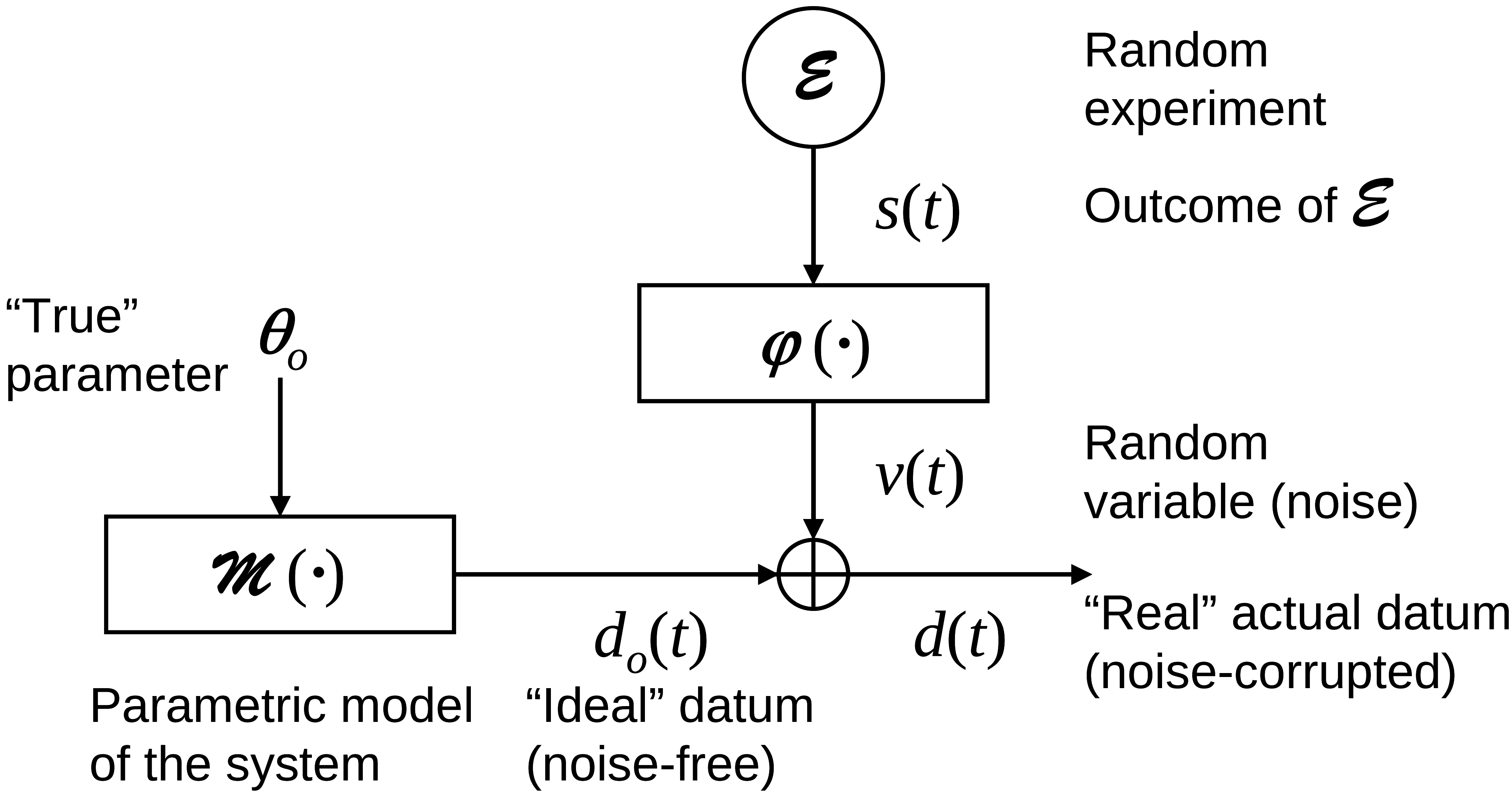
$$v = \varphi(s)$$

Example: the random variable depending on the outcome of the throw of a dice with six sides can be defined as

$$v = \varphi(s) = \begin{cases} +1 & \text{if } s \in A = \{2, 4, 6\} \\ -1 & \text{if } s \in B = \{1, 3, 5\} \end{cases}$$

A **random source of data** produces data that, besides the process under investigation characterized by the unknown true value θ_o of the variable to be estimated, are also functions of a random variable; in particular, at the time instant t, the datum $d(t)$ depends on the random variable $v(t)$.

Random source of data:



Probabilistic description of data

In the *probabilistic* (or *classical* or *statistical*) framework, data d are assumed to be produced by a random source of data $\mathcal{S}$, influenced by:

- the outcome s of a random experiment $\mathcal{E}$
- the "true" value θ_o of the unknown variable to be estimated

$$d = d\,(s, \theta_o)$$

$$\Downarrow$$

data d are random variables, since they are functions of the outcome s

$$\Downarrow$$

A full probabilistic description of data is constituted by

- its **probability distribution** $F(q) = Prob\,\{d\,(s, \theta_o) \leq q\}$ or
- its **probability density function** $f(q) = \dfrac{dF(q)}{dq}$, often denoted by p.d.f.

Estimator characteristics

A random source of data $\mathcal{S}$, influenced by the outcome s of a random experiment $\mathcal{E}$ and by the "true" value θ_o of the unknown variable to be estimated, produces data d:

$$d = d\,(s, \theta_o)$$

$$\Downarrow$$

data d are random variables, since they are functions of the outcome s

$$\Downarrow$$

the estimator $f(\cdot)$ and the estimate $\hat{\theta}$ are random variables too, being functions of d:

$$\hat{\theta} = f(d) = f(d\,(s, \theta_o))$$

$$\Downarrow$$

the quality of $f(\cdot)$ and $\hat{\theta}$ depends on their probabilistic characteristics.

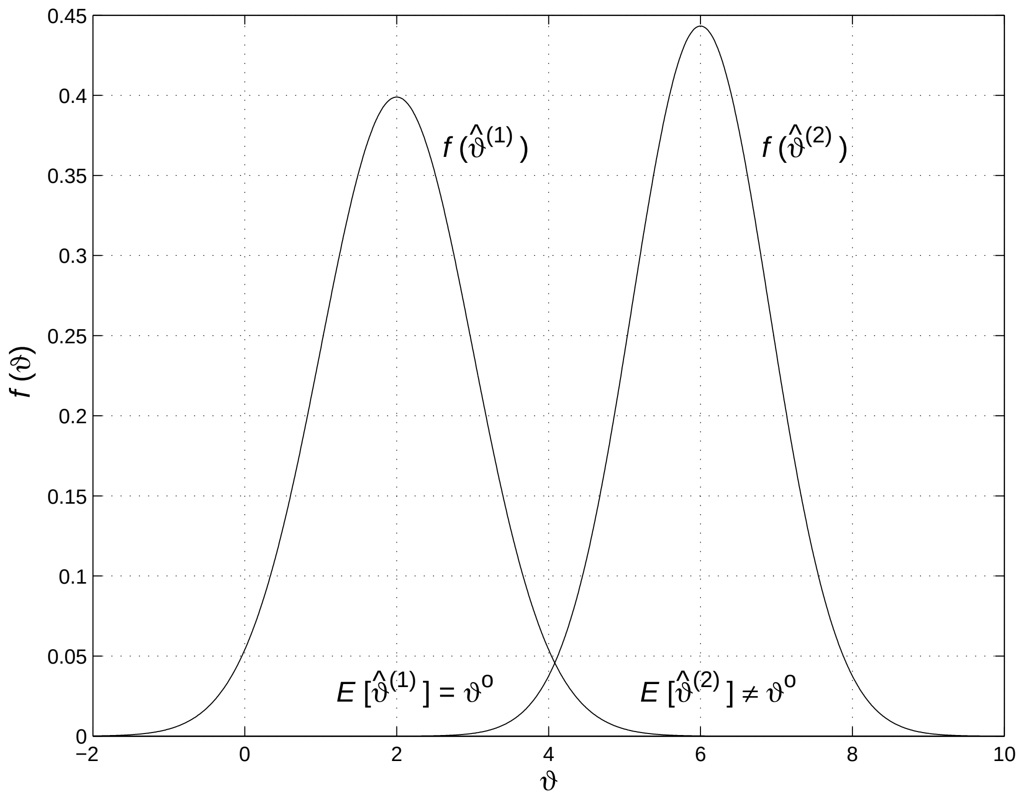
Estimator probabilistic characteristics

- No bias (in order to avoid to introduce any systematic estimation error)
- Minimum variance (smaller scattering around the mean value guarantees higher probability of obtaining values close to the “true” value θ_o)
- Asymptotic characteristics (for $N \to \infty$):
  - quadratic-mean convergence
  - almost-sure convergence
  - consistency

Estimator probabilistic characteristics

An estimator is said to be **unbiased** (or **correct**) if

$$E\left[\hat{\theta}\right] = \theta_o$$

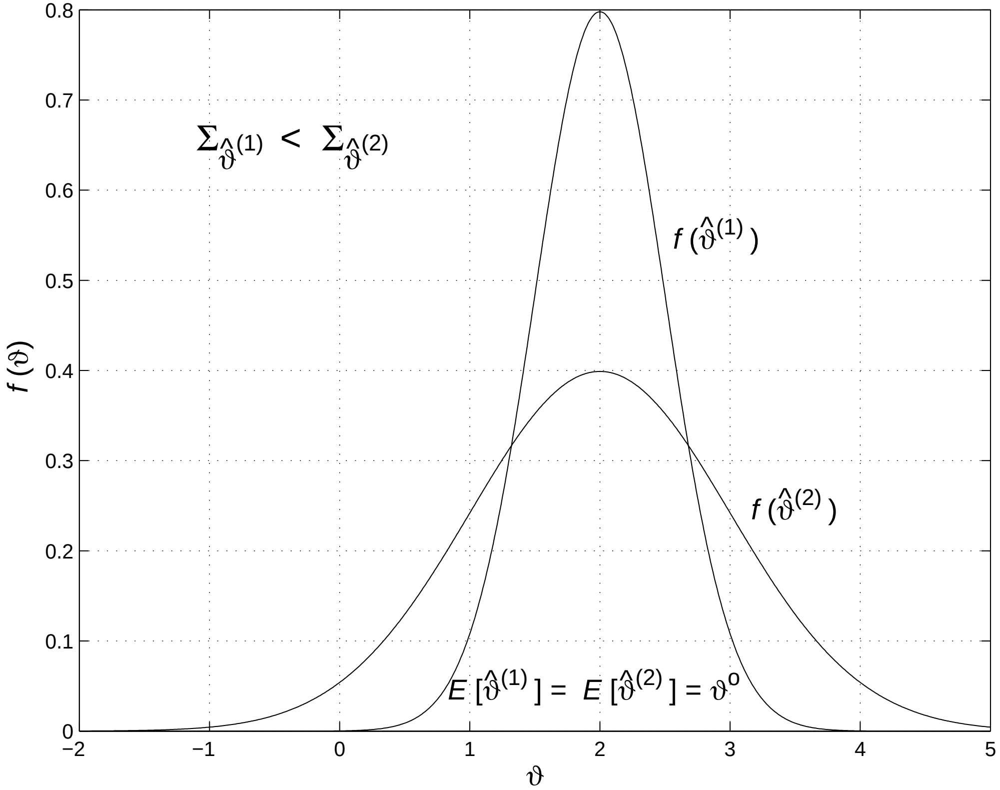


An unbiased estimator does not introduce any systematic estimation error.

Estimator probabilistic characteristics

An unbiased estimator $\hat{\theta}^{(1)}$ is said to be **efficient** (or with **minimum variance**) if

$$Var[\hat{\theta}^{(1)}] \leq Var[\hat{\theta}^{(2)}], \quad \forall \hat{\theta}^{(2)} \neq \hat{\theta}^{(1)}$$



Smaller scattering around the mean value $\Rightarrow$ higher probability of approaching θ_o.

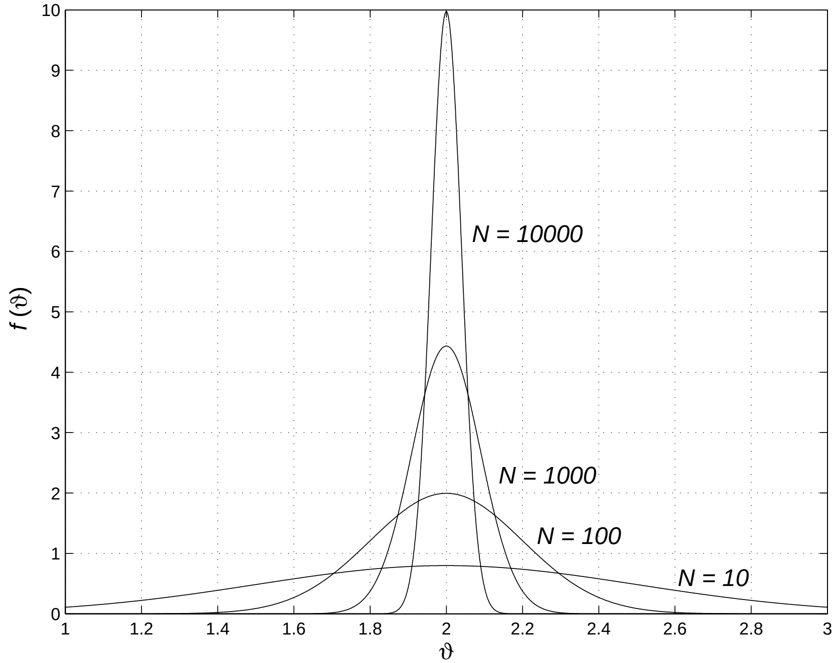
Estimator probabilistic characteristics

An unbiased estimator **converges in quadratic mean to θ_o**, i.e., $\underset{N\to\infty}{l.i.m.}\,\hat{\theta}_N = \theta_o$, if

$$\lim_{N\to\infty} E\left[\|\hat{\theta}_N - \theta_o\|^2\right] = 0$$

where $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}, \quad \forall x \in \mathbb{R}^n$, is the Euclidean norm.

An unbiased estimator such that $\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] = 0$ converges in quadratic mean:



Sure and almost-sure convergence, consistency

An estimator is function of both the outcome s of a random experiment $\mathcal{E}$ and θ_o:

$$\hat{\theta} = f(d) = f(d(s, \theta_o)) \quad \Rightarrow \quad \hat{\theta} = \hat{\theta}(s, \theta_o)$$

If a particular outcome $\overline{s} \in S$ is considered and the sequence of estimates $\hat{\theta}_N(\overline{s}, \theta_o)$ is evaluated for increasing N, a numerical series $\hat{\theta}_1(\overline{s}, \theta_o)$, $\hat{\theta}_2(\overline{s}, \theta_o)$, . . ., is derived that may converge to θ_o for some $\overline{s}$, and may not converge for some other $\overline{s}$.

Let A be the set of outcomes $\overline{s}$ guaranteeing the convergence to θ_o:

- if $A \equiv S$, then we have **sure convergence**, since it holds $\forall \overline{s} \in S$;
- if $A \subset S$, considering A like an event, the probability $P(A)$ may be defined; if A is such that $P(A) = 1$, we say that $\hat{\theta}_N$ converges to θ_o *with probability* 1:

$$\lim_{N \to \infty} \hat{\theta}_N = \theta_o \qquad w.p.1$$

  we have **almost-sure convergence** $\Rightarrow$ the algorithm is said to be **consistent**.

Example

Problem: N scalar data d_i with the same mean value $E[d_i] = \theta_o$, with variances $Var[d_i]$ possibly different but bounded ($\exists \sigma \in \mathbb{R}_+ : Var[d_i] \leq \sigma^2 < \infty, \forall i$); data are uncorrelated, i.e.:

$$E[\{d_i - E[d_i]\}\{d_j - E[d_j]\}] = 0, \quad \forall i \neq j$$

Estimator #1 (sample mean):

$$\boxed{\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} d_i}$$

- it is an unbiased estimator:

$$E\left[\hat{\theta}_N\right] = E\left[\frac{1}{N}\sum_{i=1}^{N} d_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[d_i] = \frac{1}{N}\sum_{i=1}^{N} \theta_o = \theta_o$$

- it converges in quadratic mean:

$$
\begin{aligned}
Var\left[\hat{\theta}_N\right] &= E\left[\left(\hat{\theta}_N - E\left[\hat{\theta}_N\right]\right)^2\right] = E\left[\left(\tfrac{1}{N}\textstyle\sum_{i=1}^N d_i - \theta_o\right)^2\right] = \\
&= E\left[\left(\tfrac{1}{N}\textstyle\sum_{i=1}^N d_i - \tfrac{1}{N}\sum_{i=1}^N \theta_o\right)^2\right] = E\left[\left(\tfrac{1}{N}\textstyle\sum_{i=1}^N (d_i - \theta_o)\right)^2\right] = \\
&= E\left[\tfrac{1}{N^2}\left(\textstyle\sum_{i=1}^N (d_i - \theta_o)\right)^2\right] = \tfrac{1}{N^2}E\left[\left(\textstyle\sum_{i=1}^N (d_i - \theta_o)\right)^2\right] = \\
&= \tfrac{1}{N^2}E\left[\textstyle\sum_{i=1}^N (d_i - \theta_o)^2 + \sum_{i=1}^N (d_i - \theta_o)\sum_{j=1,j\neq i}^N (d_j - \theta_o)\right] = \\
&= \tfrac{1}{N^2}\left\{\textstyle\sum_{i=1}^N E\left[(d_i - \theta_o)^2\right] + \sum_{i=1}^N E\left[(d_i - \theta_o)\sum_{j=1,j\neq i}^N (d_j - \theta_o)\right]\right\} = \\
&= \tfrac{1}{N^2}\textstyle\sum_{i=1}^N Var\left[d_i\right] \leq \tfrac{1}{N^2}\sum_{i=1}^N \sigma^2 = \sigma^2/N
\end{aligned}
$$

$$\Downarrow$$

$$\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] \leq \lim_{N\to\infty}\frac{\sigma^2}{N} = 0$$

$$\Downarrow$$

the algorithm converges in quadratic mean, since it is unbiased and with $\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] = 0$.

Estimator #2:

$$\boxed{\hat{\theta}_N = d_j}$$

- it is an unbiased estimator:

$$E\left[\hat{\theta}_N\right] = E\left[d_j\right] = \theta_o$$

- it does not converge in quadratic mean:

$$Var\left[\hat{\theta}_N\right] = E\left[\left(\hat{\theta}_N - E\left[\hat{\theta}_N\right]\right)^2\right] = E\left[\left(d_j - \theta_o\right)^2\right] = Var\left[d_j\right] \leq \sigma^2$$

and then it does not vary with the number N of data

$$\Downarrow$$

the estimation uncertainty is constant and, in particular, it does not decrease when the number of data grows.

Estimator #3 (weighted sample mean):

$$\boxed{\hat{\theta}_N = \sum_{i=1}^{N} \alpha_i d_i}$$

- it is an unbiased estimator if and only if $\sum_{i=1}^{N} \alpha_i = 1$, because

$$E\left[\hat{\theta}_N\right] = E\left[\sum_{i=1}^{N} \alpha_i d_i\right] = \sum_{i=1}^{N} \alpha_i E\left[d_i\right] = \theta_o \sum_{i=1}^{N} \alpha_i = \theta_o \;\Leftrightarrow\; \sum_{i=1}^{N} \alpha_i = 1$$

  Note: the algorithm #1 corresponds to the case $\alpha_i = \frac{1}{N}$, $\forall i$;
  the algorithm #2 corresponds to the case $\alpha_j = 1$ and $\alpha_i = 0$, $\forall i \neq j$

- it can be proven that the minimum variance unbiased estimator has weights

$$\alpha_i = \frac{\alpha}{Var\left[d_i\right]}, \quad \alpha = \left[\sum_{i=1}^{N} \frac{1}{Var\left[d_i\right]}\right]^{-1}$$

  intuitively, more uncertain data are considered as less trusted, with lower weights

- the variance of the minimum variance unbiased estimator is

$$
\begin{aligned}
Var\left[\hat{\theta}_N\right] &= E\left[\left(\hat{\theta}_N - E\left[\hat{\theta}_N\right]\right)^2\right] = E\left[\left(\sum_{i=1}^N \alpha_i d_i - \theta_o\right)^2\right] = \\
&= E\left[\left(\sum_{i=1}^N \alpha_i d_i - \sum_{i=1}^N \alpha_i \theta_o\right)^2\right] = E\left[\left(\sum_{i=1}^N \alpha_i \left(d_i - \theta_o\right)\right)^2\right] = \\
&= E\left[\sum_{i=1}^N \alpha_i^2 (d_i - \theta_o)^2 + \sum_{i=1}^N \alpha_i (d_i - \theta_o) \sum_{j=1, j\neq i}^N \alpha_j (d_j - \theta_o)\right] = \\
&= \sum_{i=1}^N \alpha_i^2 E\left[(d_i - \theta_o)^2\right] + \sum_{i=1}^N \alpha_i E\left[(d_i - \theta_o) \sum_{j=1, j\neq i}^N \alpha_j (d_j - \theta_o)\right] = \\
&= \sum_{i=1}^N \alpha_i^2 Var\left[d_i\right] = \sum_{i=1}^N \frac{\alpha^2}{Var[d_i]^2} Var[d_i] = \alpha^2 \sum_{i=1}^N \frac{1}{Var[d_i]} = \\
&= \alpha = \left[\sum_{i=1}^N \frac{1}{Var\left[d_i\right]}\right]^{-1} \leq \left[\sum_{i=1}^N \frac{1}{\sigma^2}\right]^{-1} = \frac{\sigma^2}{N}
\end{aligned}
$$

- the minimum variance unbiased algorithm converges in quadratic mean, since

$$
\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] \leq \lim_{N\to\infty} \frac{\sigma^2}{N} = 0
$$

Cramér-Rao inequality

The estimation precision has its own intrinsic limits, due to the random source of data: in fact, the variance of any estimator cannot be less than a certain value, since data are always affected by noises and the corresponding uncertainty reflects into a structural estimate uncertainty, which cannot be suppressed simply by changing the estimator:

- in the scalar case $\theta \in \mathbb{R}$, the following **Cramér-Rao inequality** holds for any unbiased estimator $\hat{\theta}$:

$$Var\left[\,\hat{\theta}\,\right] \geq m^{-1}$$

where m is the **Fisher information quantity** defined as

$$m = E\left[\left\{\frac{\partial}{\partial\theta}\ln f(d^{(\theta)},\theta)\right\}^2\right]_{\theta=\theta_o} = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(d^{(\theta)},\theta)\right]_{\theta=\theta_o} \geq 0$$

$d^{(\theta)} \in \mathbb{R}^N$ are the data generated by the random source for a generic value θ of the unknown variable, not necessarily the “true” value θ_o; $f(q,\theta)$, $q \in \mathbb{R}^N$, is the probability density function of q;

- in the vector case $\theta \in \mathbb{R}^n$, for any unbiased estimator $\hat{\theta}$, the **Cramér-Rao inequality** becomes

$$Var\left[\,\hat{\theta}\,\right] \geq M^{-1}$$

where M is the nonsingular **Fisher information matrix**

$$M = [m_{ij}] \in \mathbb{R}^{n \times n}$$

$$m_{ij} = -E\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f(d^{(\theta)}, \theta)\right]_{\theta=\theta_o}, \quad \forall i, j = 1, 2, \ldots, n$$

From this inequality it follows that

$$Var\left[\,\hat{\theta}_i\,\right] \geq \left[M^{-1}\right]_{ii}, \quad \forall i = 1, 2, \ldots, n$$

An unbiased estimator is **efficient** if it provides the minimum variance, i.e., if its variance achieves the minimal theoretic value assessed by the Cramér-Rao inequality:

$$Var\left[\,\hat{\theta}\,\right] = m^{-1} \quad \text{or} \quad Var\left[\,\hat{\theta}\,\right] = M^{-1}$$

Least Squares estimation method

Linear regression problem: given the measurements of $n+1$ real variables $y(t)$, $u_1(t), \ldots, u_n(t)$ over a time interval (e.g., for $t = 1, 2, \ldots, N$), find if possible the values of n real parameters $\theta_1, \theta_2, \ldots, \theta_n$ such that the following relationship holds

$$y(t) = \theta_1 u_1(t) + \ldots + \theta_n u_n(t)$$

In matrix terms, by defining the real vectors

$$\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^n, \quad \varphi(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{bmatrix} \in \mathbb{R}^n \quad \Rightarrow \quad y(t) = \varphi(t)^T \theta$$

In the actual problems, there exists always a nonzero error $\varepsilon(t) = y(t) - \varphi(t)^T \theta$

$$\Downarrow$$

by defining $J(\theta) = \sum_{t=1}^{N} \varepsilon(t)^2$, the problem is solved by finding $\theta^* = \arg\min_{\theta \in \mathbb{R}^n} J(\theta)$.

In order to find the minimum of the figure of merit J, we have to require that

$$\frac{dJ(\theta)}{d\theta} = \left[\begin{array}{ccc} \frac{dJ(\theta)}{d\theta_1} & \cdots & \frac{dJ(\theta)}{d\theta_n} \end{array}\right] = 0 \qquad \Leftrightarrow$$

$$\begin{aligned} \frac{dJ(\theta)}{d\theta_i} &= \frac{d}{d\theta_i}\left[\sum_{t=1}^{N} \varepsilon(t)^2\right] = \sum_{t=1}^{N} \frac{d}{d\theta_i}\left[\varepsilon(t)^2\right] = \sum_{t=1}^{N} \frac{d}{d\theta_i}\left[\left(y(t) - \varphi(t)^T\,\theta\right)^2\right] = \\ &= -2\sum_{t=1}^{N}\left(y(t) - \varphi(t)^T\,\theta\right) u_i(t) = 0, \qquad i = 1, 2, \ldots, n \qquad \Leftrightarrow \end{aligned}$$

$$\frac{dJ(\theta)}{d\theta} = -2\sum_{t=1}^{N}\left(y(t) - \varphi(t)^T\,\theta\right)\varphi(t)^T = 0 \qquad \Leftrightarrow$$

$$\sum_{t=1}^{N}\left(y(t)\,\varphi(t)^T - \varphi(t)^T\,\theta\varphi(t)^T\right) = \sum_{t=1}^{N} y(t)\,\varphi(t)^T - \sum_{t=1}^{N}\varphi(t)^T\,\theta\varphi(t)^T = 0 \qquad \Leftrightarrow$$

$$\sum_{t=1}^{N}\varphi(t)^T\,\theta\varphi(t)^T = \sum_{t=1}^{N} y(t)\,\varphi(t)^T \qquad \Leftrightarrow$$

$$\sum_{t=1}^{N}\left[\varphi(t)\,\varphi(t)^T\right]\theta = \sum_{t=1}^{N}\varphi(t)\,y(t)$$

The relationship

$$\sum_{t=1}^{N}\left[\varphi(t)\,\varphi(t)^T\right]\theta=\sum_{t=1}^{N}\varphi(t)\,y(t)$$

is a system of n scalar equations involving n scalar unknowns $\theta_1,\theta_2,\ldots,\theta_n$ that is called **normal equation system**:

- if the matrix $\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T$ is nonsingular ($\Leftrightarrow \det\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\neq 0$, known as *identifiability condition*), then the normal equation system has a single *unique* solution given by the **Least Squares (LS) estimate**:

$$\boxed{\hat{\theta}=\left[\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\right]^{-1}\left[\sum_{t=1}^{N}\varphi(t)\,y(t)\right]}$$

- if $\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T$ is singular, it can be proved that the normal equations have an infinite number of solutions, due to their particular structure.

The stationarity condition $\frac{dJ(\theta)}{d\theta}=0$ does not guarantee that $\hat{\theta}$ is a minimum of $J(\theta)$
$\Rightarrow$ we have to consider the Hessian matrix

$$\begin{aligned}
\frac{d^2 J(\theta)}{d\theta^2} &= \frac{d}{d\theta}\left[\frac{dJ(\theta)}{d\theta}\right]^T = \frac{d}{d\theta}\left[-2\sum_{t=1}^{N}\left(y(t)-\varphi(t)^T\,\theta\right)\varphi(t)^T\right]^T = \\
&= \frac{d}{d\theta}\left[-2\sum_{t=1}^{N}\left(y(t)\,\varphi(t)^T-\theta^T\varphi(t)\,\varphi(t)^T\right)^T\right] = \\
&= \frac{d}{d\theta}\left[-2\sum_{t=1}^{N}y(t)\,\varphi(t)+2\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\,\theta\right] = \\
&= 2\sum_{t=1}^{N}\frac{d}{d\theta}\varphi(t)\,\varphi(t)^T\,\theta = 2\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T
\end{aligned}$$

that turns out to be positive semidefinite, since $\forall x\in\mathbb{R}^n$

$$x^T\frac{d^2 J(\theta)}{d\theta^2}x = x^T 2\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T\,x = 2\sum_{t=1}^{N}x^T\varphi(t)\,\varphi(t)^T\,x = 2\sum_{t=1}^{N}\left(x^T\varphi(t)\right)^2 \geq 0$$

$$\Downarrow$$

$\hat{\theta}$ is certainly a (local or global) minimum of $J(\theta)$.

The Taylor series expansion of $J(\theta)$ in the neighborhood of $\hat{\theta}$ allows to understand if $\hat{\theta}$ is a local or global minimum:

$$J(\theta) = J(\hat{\theta}) + \left.\frac{dJ(\theta)}{d\theta}\right|_{\hat{\theta}}(\theta-\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^T \left.\frac{d^2J(\theta)}{d\theta^2}\right|_{\hat{\theta}}(\theta-\hat{\theta}) + \ldots = J(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^T \left.\frac{d^2J(\theta)}{d\theta^2}\right|_{\hat{\theta}}(\theta-\hat{\theta})$$

since the term $\left.\frac{dJ(\theta)}{d\theta}\right|_{\theta=\hat{\theta}}$ is zero ($\hat{\theta}$ is a minimum) as well as all the $J(\theta)$ derivatives of order greater than two ($J(\theta)$ is a quadratic function of θ)

$$\Downarrow$$

$$J(\theta) - J(\hat{\theta}) = \frac{1}{2}(\theta-\hat{\theta})^T \left.\frac{d^2J(\theta)}{d\theta^2}\right|_{\hat{\theta}}(\theta-\hat{\theta}), \qquad \left.\frac{d^2J(\theta)}{d\theta^2}\right|_{\hat{\theta}} = 2\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T,$$

is a positive semidefinite quadratic form, since $\left.\frac{d^2J(\vartheta)}{d\vartheta^2}\right|_{\hat{\theta}}$ is positive semidefinite:

- if $\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T$ is nonsingular $\Rightarrow$ $\left.\frac{d^2J(\theta)}{d\theta^2}\right|_{\hat{\theta}}$ is positive definite $\Rightarrow$ the quadratic form is positive definite and it is a paraboloid with a unique minimum $\Rightarrow$ $\hat{\theta}$ is the global minimum of $J(\theta)$;
- if $\sum_{t=1}^{N}\varphi(t)\,\varphi(t)^T$ is singular $\Rightarrow$ the quadratic form is positive semidefinite and it has an infinite number of local minima, aligned over a line tangent to $J(\theta)$.

The obtained results may be rewritten in a compact matrix form by defining:

$$\Phi = \begin{bmatrix} \varphi(1)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix} = \begin{bmatrix} u_1(1) & \ldots & u_n(1) \\ \vdots & & \vdots \\ u_1(N) & \ldots & u_n(N) \end{bmatrix} \in \mathbb{R}^{N\times n}, \quad y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \in \mathbb{R}^N$$

$$\Downarrow$$

$$y(t) = \varphi(t)^T \theta, \quad t = 1, 2, \ldots, N \qquad \Leftrightarrow \qquad \boxed{\mathbf{y} = \boldsymbol{\Phi\theta}}$$

$$\Downarrow$$

$$\textstyle\sum_{t=1}^{N} \varphi(t)\, \varphi(t)^T = \Phi^T \Phi, \quad \sum_{t=1}^{N} \varphi(t)\, y(t) = \Phi^T y$$

$$\Downarrow$$

the normal equation system becomes:

$$\Phi^T \Phi \theta = \Phi^T y$$

and, if $\Phi^T \Phi$ is nonsingular (*identifiability condition*), it has a unique solution given by the least squares estimate:

$$\boxed{\hat{\boldsymbol{\theta}}_{\mathbf{LS}} = \left[\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right]^{-\mathbf{1}} \boldsymbol{\Phi}^T \mathbf{y}}$$

Proof:

$$\sum_{t=1}^{N} \varphi(t)\,\varphi(t)^T = \sum_{t=1}^{N} \begin{bmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{bmatrix} \begin{bmatrix} u_1(t) & \cdots & u_n(t) \end{bmatrix} = \sum_{t=1}^{N} \begin{bmatrix} u_1^2(t) & \cdots & u_1(t)\,u_n(t) \\ \vdots & \ddots & \vdots \\ u_n(t)\,u_1(t) & \cdots & u_n^2(t) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{t=1}^{N} u_1^2(t) & \cdots & \sum_{t=1}^{N} u_1(t)\,u_n(t) \\ \vdots & \ddots & \vdots \\ \sum_{t=1}^{N} u_n(t)\,u_1(t) & \cdots & \sum_{t=1}^{N} u_n^2(t) \end{bmatrix}$$

$$\Phi^T \Phi = \begin{bmatrix} \varphi(1) & \cdots & \varphi(N) \end{bmatrix} \begin{bmatrix} \varphi(1)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix} = \begin{bmatrix} u_1(1) & \dots & u_1(N) \\ \vdots & \ddots & \vdots \\ u_n(1) & \dots & u_n(N) \end{bmatrix} \begin{bmatrix} u_1(1) & \dots & u_n(1) \\ \vdots & \ddots & \vdots \\ u_1(N) & \dots & u_n(N) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{t=1}^{N} u_1^2(t) & \cdots & \sum_{t=1}^{N} u_1(t)\,u_n(t) \\ \vdots & \ddots & \vdots \\ \sum_{t=1}^{N} u_n(t)\,u_1(t) & \cdots & \sum_{t=1}^{N} u_n^2(t) \end{bmatrix} = \sum_{t=1}^{N} \varphi(t)\,\varphi(t)^T$$

$$\sum_{t=1}^{N} \varphi(t)\,y(t) = \sum_{t=1}^{N} \begin{bmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{bmatrix} y(t) = \begin{bmatrix} \sum_{t=1}^{N} u_1(t)\,y(t) \\ \vdots \\ \sum_{t=1}^{N} u_n(t)\,y(t) \end{bmatrix}$$

$$\Phi^T y = \begin{bmatrix} \varphi(1) & \cdots & \varphi(N) \end{bmatrix} \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} u_1(1) & \dots & u_1(N) \\ \vdots & \ddots & \vdots \\ u_n(1) & \dots & u_n(N) \end{bmatrix} \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{N} u_1(t)\,y(t) \\ \vdots \\ \sum_{t=1}^{N} u_n(t)\,y(t) \end{bmatrix}$$

Probabilistic characteristics of least squares estimator

Assumptions:

- the identifiability condition holds: $\exists \left[\Phi^T \Phi\right]^{-1}$;
- the random source of data has the following structure

$$y(t) = \varphi(t)^T \theta_o + v(t)\,, \quad t = 1, 2, \ldots, N$$

where $v(t)$ is a zero-mean random disturbance $\Rightarrow$

the relationship between y and $u_1, u_2, \ldots, u_n$ is assumed to be linear $\Rightarrow$

there exists a “true” value θ_o of the unknown variable;

in compact matrix form, it results that:

$$y = \Phi\theta_o + v$$

where $v = \begin{bmatrix} v(1) \\ \vdots \\ v(N) \end{bmatrix} \in \mathbb{R}^N$ is a vector random variable with $E\left[v\right] = \mathbf{0}$.

Under these assumptions, the least squares estimator becomes:

$$\begin{aligned}\hat{\theta} &= [\Phi^T\Phi]^{-1}\Phi^T y = [\Phi^T\Phi]^{-1}\Phi^T(\Phi\theta_o + v) = \\ &= [\Phi^T\Phi]^{-1}\Phi^T\Phi\theta_o + [\Phi^T\Phi]^{-1}\Phi^T v = \theta_o + [\Phi^T\Phi]^{-1}\Phi^T v\end{aligned}$$

and it has the following probabilistic characteristics:

- it is **unbiased**, since its mean value $E[\hat{\theta}\,] = \theta_o$

$$\begin{aligned}E[\hat{\theta}\,] &= E\Big[\big[\Phi^T\Phi\big]^{-1}\Phi^T y\Big] = \big[\Phi^T\Phi\big]^{-1}\Phi^T E[y] = \big[\Phi^T\Phi\big]^{-1}\Phi^T E[\Phi\theta_o + v] = \\ &= \big[\Phi^T\Phi\big]^{-1}\Phi^T(\Phi\theta_o + E[v]) = \big[\Phi^T\Phi\big]^{-1}\Phi^T\Phi\theta_o = \theta_o\end{aligned}$$

- if v is a vector of zero-mean random variables that are uncorrelated and with the same variance σ_v^2 ($Var[v] = E\left[vv^T\right] = \sigma_v^2 I_N$), as in the case of disturbance $v(\cdot)$ given by a white noise $WN(0,\sigma_v^2) \quad\Rightarrow\quad Var[\hat{\theta}\,] = \sigma_v^2[\Phi^T\Phi]^{-1}$

$$\begin{aligned}Var[\hat{\theta}\,] &= E\Big[(\hat{\theta} - E[\hat{\theta}])(\hat{\theta} - E[\hat{\theta}])^T\Big] = E\Big[(\hat{\theta} - \theta_o)(\hat{\theta} - \theta_o)^T\Big] = \\ &= E\Big[\big([\Phi^T\Phi]^{-1}\Phi^T v\big)\big([\Phi^T\Phi]^{-1}\Phi^T v\big)^T\Big] = E\Big[[\Phi^T\Phi]^{-1}\Phi^T vv^T\Phi[\Phi^T\Phi]^{-1}\Big] = \\ &= [\Phi^T\Phi]^{-1}\Phi^T E\left[vv^T\right]\Phi[\Phi^T\Phi]^{-1} = [\Phi^T\Phi]^{-1}\Phi^T\sigma_v^2 I_N\Phi[\Phi^T\Phi]^{-1} = \\ &= \sigma_v^2[\Phi^T\Phi]^{-1}\Phi^T\Phi[\Phi^T\Phi]^{-1} = \sigma_v^2[\Phi^T\Phi]^{-1}\end{aligned}$$

- The variance σ_v^2 of the disturbance v is usually unknown $\Rightarrow$
  under the same previous assumptions, a “reasonable” unbiased estimate $\hat{\sigma}_v^2$ (such that $E[\hat{\sigma}_v^2] = \sigma_v^2$) can be directly derived from data as

$$\hat{\sigma}_v^2 = \frac{J(\hat{\theta})}{N-n}$$

where $N =$ measurement number, $n =$ number of unknown parameters of θ,

$$\begin{aligned}
J(\hat{\theta}) &= \textstyle\sum_{t=1}^{N} \varepsilon(t)^2 \Big|_{\theta=\hat{\theta}} = \sum_{t=1}^{N} \left[y(t) - \varphi(t)^T\,\hat{\theta}\right]^2 = [y - \Phi\hat{\theta}]^T[y - \Phi\hat{\theta}] = \\
&= \left((I_N - \Phi[\Phi^T\Phi]^{-1}\Phi^T)y\right)^T (I_N - \Phi[\Phi^T\Phi]^{-1}\Phi^T)y = \\
&= y^T(I_N - \Phi[\Phi^T\Phi]^{-1}\Phi^T)(I_N - \Phi[\Phi^T\Phi]^{-1}\Phi^T)y = \\
&= y^T(I_N - 2\Phi[\Phi^T\Phi]^{-1}\Phi^T + \Phi[\Phi^T\Phi]^{-1}\Phi^T\Phi[\Phi^T\Phi]^{-1}\Phi^T)y = \\
&= y^T(I_N - \Phi[\Phi^T\Phi]^{-1}\Phi^T)y
\end{aligned}$$

$$\Downarrow$$

$$Var[\hat{\theta}\,] = \sigma_v^2[\Phi^T\Phi]^{-1} \cong \hat{\sigma}_v^2[\Phi^T\Phi]^{-1}$$

Weighted Least Squares estimation method

With the least squares estimation method, all the errors have the same relevance, since the figure of merit to be minimized is

$$J_{LS}(\theta)=\sum_{t=1}^{N}\varepsilon(t)^2, \quad \text{where} \quad \varepsilon(t)=y(t)-\varphi(t)^T\theta, \quad t=1,2,\ldots,N.$$

However, if some measurements are more accurate than some others, different relevance can be assigned to the errors, by defining the figure of merit

$$J_{WLS}(\theta)=\sum_{t=1}^{N}q(t)\,\varepsilon(t)^2=\varepsilon^T Q\varepsilon$$

where $q(t)>0$ are the weighting coefficients (or *weights*) for $t=1,2,\ldots,N$,

$$Q=\operatorname{diag}\left(q(t)\right)=\begin{bmatrix} q(1) & 0 & \ldots & 0\\ 0 & q(2) & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & \ldots & q(N)\end{bmatrix}\in\mathbb{R}^{N\times N}, \quad \varepsilon=\begin{bmatrix}\varepsilon(1)\\ \vdots\\ \varepsilon(N)\end{bmatrix}\in\mathbb{R}^N.$$

The **Weighted Least Squares (WLS) estimate** minimizes the figure of merit $J_{WLS}(\theta)$:

$$\boxed{\hat{\boldsymbol{\theta}}_{WLS} = \left[\boldsymbol{\Phi}^T \mathbf{Q} \boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^T \mathbf{Q} \mathbf{y}}$$

If the disturbance v is a vector of zero-mean uncorrelated random variables with variance Σ_v, the estimator $\hat{\theta}_{WLS}$ has the following probabilistic characteristics:

- it is **unbiased**, since its mean value $E[\hat{\theta}_{WLS}] = \theta_o$

$$E[\hat{\theta}_{WLS}] = E\left[\left[\Phi^T Q \Phi\right]^{-1} \Phi^T Q y\right] = \left[\Phi^T Q \Phi\right]^{-1} \Phi^T Q E[y] = \left[\Phi^T Q \Phi\right]^{-1} \Phi^T Q E[\Phi \theta_o + v] =$$
$$= \left[\Phi^T Q \Phi\right]^{-1} \Phi^T Q \left(\Phi \theta_o + E[v]\right) = \left[\Phi^T Q \Phi\right]^{-1} \Phi^T Q \Phi \theta_o = \theta_o$$

- its variance is

$$Var[\hat{\theta}_{WLS}] = E[(\hat{\theta}_{WLS} - E[\hat{\theta}_{WLS}])(\hat{\theta}_{WLS} - E[\hat{\theta}_{WLS}])^T] =$$
$$= E[(\hat{\theta}_{WLS} - \theta_o)(\hat{\theta}_{WLS} - \theta_o)^T] = E\left[\left([\Phi^T Q \Phi]^{-1} \Phi^T Q v\right)\left([\Phi^T Q \Phi]^{-1} \Phi^T Q v\right)^T\right] =$$
$$= E\left[[\Phi^T Q \Phi]^{-1} \Phi^T Q v v^T Q^T \Phi [\Phi^T Q \Phi]^{-1}\right] =$$
$$= [\Phi^T Q \Phi]^{-1} \Phi^T Q E\left[v v^T\right] Q \Phi [\Phi^T Q \Phi]^{-1} = [\Phi^T Q \Phi]^{-1} \Phi^T Q \Sigma_v Q \Phi [\Phi^T Q \Phi]^{-1}$$

and then it depends on the disturbance variance Σ_v;

- it can be proved that the best choice for Q that minimizes $Var[\hat{\theta}_{WLS}]$ is

$$Q^* = \arg\min_{Q=\text{diag}(q(t))\in\mathbb{R}^{N\times N}} Var[\hat{\theta}_{WLS}] = \Sigma_v^{-1}$$

and in this case we obtain the so-called **Gauss-Markov estimate**:

$$\boxed{\hat{\boldsymbol{\theta}}_{GM} = \left[\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_v^{-1}\boldsymbol{\Phi}\right]^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_v^{-1}\mathbf{y}}$$

whose variance is

$$\begin{aligned} Var[\hat{\theta}_{GM}] &= [\Phi^T Q\Phi]^{-1}\Phi^T Q\Sigma_v Q\Phi[\Phi^T Q\Phi]^{-1} = \\ &= [\Phi^T\Sigma_v^{-1}\Phi]^{-1}\Phi^T\Sigma_v^{-1}\Sigma_v Q\Phi[\Phi^T\Sigma_v^{-1}\Phi]^{-1} \\ &= [\Phi^T\Sigma_v^{-1}\Phi]^{-1}\ ; \end{aligned}$$

If in particular it results that $\Sigma_v = \sigma_v^2 I_N \quad\Rightarrow$

$$\hat{\theta}_{GM} = \left[\Phi^T\frac{1}{\sigma_v^2}I_N\Phi\right]^{-1}\Phi^T\frac{1}{\sigma_v^2}I_N y = \left[\Phi^T\Phi\right]^{-1}\Phi^T y = \hat{\theta}_{LS}$$

Maximum Likelihood estimators

The actual data are generated by a random source, which depends on the outcome s of a random experiment and on the “true” value θ_o of the unknown to be estimated. However, if a generic value θ of the unknown parameter is considered, the data can be seen as function of both the value θ and the outcome s $\Rightarrow$

the data can be denoted by $d^{(\theta)}(s)$, with p.d.f. $f(q, \theta)$ that is function of θ too.

Let δ be the particular data observation that corresponds to a particular outcome $\overline{s}$ of the random experiment:

$$\delta = d^{(\theta)}(\overline{s})$$

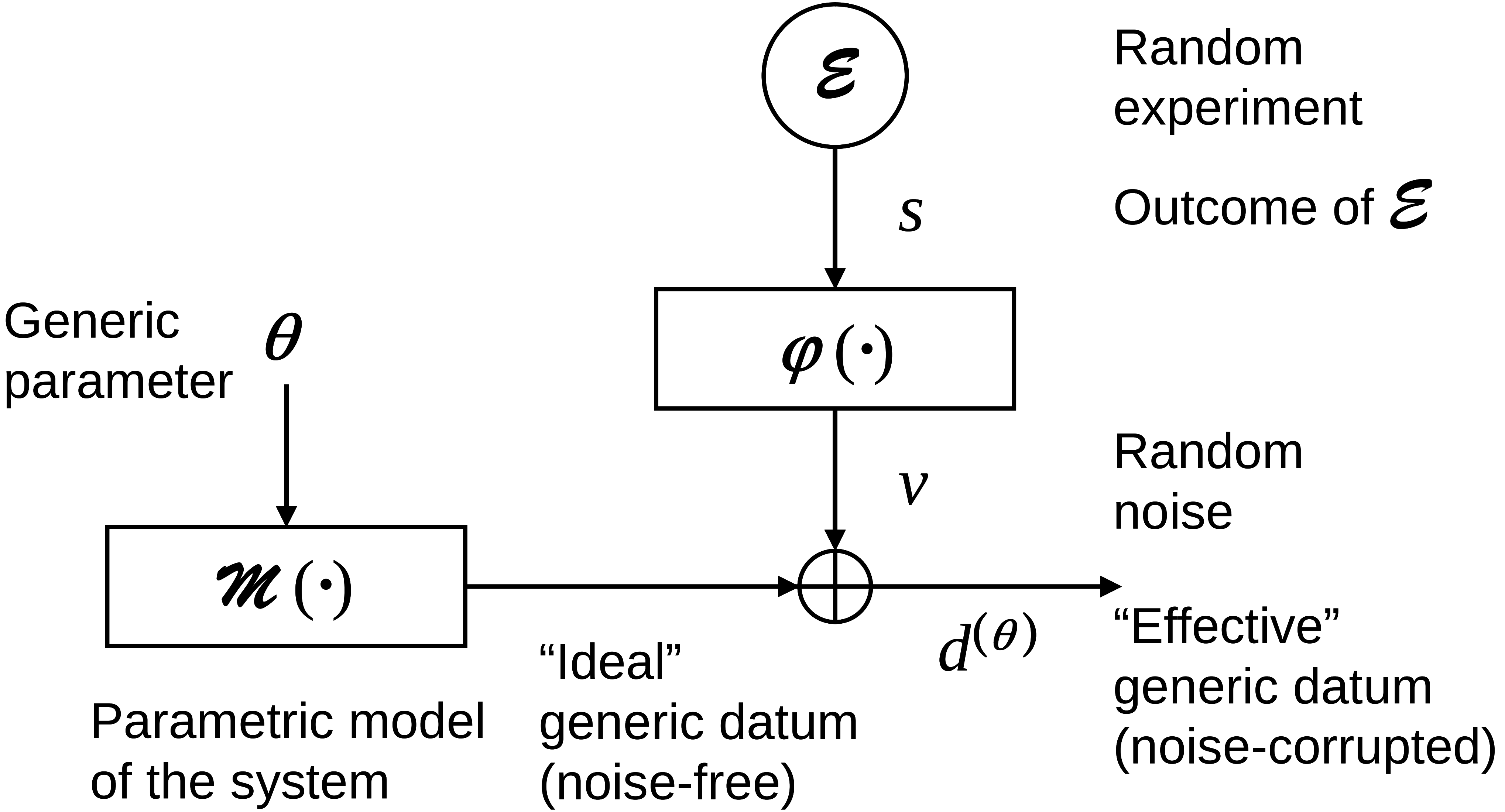
The so-called **likelihood function** is given by the p.d.f. of the data evaluated in δ:

$$L(\theta) = f(q, \theta)|_{q=\delta}$$

The **Maximum Likelihood (ML) estimate** is defined as:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \mathbb{R}^n} L(\theta)$$

Random source of data for a generic value θ of the unknown parameter:



Example: a scalar parameter $\theta_o \in \mathbb{R}$ is estimated using a unique measurement (i.e., $N = 1$), corrupted by a zero-mean Gaussian disturbance with variance σ_v^2 $\Rightarrow$ the random source of data has the following structure:

$$y = \theta_o + v$$

where the noise v is a scalar zero-mean Gaussian random variable with p.d.f.

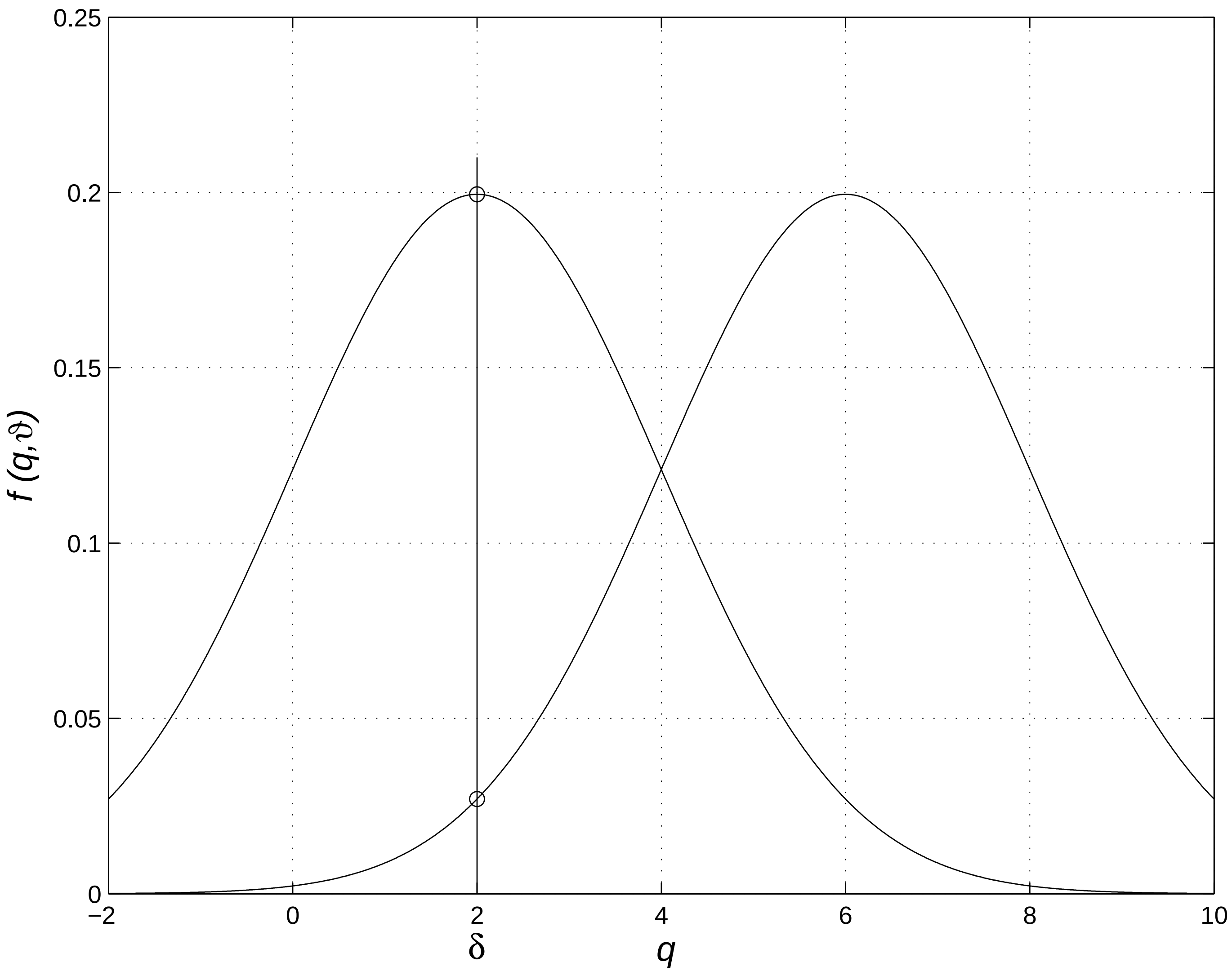
$$f(q) = \mathcal{N}\left(0, \sigma_v^2\right) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-q^2}{2\sigma_v^2}\right)$$

Since $v = y - \theta_o \Rightarrow$ the p.d.f. of data y generated by a random source where a generic value θ is considered instead of θ_o is then given by

$$f(q, \theta) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-(q-\theta)^2}{2\sigma_v^2}\right) = \mathcal{N}\left(\theta, \sigma_v^2\right) \quad \Rightarrow$$

$$L(\theta) = \left. f(q, \theta)\right|_{q=\delta} = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-(\delta-\theta)^2}{2\sigma_v^2}\right) = \mathcal{N}\left(\delta, \sigma_v^2\right)$$

$f(q, \theta)$ translates when the value of θ changes $\Rightarrow L(\theta) = \left. f(q, \theta)\right|_{q=\delta}$ varies too.



$$f(q,\theta) = \mathcal{N}\left(\theta, \sigma_v^2\right) \quad \Rightarrow \quad L(\theta) = f(q,\theta)\big|_{q=\delta} = \mathcal{N}\left(\delta, \sigma_v^2\right)$$

Maximum Likelihood estimator properties

The estimate $\hat{\theta}_{ML}$ is:

- asymptotically unbiased: $E\left(\hat{\theta}_{ML}\right) \xrightarrow[N \to \infty]{} \theta_o$
- asymptotically efficient: $\Sigma_{\hat{\theta}_{ML}} \leq \Sigma_{\hat{\theta}} \quad \forall \hat{\theta}$ if $N \to \infty$
- consistent: $\lim\limits_{N \to \infty} \Sigma_{\hat{\theta}_{ML}} = 0$
- asymptotically Gaussian (for $N \to \infty$)

Example: let us assume that the random source of data has the following structure:

$$y(t) = \psi(t, \theta_o) + v(t)\,, \quad t = 1, 2, \ldots, N \quad \Leftrightarrow \quad y = \Psi(\theta_o) + v$$

where $\psi(t, \theta_o)$ is a generic *nonlinear* function of θ_o and the disturbance v is a vector of zero-mean Gaussian random variables with variance Σ_v and p.d.f.

$$f(q) = \mathcal{N}(0, \Sigma_v) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2} q^T \Sigma_v^{-1} q\right)$$

Since $v = y - \Psi(\theta_o) \Rightarrow$ the p.d.f. of data generated by a random source where a generic value θ is considered instead of θ_o is then given by

$$f(q, \theta) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2} \left[q - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[q - \Psi(\theta)\right]\right)$$

$$\Downarrow$$

$$L(\theta) = f(q, \theta)\big|_{q=\delta} = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2} \left[\delta - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[\delta - \Psi(\theta)\right]\right)$$

$$L(\theta) = f(q,\theta)\big|_{q=\delta} = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\frac{1}{2}\left[\delta - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[\delta - \Psi(\theta)\right]\right)$$

$$\Downarrow$$

$f(q,\theta)\big|_{q=\delta}$ is an exponential function of θ

$$\Downarrow$$

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\mathbb{R}^n} L(\theta) = \arg\min_{\theta\in\mathbb{R}^n} \underbrace{\left\{\left[\delta - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[\delta - \Psi(\theta)\right]\right\}}_{R(\theta)}$$

Problem: the global minimum of $R(\theta)$ has to be found with respect to θ, but $R(\theta)$ may have many local minima if $\Psi(\theta)$ is a generic nonlinear function of the unknown variable; the standard nonlinear optimization algorithms do not guarantee to find always the global minimum.

Particular case: $\Psi(\theta) =$ *linear* function of the unknown parameters $= \Phi\theta$

$$\Downarrow$$

$R(\theta)$ is a quadratic function of θ : $R(\theta) = [\delta - \Phi\theta]^T \Sigma_v^{-1} [\delta - \Phi\theta]$

$$\Downarrow$$

there exists a unique minimum of $R(\theta)$, if $\det\left(\Phi^T \Sigma_v^{-1} \Phi\right) \neq 0$

$$\Downarrow$$

$$\hat{\theta}_{ML} = \left(\Phi^T \Sigma_v^{-1} \Phi\right)^{-1} \Phi^T \Sigma_v^{-1} \delta = \textbf{Gauss-Markov estimate} = \hat{\theta}_{GM} =$$

$=$ *Weighted Least Squares estimate* using the disturbance variance Σ_v

If $\Sigma_v = \sigma_v^2 I_N$, i.e., independent identically distributed (*i.i.d.)* disturbance:

$$\hat{\theta}_{ML} = \hat{\theta}_{GM} = \left(\Phi^T \Phi\right)^{-1} \Phi^T \delta = \textbf{Least Squares estimate}$$

Gauss-Markov estimate properties

The estimate $\hat{\theta}_{GM}$ is:

- unbiased: $E\left(\hat{\theta}_{GM}\right) = \theta_o$
- efficient: $\Sigma_{\hat{\theta}_{GM}} \leq \Sigma_{\hat{\theta}} \quad \forall \hat{\theta}$
- consistent: $\lim\limits_{N \to \infty} \Sigma_{\hat{\theta}_{GM}} = 0$
- Gaussian

Bayesian estimation method

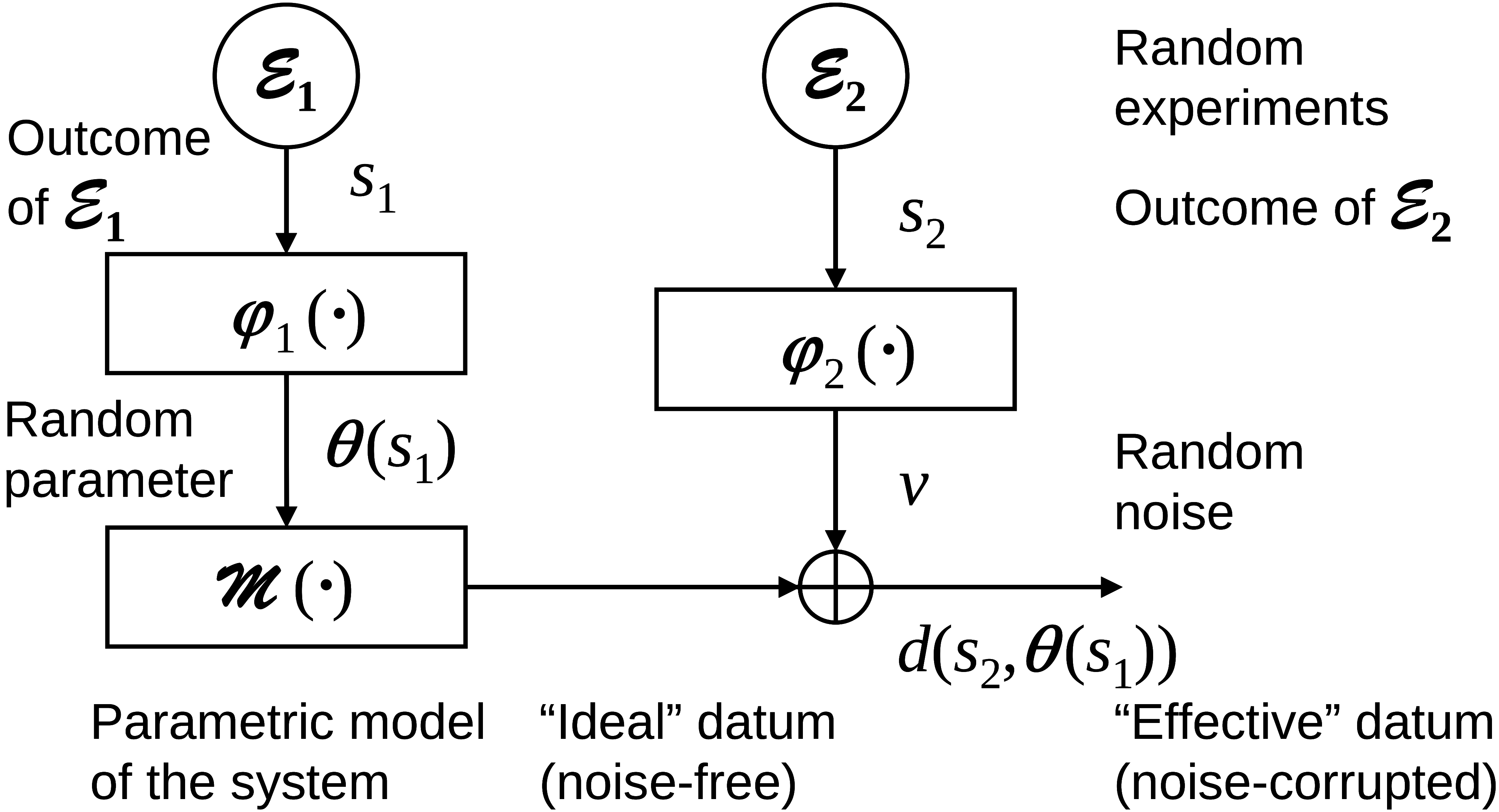
The Bayesian method allows one to take into account experimental data and *a priori* information on the unknown of the estimation problem that, if well exploited, can improve the estimate and make up for possible random errors corrupting the data:

- the unknown θ is considered as a random variable, whose a priori p.d.f. (i.e., in absence of data) has some given behaviour, mean value and variance

  $$\Downarrow$$

  the mean value is a possible initial estimate of θ, while the variance represents the a priori uncertainty;
- as new experimental data arrive, the p.d.f. of θ is updated on the basis of the new information: the mean value changes with respect to the a priori one, while the variance is expected to decrease thanks to the information provided by data.

Random source of data with a random unknown parameter θ:



A joint random experiment $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2$ is assumed to exist, whose joint outcome s is the couple of single outcomes s_1 and s_2: $s = (s_1, s_2)$:

- the unknown θ is generated by a first random source $\mathcal{S}_1$ on the basis of the outcome s_1 of the first random experiment $\mathcal{E}_1 \quad \Rightarrow \quad \theta = \theta(s_1)$;
- the data d are generated by the second random source $\mathcal{S}_2$, influenced by
  - the outcome s_2 of the second random experiment $\mathcal{E}_2$
  - the value $\theta(s_1)$ of the unknown to be estimated

$$d = d(s_2, \theta(s_1))$$

A generic estimator is a function of data $\hat{\theta} = h(d)$ and its performances improve as much as the estimate $\hat{\theta}$ is closer to the unknown to be estimated

$$\Downarrow$$

by considering as figure of merit the mean squared error (MSE)

$$J(h(\cdot)) = E[\|\theta - h(d)\|^2]$$

the Bayesian optimal estimator is the particular function $h^*(\cdot)$ such that

$$E[\|\theta - h^*(d)\|^2] \leq E[\|\theta - h(d)\|^2], \quad \forall h(\cdot)$$

It can be proved that such an optimal estimator exists and it is given by:

$$h^*(x) = E\left[\theta | d = x\right]$$

where x is the current value that the data d may take.

The **Bayesian estimator** (or **conditional mean estimator**) is the function

$$\boxed{\hat{\boldsymbol{\theta}} = \mathbf{E}\left[\boldsymbol{\theta} | \mathbf{d}\right]}$$

and the **Bayesian estimate** (or **conditional mean estimate**) is the numeric value

$$\hat{\theta} = E\left[\theta | d = \delta\right]$$

where δ is the value of the data d that corresponds to a particular outcome of the joint random experiment $\mathcal{E}$.

Bayesian estimator in the Gaussian case

Assumption: the data d and the unknown θ are scalar random variables with zero mean value and both are individually and jointly Gaussian:

$$\begin{bmatrix} d \\ \theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = Var\begin{bmatrix} d \\ \theta \end{bmatrix} = \begin{bmatrix} \sigma_{dd} & \sigma_{d\theta} \\ \sigma_{\theta d} & \sigma_{\theta\theta} \end{bmatrix}\right) \Rightarrow \text{their joint p.d.f. is given by:}$$

$$f(d,\theta) = C \exp\left\{-\frac{1}{2}\left[\, d \quad \theta \,\right] \Sigma^{-1} \left[\, d \quad \theta \,\right]^T\right\}, \quad C : \text{suitable constant}$$

Since

$$\det\Sigma = \det\begin{bmatrix} \sigma_{dd} & \sigma_{d\theta} \\ \sigma_{\theta d} & \sigma_{\theta\theta} \end{bmatrix} = \sigma_{dd}\sigma_{\theta\theta} - \sigma_{d\theta}^2 = \sigma_{dd}\left(\sigma_{\theta\theta} - \frac{\sigma_{\theta d}^2}{\sigma_{dd}}\right) = \sigma_{dd}\,\sigma^2,$$

$$\text{where} \quad \sigma^2 = \sigma_{\theta\theta} - \sigma_{\theta d}^2/\sigma_{dd} \leq \sigma_{\theta\theta}$$

$$\Downarrow$$

$$\Sigma^{-1} = \frac{1}{\det\Sigma}\begin{bmatrix} \sigma_{\theta\theta} & -\sigma_{d\theta} \\ -\sigma_{\theta d} & \sigma_{dd} \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} \sigma_{\theta\theta}/\sigma_{dd} & -\sigma_{d\theta}/\sigma_{dd} \\ -\sigma_{\theta d}/\sigma_{dd} & 1 \end{bmatrix}$$

$$\Downarrow$$

$$f(d,\theta) = C\exp\left\{-\frac{1}{2\sigma^2}\begin{bmatrix} d & \theta \end{bmatrix}\begin{bmatrix} \sigma_{\theta\theta}/\sigma_{dd} & -\sigma_{d\theta}/\sigma_{dd} \\ -\sigma_{\theta d}/\sigma_{dd} & 1 \end{bmatrix}\begin{bmatrix} d \\ \theta \end{bmatrix}\right\} =$$

$$= C\exp\left\{-\frac{1}{2\sigma^2}\begin{bmatrix} d & \theta \end{bmatrix}\begin{bmatrix} \sigma_{\theta\theta}/\sigma_{dd}\, d - \sigma_{d\theta}/\sigma_{dd}\,\theta \\ -\sigma_{\theta d}/\sigma_{dd}\, d + \theta \end{bmatrix}\right\} =$$

$$= C\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{\sigma_{\theta\theta}}{\sigma_{dd}}d^2 - 2\frac{\sigma_{\theta d}}{\sigma_{dd}}d\theta + \theta^2\right)\right\}$$

The p.d.f. of the data d is given by:

$$f(d) = C'\exp\left\{-\frac{d^2}{2\sigma_{dd}}\right\}, \quad C' \text{ : suitable constant}$$

$$\Downarrow$$

the p.d.f. of the unknown θ conditioned by data d is equal to:

$$f(\theta|d) = \frac{f(d,\theta)}{f(d)} = \frac{C}{C'}\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{\sigma_{\theta\theta}}{\sigma_{dd}}d^2 - 2\frac{\sigma_{\theta d}}{\sigma_{dd}}d\theta + \theta^2\right) + \frac{d^2}{2\sigma_{dd}}\right\} =$$

$$= C''\exp\left\{-\frac{1}{2\sigma^2}\left[\frac{\sigma_{d\theta}^2}{\sigma_{dd}^2}d^2 - 2\frac{\sigma_{\theta d}}{\sigma_{dd}}d\theta + \theta^2\right]\right\} = C''\exp\left\{-\frac{1}{2\sigma^2}\left[\theta - \frac{\sigma_{\theta d}}{\sigma_{dd}}d\right]^2\right\}$$

$$\Downarrow$$

$$f(\theta|d) = C''\exp\left\{-\frac{1}{2\sigma^2}\left[\theta - \frac{\sigma_{\theta d}}{\sigma_{dd}}d\right]^2\right\} \sim \mathcal{N}\left(\frac{\sigma_{\theta d}}{\sigma_{dd}}d,\ \sigma^2\right)$$

The Bayesian estimator is the function

$$\boxed{\hat{\theta} = E\left[\theta|\, d\right] = \frac{\boldsymbol{\sigma}_{\theta d}}{\boldsymbol{\sigma}_{dd}}d}$$

while the Bayesian estimate corresponding to the particular observation δ of data d is the numerical value

$$\hat{\theta} = E\left[\theta|\, d = \delta\right] = \frac{\sigma_{\theta d}}{\sigma_{dd}}\delta$$

Since $E\left[d\right] = E\left[\theta\right] = 0 \Rightarrow$

$$E[\hat{\theta}] = E\left[\frac{\sigma_{\theta d}}{\sigma_{dd}}d\right] = \frac{\sigma_{\theta d}}{\sigma_{dd}}E\left[d\right] = 0$$

$$Var[\hat{\theta}\,] = E[(\hat{\theta} - E[\hat{\theta}\,])^2] = E[\hat{\theta}^2] = E\left[\frac{\sigma_{\theta d}^2}{\sigma_{dd}^2}d^2\right] = \frac{\sigma_{\theta d}^2}{\sigma_{dd}^2}E\left[d^2\right] = \frac{\sigma_{\theta d}^2}{\sigma_{dd}}$$

$$J(\hat{\theta}) = Var[\theta - \hat{\theta}\,] = E[(\theta - \hat{\theta})^2] = E[(\theta - \frac{\sigma_{\theta d}}{\sigma_{dd}}d)^2] = E[\theta^2 - 2\frac{\sigma_{\theta d}}{\sigma_{dd}}\theta d + \frac{\sigma_{\theta d}^2}{\sigma_{dd}^2}d^2] =$$

$$= E\left[\theta^2\right] - 2\frac{\sigma_{\theta d}}{\sigma_{dd}}E[\theta d] + \frac{\sigma_{\theta d}^2}{\sigma_{dd}^2}E\left[d^2\right] = \sigma_{\theta\theta} - 2\frac{\sigma_{\theta d}}{\sigma_{dd}}\sigma_{\theta d} + \frac{\sigma_{\theta d}^2}{\sigma_{dd}^2}\sigma_{dd} =$$

$$= \sigma_{\theta\theta} - 2\frac{\sigma_{\theta d}^2}{\sigma_{dd}} + \frac{\sigma_{\theta d}^2}{\sigma_{dd}} = \sigma_{\theta\theta} - \frac{\sigma_{\theta d}^2}{\sigma_{dd}} = \sigma^2$$

Optimal linear estimator

Assumption: both the data d and the unknown θ are scalar random variables with zero mean value and variance matrix $Var\begin{bmatrix} d \\ \theta \end{bmatrix} = \begin{bmatrix} \sigma_{dd} & \sigma_{d\theta} \\ \sigma_{\theta d} & \sigma_{\theta\theta} \end{bmatrix}$.

Goal: estimate θ by means of a linear estimator whose structure is

$$\hat{\theta} = \alpha d + \beta$$

with α, β real parameters, estimated by minimizing the mean squared error (MSE):

$$J = E[(\theta - \hat{\theta})^2] = E[(\theta - \alpha d - \beta)^2] = J(\alpha, \beta)$$

$$\Updownarrow \quad \text{gradient}\, J(\alpha,\beta)=\mathbf{0},\ \text{Hessian}\, J(\alpha,\beta)\geq 0$$

$$\frac{\partial J}{\partial \alpha} = \frac{\partial}{\partial \alpha} E\left[(\theta - \alpha d - \beta)^2\right] = E\left[\frac{\partial}{\partial \alpha}(\theta - \alpha d - \beta)^2\right] = E\left[-2(\theta - \alpha d - \beta)d\right] =$$
$$= -2E\left[\theta d\right] + 2\alpha E\left[d^2\right] + 2\beta E\left[d\right] = -2\sigma_{d\theta} + 2\alpha\sigma_{dd} = 0$$
$$\frac{\partial J}{\partial \beta} = \frac{\partial}{\partial \beta} E\left[(\theta - \alpha d - \beta)^2\right] = E\left[\frac{\partial}{\partial \beta}(\theta - \alpha d - \beta)^2\right] = E\left[-2(\theta - \alpha d - \beta)\right] =$$
$$= -2E\left[\theta\right] + 2\alpha E\left[d\right] + 2\beta = 2\beta = 0$$

$$\Updownarrow$$

$$\begin{cases} \alpha = \sigma_{\theta d} / \sigma_{dd} \\ \beta = 0 \end{cases} \quad \Rightarrow \quad \hat{\theta} = \frac{\sigma_{d\theta}}{\sigma_{dd}} d \equiv E\left[\theta | \, d\right]$$

Generalizations

- If the data d and the unknown θ are scalar random variables with nonzero mean value ($E[d] = \bar{d} \in \mathbb{R}$, $E[\theta] = \bar{\theta} \in \mathbb{R}$) and variance matrix $Var\begin{bmatrix} d \\ \theta \end{bmatrix} = \begin{bmatrix} \sigma_{dd} & \sigma_{d\theta} \\ \sigma_{\theta d} & \sigma_{\theta\theta} \end{bmatrix}$, the Bayesian estimator and the optimal linear estimator are given by:

$$\hat{\theta} = \bar{\theta} + \frac{\sigma_{\theta d}}{\sigma_{dd}} \left(d - \bar{d}\right)$$

$$J(\hat{\theta}) = Var[\theta - \hat{\theta}\,] = E[(\theta - \hat{\theta})^2] = \sigma_{\theta\theta} - \frac{\sigma_{\theta d}^2}{\sigma_{dd}} = \sigma^2$$

- If the data d and the unknown θ are vector random variables with nonzero mean value ($E[d] = \bar{d} \in \mathbb{R}^N$, $E[\theta] = \bar{\theta} \in \mathbb{R}^n$) and variance matrix $Var\begin{bmatrix} d \\ \theta \end{bmatrix} = \begin{bmatrix} \Sigma_{dd} & \Sigma_{d\theta} \\ \Sigma_{\theta d} & \Sigma_{\theta\theta} \end{bmatrix}$, the Bayesian estimator and the optimal linear estimator are given by:

$$\hat{\theta} = \bar{\theta} + \Sigma_{\theta d} \Sigma_{dd}^{-1} \left(d - \bar{d}\right)$$

$$Var[\theta - \hat{\theta}\,] = E[(\theta - \hat{\theta})(\theta - \hat{\theta})^T] = \Sigma_{\theta\theta} - \Sigma_{\theta d} \Sigma_{dd}^{-1} \Sigma_{d\theta}$$

Proof for the scalar case.

If $E[d]=\bar{d}\neq 0\in\mathbb{R}$ and/or $E[\theta]=\bar{\theta}\neq 0\in\mathbb{R}$, then define the random variables

$$d' = d-\bar{d} \quad\Rightarrow\quad E[d'] = E\left[d-\bar{d}\right] = E[d]-\bar{d}=0$$

$$\theta' = \theta-\bar{\theta} \quad\Rightarrow\quad E\left[\theta'\right] = E\left[\theta-\bar{\theta}\right] = E[\theta]-\bar{\theta}=0$$

The Bayesian estimate $\widehat{\theta'}$ of θ' based on d' is given by:

$$\widehat{\theta'}=E\left[\theta'\middle|\, d'\right]=E\left[\theta-\bar{\theta}\middle|\, d'\right]=E[\theta|\, d']-\bar{\theta}=E[\theta|\, d]-\bar{\theta}=\hat{\theta}-\bar{\theta}=\frac{\sigma_{\theta'd'}}{\sigma_{d'd'}}d'$$

where

$$\sigma_{\theta'd'} = E[(\theta'-E\left[\theta'\right])(d'-E[d'])] = E[\theta' d'] = E[(\theta-\bar{\theta})(d-\bar{d})] = \sigma_{\theta d}$$

$$\sigma_{d'd'} = E[(d'-E[d'])^2] = E[(d')^2] = E[(d-\bar{d})^2] = \sigma_{dd}$$

and then:

$$\widehat{\theta'} = \frac{\sigma_{\theta'd'}}{\sigma_{d'd'}}d' = \frac{\sigma_{\theta d}}{\sigma_{dd}}(d-\bar{d}) = \hat{\theta}-\bar{\theta} \quad\Rightarrow\quad \hat{\theta}=\bar{\theta}+\frac{\sigma_{\theta d}}{\sigma_{dd}}(d-\bar{d})$$

Remarks

Remark #1:

- Using the *a priori* information only (i.e., in absence of data), a reasonable initial estimate of the unknown is given by the **a priori estimate**

$$\hat{\theta} = \hat{\theta}^{prior} = E[\theta] = \bar{\theta}$$

and the corresponding **a priori uncertainty** is $Var[\theta] = \Sigma_{\theta\theta}$

- Using also the *a posteriori* information (i.e., the data), the estimate changes and the **a posteriori estimate** in the scalar case is given by

$$\hat{\theta} = \hat{\theta}^{posterior} = \bar{\theta} + \frac{\sigma_{\theta d}}{\sigma_{dd}}\left(d - \bar{d}\right) = \hat{\theta}^{prior} + \frac{\sigma_{\theta d}}{\sigma_{dd}}\left(d - \bar{d}\right)$$

  - if $\sigma_{\theta d} = 0$, i.e., if d and θ are uncorrelated $\Rightarrow \hat{\theta}^{posterior} = \hat{\theta}^{prior}$
  - if $\sigma_{\theta d} > 0 \Rightarrow \hat{\theta}^{posterior} - \hat{\theta}^{prior}$ and $d - \bar{d}$ have the same sign
  - if $\sigma_{\theta d} < 0 \Rightarrow \hat{\theta}^{posterior} - \hat{\theta}^{prior}$ and $d - \bar{d}$ have opposite sign

Remark #2: the a posteriori estimate in the scalar case is given by

$$\hat{\theta} = \hat{\theta}^{posterior} = \bar{\theta} + \frac{\sigma_{\theta d}}{\sigma_{dd}}\left(d - \bar{d}\right) = \hat{\theta}^{prior} + \frac{\sigma_{\theta d}}{\sigma_{dd}}\left(d - \bar{d}\right)$$

- if σ_{dd} is high, i.e., if the observation d is affected by great uncertainty $\Rightarrow$ $\hat{\theta}$ mainly depends on $\hat{\theta}^{prior}$ instead on the term $\frac{\sigma_{\theta d}}{\sigma_{dd}}\left(d - \bar{d}\right)$
- if σ_{dd} is low, i.e., if the observation d is affected by small uncertainty $\Rightarrow$ $\hat{\theta}$ strongly depends on the term $\frac{\sigma_{\theta d}}{\sigma_{dd}}\left(d - \bar{d}\right)$ that corrects $\hat{\theta}^{prior}$

Remark #3: the estimation error variance represents the **a posteriori uncertainty**:

$$J(\hat{\theta}) = Var[\theta - \hat{\theta}\,] = E[(\theta - \hat{\theta})^2] = \sigma_{\theta\theta} - \frac{\sigma_{\theta d}^2}{\sigma_{dd}} = \sigma_{\theta\theta}\left(1 - \frac{\sigma_{\theta d}^2}{\sigma_{\theta\theta}\sigma_{dd}}\right) = \sigma_{\theta\theta}\left(1 - \rho^2\right)$$

where $\rho = \frac{\sigma_{\theta d}}{\sqrt{\sigma_{\theta\theta}\sigma_{dd}}}$ is the correlation coefficient between θ and d, such that $|\rho| \leq 1$

- if $\rho = 0$, i.e., if d and θ are uncorrelated $\Rightarrow$ the a posteriori uncertainty turns out to be equal to the a priori one
- if $\rho \neq 0$ $\Rightarrow$ the a posteriori uncertainty is smaller than the a priori one

Geometrical interpretation

- Let $\mathbb{G}$ be the set of the real scalar random variables v with zero mean value, whose value $v(s)$ depends on the outcome s of the underlying random experiment $\mathcal{E}$.
- Let $\mathcal{G}$ be the vector space defined on $\mathbb{G}$ such that, $\forall v_1, v_2 \in \mathbb{G}$ and $\forall \mu \in \mathbb{R}$, then $v_1 + v_2 \in \mathbb{G}$ and $\mu v_1 \in \mathbb{G}$; let $\mathcal{G}$ be equipped with the inner (or scalar) product:

$$\langle v_1, v_2 \rangle = E\left[v_1 v_2\right]$$

that satisfies the following properties, $\forall v, v_1, v_2 \in \mathbb{G}$ and $\forall \mu \in \mathbb{R}$:

$$\left.\begin{array}{rl} (i) & \langle v, v \rangle = Var[v] \geq 0 \text{ (nonnegativity)} \\ (ii) & \langle v, v \rangle = 0 \text{ if and only if } v \sim (0,0) \end{array}\right\} \text{ (positive-definiteness)}$$

$$\begin{array}{rl} (iii) & \langle v, v_1 + v_2 \rangle = \langle v, v_1 \rangle + \langle v, v_2 \rangle \text{ (additivity)} \\ (iv) & \langle v_1, \mu v_2 \rangle = \mu \langle v_1, v_2 \rangle \text{ (homogeneity)} \\ (v) & \langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle \text{ (symmetry)} \end{array}$$

Such an inner product allows to naturally define a norm on $\mathcal{G}$ as:

$$\|v\| = \sqrt{\langle v, v \rangle} = \sqrt{Var[v]}$$

- Any random variable v is a vector in the space $\mathcal{G}$ with "length" $\|v\| = \sqrt{Var[v]}$
- Given two random variables v_1 and v_2, the angle α between the corresponding vectors in $\mathcal{G}$ is involved in the inner product, since:

$$\langle v_1, v_2 \rangle = \|v_1\| \; \|v_2\| \; \cos\alpha$$

$$\Downarrow$$

$$\cos\alpha = \frac{\langle v_1, v_2 \rangle}{\|v_1\| \; \|v_2\|} = \frac{E\left[v_1 v_2\right]}{\sqrt{Var[v_1]}\,\sqrt{Var[v_2]}} = \rho$$

  - $\rho = 0 \Leftrightarrow v_1$ and v_2 are uncorrelated $\Leftrightarrow$ the corresponding vectors in $\mathcal{G}$ are orthogonal, i.e., $v_1 \perp v_2$
  - $\rho = \pm 1 \Leftrightarrow$ the vectors corresponding to v_1 and v_2 are parallel, i.e., $v_1 // v_2$:
    if $v_2 = \alpha v_1 + \beta$, with $\alpha, \beta \in \mathbb{R}$ and $\alpha > 0$, then $\rho = +1$
    if $v_2 = \alpha v_1 + \beta$, with $\alpha, \beta \in \mathbb{R}$ and $\alpha < 0$, then $\rho = -1$

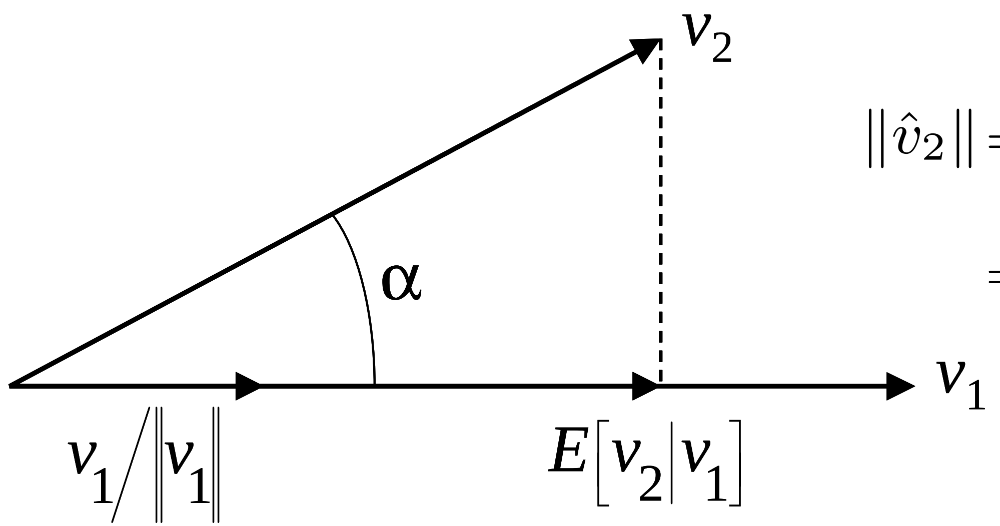
- In the scalar Gaussian case, the Bayesian estimate of v_2 based on v_1 is:

$$\hat{v}_2 = E\left[v_2 \middle| v_1\right] = \frac{\sigma_{21}}{\sigma_{11}} v_1, \quad \text{where } \sigma_{21} = E\left[v_1 v_2\right], \sigma_{11} = Var[v_1]$$

$$\Downarrow$$

$$\hat{v}_2 = \frac{E[v_1 v_2]}{Var[v_1]} v_1 = \frac{\langle v_1, v_2 \rangle}{\|v_1\|^2} v_1 = \frac{1}{\|v_1\|} \underbrace{\frac{\langle v_1, v_2 \rangle}{\|v_1\|} \frac{1}{\|v_2\|}}_{\cos\alpha} \|v_2\| v_1 = \underbrace{\|v_2\| \cos\alpha}_{\|\hat{v}_2\|} \frac{v_1}{\|v_1\|}$$

the Bayesian estimate $\hat{v}_2$ has the same direction of v_1 with “length” $\|v_2\| \cos\alpha$, i.e., $\hat{v}_2$ is the orthogonal projection of v_2 over v_1



$$\|\hat{v}_2\| = \left\| \frac{\sigma_{21}}{\sigma_{11}} v_1 \right\| = \frac{\sigma_{21}}{\sigma_{11}} \|v_1\|$$

$$= \frac{\sigma_{21}}{\sigma_{11}} \sqrt{\sigma_{11}} = \frac{\sigma_{21}}{\sqrt{\sigma_{11}}}$$

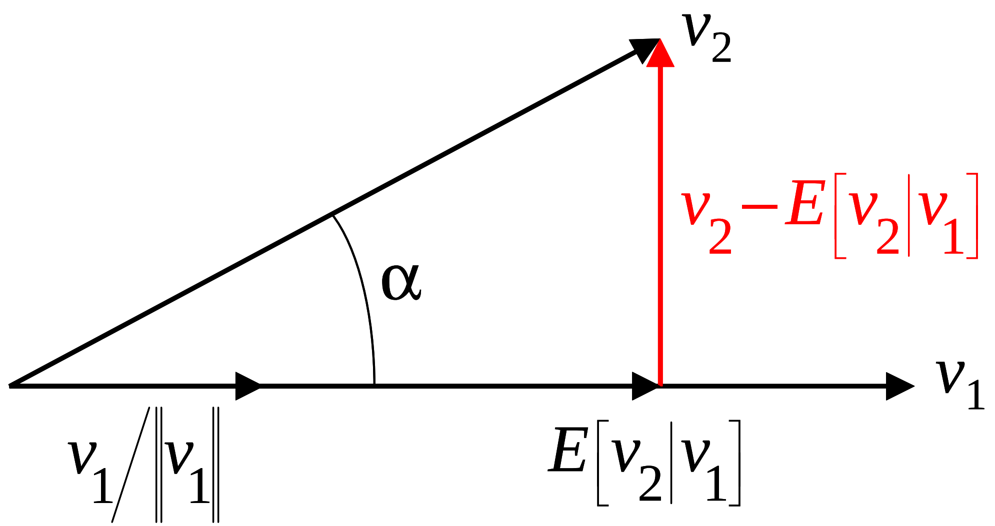
- The estimation error variance of v_2 given v_1 (i.e., the a posteriori uncertainty) is:

$$Var[v_2 - E[v_2|v_1]] = \sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}, \text{ with } \sigma_{22} = Var[v_2], \sigma_{21} = E[v_1 v_2], \sigma_{11} = Var[v_1]$$

$$\Downarrow$$

$$Var[v_2 - E[v_2|v_1]] = Var[v_2] - \frac{E[v_1 v_2]^2}{Var[v_1]} = \|v_2\|^2 - \|E[v_2|v_1]\|^2 = \|v_2 - E[v_2|v_1]\|^2$$

i.e., it can be computed by evaluating the "length" of the vector $v_2 - E[v_2|v_1]$ through the Pythagorean theorem



- The generalization of the geometric interpretation to the vector case is straightforward

Recursive Bayesian estimation: scalar case

Assumptions: the unknown θ is a scalar random variable with zero mean value; the data vector d is a random variable having 2 components $d(1), d(2)$, with zero mean value:

$$\underbrace{\begin{bmatrix} \theta \\ d(1) \\ d(2) \end{bmatrix}}_{d\ \text{for}\ d(1),d(2)} \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = Var \begin{bmatrix} \theta \\ d(1) \\ d(2) \end{bmatrix} = \begin{bmatrix} \sigma_{\theta\theta} & \overbrace{\sigma_{\theta 1} \quad \sigma_{\theta 2}}^{\Sigma_{\theta d}} \\ \underbrace{\begin{matrix} \sigma_{1\theta} \\ \sigma_{2\theta} \end{matrix}}_{\Sigma_{d\theta} = \Sigma_{\theta d}^T} & \underbrace{\begin{matrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{matrix}}_{\Sigma_{dd}} \end{bmatrix} \right), \qquad \begin{cases} \sigma_{\theta 1} = \sigma_{1\theta} \\ \sigma_{\theta 2} = \sigma_{2\theta} \\ \sigma_{12} = \sigma_{21} \end{cases}$$

- The optimal linear estimate of θ based on $d(1)$ only is given by:

$$E\left[\theta \mid d(1)\right] = \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1)$$

- The optimal linear estimate of θ based on $d(1)$ and $d(2)$ is given by:

$$E\left[\theta \mid d(1), d(2)\right] = \Sigma_{\theta d} \Sigma_{dd}^{-1} d = \begin{bmatrix} \sigma_{\theta 1} & \sigma_{\theta 2} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} d(1) \\ d(2) \end{bmatrix}$$

Since

$$\det \Sigma_{dd} = \det \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \sigma_{11}\sigma_{22} - \sigma_{21}^2 = \sigma_{11}\left(\sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}\right) = \sigma_{11}\ \sigma^2,$$

$$\text{where} \quad \sigma^2 = \sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}$$

$$\Downarrow$$

$$\Sigma_{dd}^{-1} = \frac{1}{\det \Sigma_{dd}} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} \sigma_{22}/\sigma_{11} & -\sigma_{12}/\sigma_{11} \\ -\sigma_{21}/\sigma_{11} & 1 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{aligned} E\left[\theta|\, d(1), d(2)\right] &= \Sigma_{\theta d}\Sigma_{dd}^{-1} d = \left[\ \sigma_{\theta 1} \quad \sigma_{\theta 2}\ \right] \frac{1}{\sigma^2} \begin{bmatrix} \sigma_{22}/\sigma_{11} & -\sigma_{12}/\sigma_{11} \\ -\sigma_{21}/\sigma_{11} & 1 \end{bmatrix} \begin{bmatrix} d(1) \\ d(2) \end{bmatrix} = \\ &= \frac{1}{\sigma^2} \left[\sigma_{\theta 1}\frac{\sigma_{22}}{\sigma_{11}} - \sigma_{\theta 2}\frac{\sigma_{21}}{\sigma_{11}} \quad \sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}} \right] \begin{bmatrix} d(1) \\ d(2) \end{bmatrix} = \\ &= \frac{1}{\sigma^2}\left(\sigma_{\theta 1}\frac{\sigma_{22}}{\sigma_{11}} - \sigma_{\theta 2}\frac{\sigma_{21}}{\sigma_{11}}\right) d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}}\right) d(2) \end{aligned}$$

By adding and subtracting the term $E\left[\theta|\, d(1)\right] = \dfrac{\sigma_{\theta 1}}{\sigma_{11}} d(1)$ and recalling that $\sigma_{12} = \sigma_{21}$ and $\sigma^2 = \sigma_{22} - \dfrac{\sigma_{21}^2}{\sigma_{11}}$, it results that:

$$
\begin{aligned}
&E\left[\theta|\, d(1), d(2)\right] = \\
&= \frac{1}{\sigma^2}\left(\sigma_{\theta 1}\frac{\sigma_{22}}{\sigma_{11}} - \sigma_{\theta 2}\frac{\sigma_{21}}{\sigma_{11}}\right) d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}}\right) d(2) + \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) - \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) = \\
&= \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 1}\frac{\sigma_{22}}{\sigma_{11}} - \sigma_{\theta 2}\frac{\sigma_{21}}{\sigma_{11}} - \frac{\sigma_{\theta 1}}{\sigma_{11}}\sigma^2\right) d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}}\right) d(2) = \\
&= \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 1}\frac{\sigma_{22}}{\sigma_{11}} - \sigma_{\theta 2}\frac{\sigma_{21}}{\sigma_{11}} - \frac{\sigma_{\theta 1}}{\sigma_{11}}\sigma_{22} + \frac{\sigma_{\theta 1}}{\sigma_{11}}\frac{\sigma_{21}^2}{\sigma_{11}}\right) d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}}\right) d(2) \\
&= \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) + \frac{1}{\sigma^2}\left(-\sigma_{\theta 2}\frac{\sigma_{21}}{\sigma_{11}} + \sigma_{\theta 1}\frac{\sigma_{21}^2}{\sigma_{11}^2}\right) d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}}\right) d(2) = \\
&= \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) - \frac{1}{\sigma^2}\frac{\sigma_{21}}{\sigma_{11}}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{21}}{\sigma_{11}}\right) d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{12}}{\sigma_{11}}\right) d(2) = \\
&= \frac{\sigma_{\theta 1}}{\sigma_{11}} d(1) + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{21}}{\sigma_{11}}\right) \left[d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)\right] = \\
&= E\left[\theta|\, d(1)\right] + \frac{1}{\sigma^2}\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{21}}{\sigma_{11}}\right) \left[d(2) - E\left[\, d(2)|\, d(1)\right]\right]
\end{aligned}
$$

Definition: given two scalar random variables $d(1)$ and $d(2)$ with zero mean value, the **innovation of $d(2)$ given $d(1)$** is the scalar random variable defined by:

$$e = d(2) - E\left[\, d(2) |\, d(1)\right] = d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)$$

- $E[e] = E\left[d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)\right] = E[d(2)] - \frac{\sigma_{21}}{\sigma_{11}} E[d(1)] = 0$
- $\sigma_{ee} = Var[e] = E\left[(e - E\,[e])^2\right] = E\left[e^2\right] = E\left[\left(d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)\right)^2\right] =$
  $= E\left[d^2(2) - 2\frac{\sigma_{21}}{\sigma_{11}} d(2)d(1) + \frac{\sigma_{21}^2}{\sigma_{11}^2} d^2(1)\right] = E\left[d^2(2)\right] - 2\frac{\sigma_{21}}{\sigma_{11}} E[d(2)d(1)] + \frac{\sigma_{21}^2}{\sigma_{11}^2} E\left[d^2(1)\right]$
  $= \sigma_{22} - 2\frac{\sigma_{21}}{\sigma_{11}}\sigma_{21} + \frac{\sigma_{21}^2}{\sigma_{11}^2}\sigma_{11} = \sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}} = \sigma^2 = \sigma_{22}\left(1 - \frac{\sigma_{21}^2}{\sigma_{11}\sigma_{22}}\right) = \sigma_{22}\left(1 - \rho_{21}^2\right) \leq \sigma_{22}$
- $\sigma_{\theta e} = E[\theta e] = E\left[\theta\left(d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)\right)\right] = E[\theta d(2)] - \frac{\sigma_{21}}{\sigma_{11}} E[\theta d(1)] = \sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{21}}{\sigma_{11}}$
- $\sigma_{1e} = E[d(1)e] = E\left[d(1)\left(d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)\right)\right] = E[d(1)d(2)] - \frac{\sigma_{21}}{\sigma_{11}} E\left[d^2(1)\right] =$
  $= \sigma_{12} - \frac{\sigma_{21}}{\sigma_{11}}\sigma_{11} = 0 \quad \Leftrightarrow \quad d(1)$ and e are uncorrelated, as well as $E\left[\, d(2) |\, d(1)\right]$ and e are

From the definition, it follows that: $d(2) = E\left[\, d(2) |\, d(1)\right] + e \quad \Rightarrow$
the term e represents the only new information provided by $d(2)$ with respect to $d(1)$

By exploiting the definition and the properties of the innovation e, it follows that:

$$E[\theta|\, d(1), d(2)] = E[\theta|\, d(1)] + \underbrace{\frac{1}{\sigma^2}}_{1/\sigma_{ee}} \underbrace{\left(\sigma_{\theta 2} - \sigma_{\theta 1}\frac{\sigma_{21}}{\sigma_{11}}\right)}_{\sigma_{\theta e}} \underbrace{[d(2) - E[\, d(2)|\, d(1)]]}_{e} =$$

$$= E[\theta|\, d(1)] + \frac{\sigma_{\theta e}}{\sigma_{ee}} e =$$

$$= E[\theta|\, d(1)] + E[\theta|\, e]$$

i.e., the optimal linear estimate of θ based on $d(1)$ and $d(2)$ is equal to the sum of:

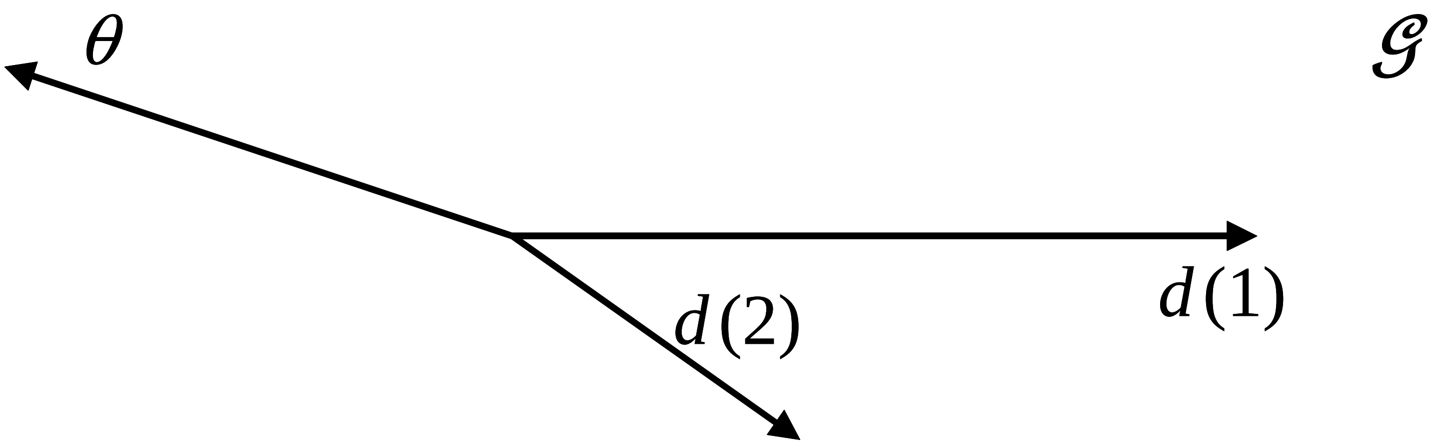
- the optimal linear estimate of θ based on the observation $d(1)$ only
- the optimal linear estimate of θ based on the innovation $e = d(2) - \frac{\sigma_{21}}{\sigma_{11}} d(1)$, which depends on data $d(1)$ and $d(2)$

It can be proved as well that:

$$\boxed{\mathbf{E}\,[\boldsymbol{\theta}|\, \mathbf{d(1)}, \mathbf{e}] = \mathbf{E}\,[\boldsymbol{\theta}|\, \mathbf{d(1)}, \mathbf{d(2)}] = \mathbf{E}\,[\boldsymbol{\theta}|\, \mathbf{d(1)}] + \mathbf{E}\,[\boldsymbol{\theta}|\, \mathbf{e}]}$$

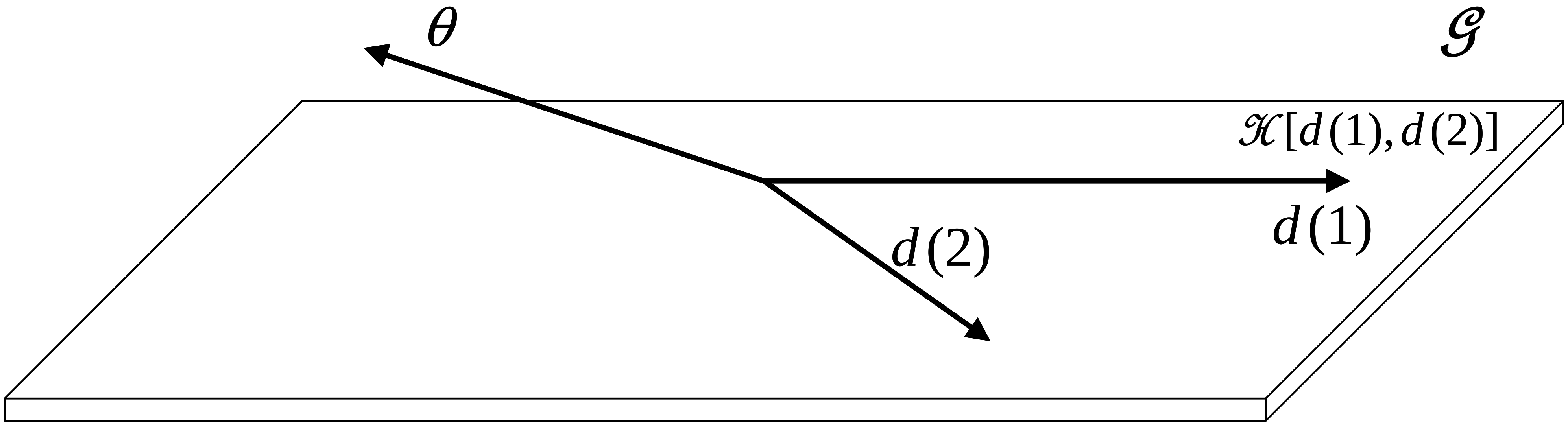
Geometrical interpretation

- Let us consider any random variable as a vector in the normed vector space $\mathcal{G}$
  $\Rightarrow$ the Bayesian estimate of θ based on d is the orthogonal projection of θ over d



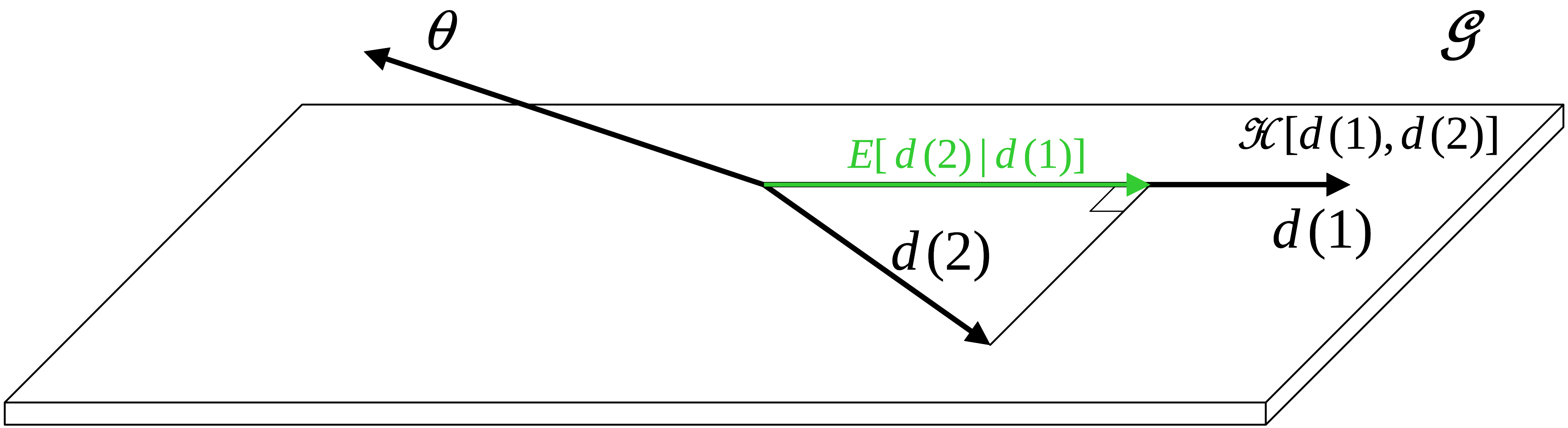
Geometrical interpretation

- Let us consider any random variable as a vector in the normed vector space $\mathcal{G}$ $\Rightarrow$ the Bayesian estimate of θ based on d is the orthogonal projection of θ over d
- Let $\mathcal{H}[d(1), d(2)]$ be the plane defined by the vectors $d(1)$ and $d(2)$



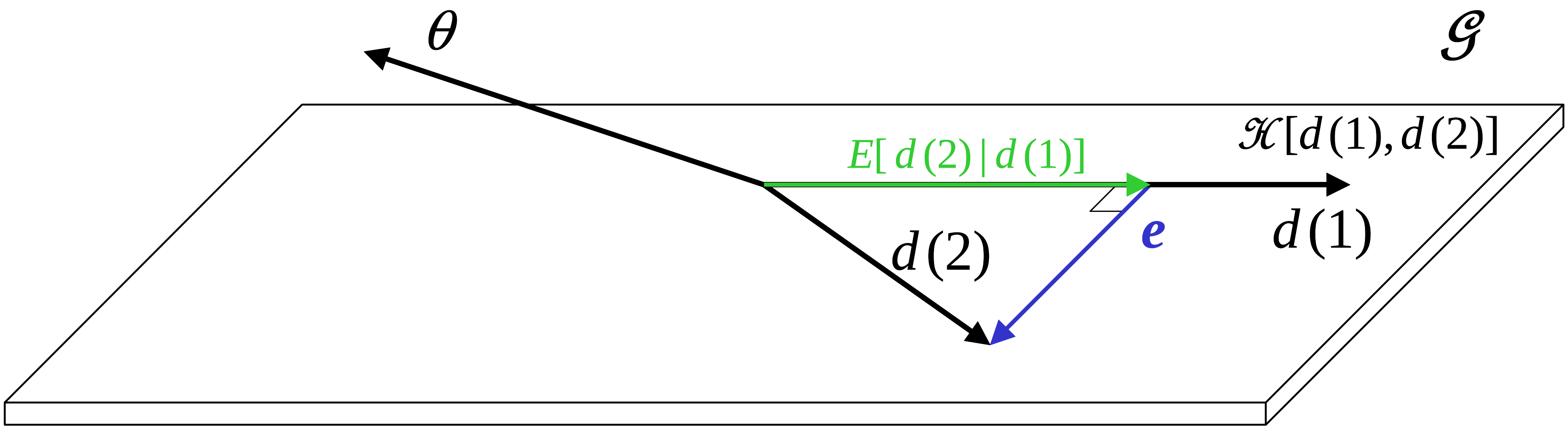
Geometrical interpretation

- Let us consider any random variable as a vector in the normed vector space $\mathcal{G}$ $\Rightarrow$ the Bayesian estimate of θ based on d is the orthogonal projection of θ over d
- Let $\mathcal{H}[d(1), d(2)]$ be the plane defined by the vectors $d(1)$ and $d(2)$
- The Bayesian estimate $E[\,d(2)|\,d(1)]$ is the projection of $d(2)$ over $d(1)$

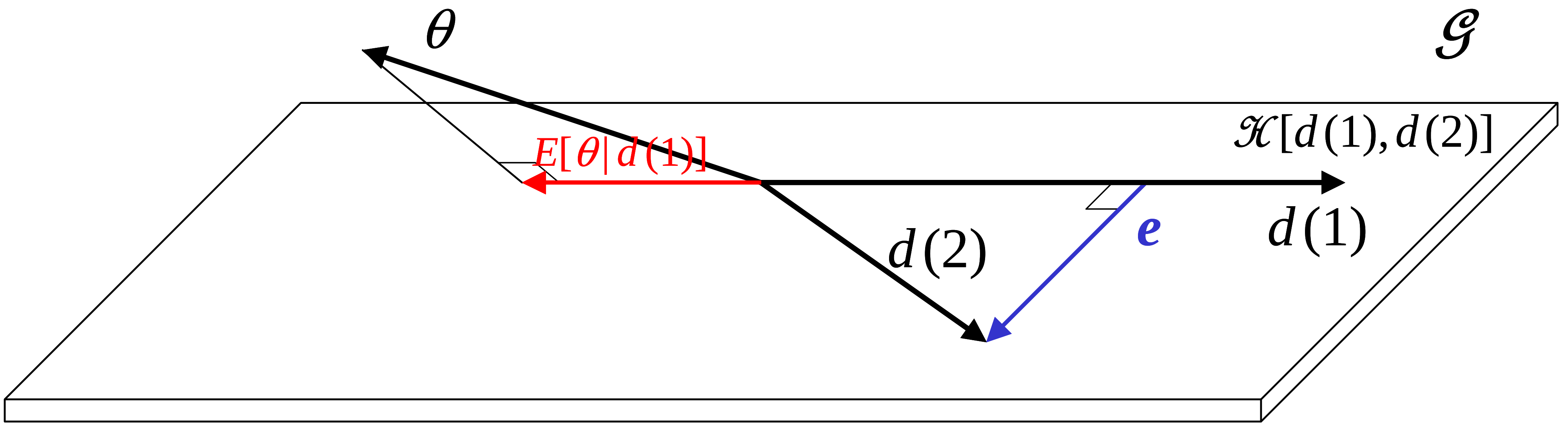


Geometrical interpretation

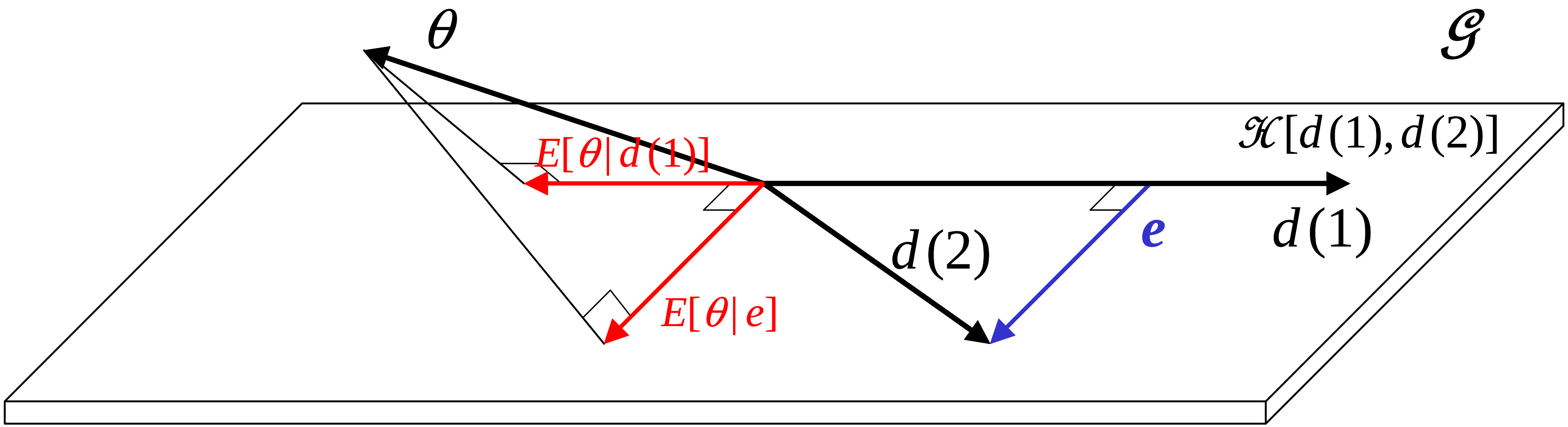
- Let us consider any random variable as a vector in the normed vector space $\mathcal{G}$ $\Rightarrow$ the Bayesian estimate of θ based on d is the orthogonal projection of θ over d
- Let $\mathcal{H}[d(1), d(2)]$ be the plane defined by the vectors $d(1)$ and $d(2)$
- The Bayesian estimate $E[\,d(2)|\,d(1)]$ is the projection of $d(2)$ over $d(1)$
- The innovation $e = d(2) - E[\,d(2)|\,d(1)]$ is the vector given by the difference between $d(2)$ and the projection of $d(2)$ over $d(1)$ and it is orthogonal to $d(1)$



- The Bayesian estimate $E[\theta|\, d(1)]$ is the orthogonal projection of θ over $d(1)$

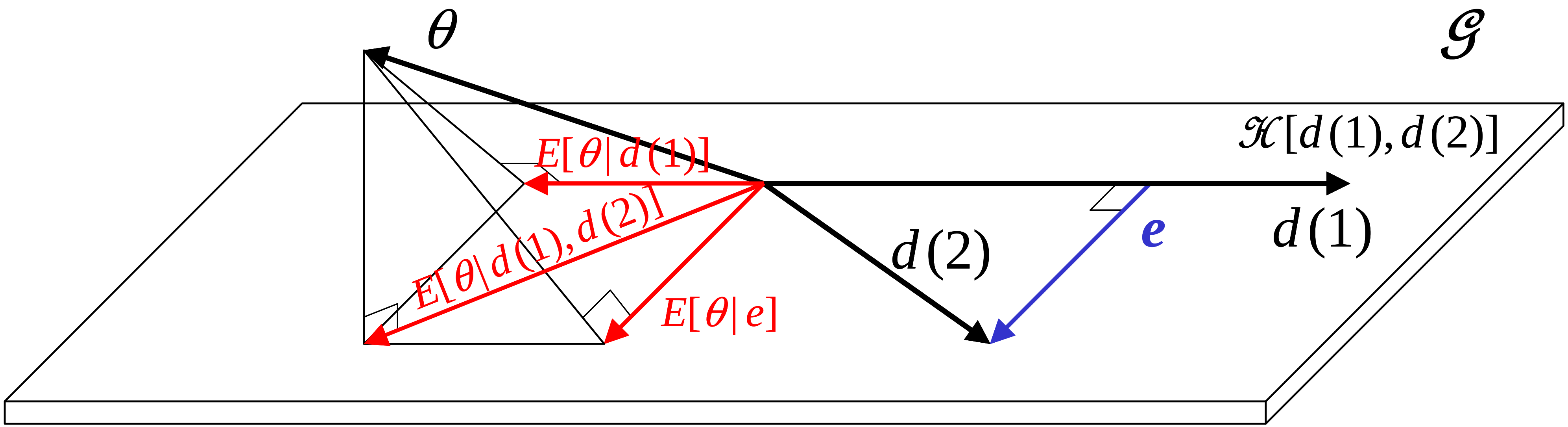


- The Bayesian estimate $E[\theta|\, d(1)]$ is the orthogonal projection of θ over $d(1)$
- The Bayesian estimate $E[\theta|\, e]$ is the orthogonal projection of θ over e and then it is orthogonal to $E[\theta|\, d(1)]$



- The Bayesian estimate $E[\theta|\, d(1)]$ is the orthogonal projection of θ over $d(1)$
- The Bayesian estimate $E[\theta|\, e]$ is the orthogonal projection of θ over e and then it is orthogonal to $E[\theta|\, d(1)]$
- The Bayesian estimate $E[\theta|\, d(1), d(2)]$ is the orthogonal projection of θ over the plane $\mathcal{H}[d(1), d(2)]$ and it is the vector sum of $E[\theta|\, d(1)]$ and $E[\theta|\, e]$:

$$E[\theta|\, d(1), d(2)] = E[\theta|\, d(1)] + E[\theta|\, e] = E[\theta|\, d(1), e]$$



Recursive Bayesian estimation: vector case

- If the unknown θ and the data d are vector random variables with zero mean value:

$$\begin{bmatrix}\theta\\ \underbrace{\begin{matrix}d(1)\\ d(2)\end{matrix}}_{d}\end{bmatrix} \sim \left(\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}, \Sigma = Var\begin{bmatrix}\theta\\ d(1)\\ d(2)\end{bmatrix} = \begin{bmatrix}\Sigma_{\theta\theta} & \overbrace{\begin{matrix}\Sigma_{\theta 1} & \Sigma_{\theta 2}\end{matrix}}^{\Sigma_{\theta d}} \\ \underbrace{\begin{matrix}\Sigma_{1\theta}\\ \Sigma_{2\theta}\end{matrix}}_{\Sigma_{d\theta}=\Sigma_{\theta d}^T} & \underbrace{\begin{matrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{matrix}}_{\Sigma_{dd}}\end{bmatrix}\right), \quad \begin{cases}\Sigma_{\theta 1} = \Sigma_{1\theta}^T\\ \Sigma_{\theta 2} = \Sigma_{2\theta}^T\\ \Sigma_{12} = \Sigma_{21}^T\end{cases}$$

by defining the innovation of $d(2)$ given $d(1)$ as the vector random variable:

$$e = d(2) - E\left[\, d(2)|\, d(1)\right] = d(2) - \Sigma_{21}\Sigma_{11}^{-1}d(1)$$

the optimal linear estimate of θ based on $d(1)$ and $d(2)$ is given by:

$$E[\theta|\, d(1), d(2)] = \Sigma_{\theta 1}\Sigma_{11}^{-1}d(1) + \Sigma_{\theta e}\Sigma_{ee}^{-1}e = E[\theta|\, d(1)] + E[\theta|\, e]$$

where

$$\Sigma_{ee} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, \quad \Sigma_{\theta e} = \Sigma_{\theta 2} - \Sigma_{\theta 1}\Sigma_{11}^{-1}\Sigma_{12}$$

- If the unknown θ and the data d are vector random variables with nonzero mean value:

$$\begin{bmatrix} \theta \\ d(1) \\ d(2) \end{bmatrix} \sim \left(\begin{bmatrix} \bar{\theta} \\ \bar{d}(1) \\ \bar{d}(2) \end{bmatrix}, \Sigma = Var \begin{bmatrix} \theta \\ d(1) \\ d(2) \end{bmatrix} = \begin{bmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta 1} & \Sigma_{\theta 2} \\ \Sigma_{1\theta} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{2\theta} & \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right), \quad \begin{cases} \Sigma_{\theta 1} = \Sigma_{1\theta}^T \\ \Sigma_{\theta 2} = \Sigma_{2\theta}^T \\ \Sigma_{12} = \Sigma_{21}^T \end{cases}$$

by defining the innovation of $d(2)$ given $d(1)$ as the vector random variable:

$$e = d(2) - \bar{d}(2) - E[d(2) - \bar{d}(2) \,|\, d(1) - \bar{d}(1)] = d(2) - \bar{d}(2) - \Sigma_{21}\Sigma_{11}^{-1}[d(1) - \bar{d}(1)]$$

the optimal linear estimate of θ based on $d(1)$ and $d(2)$ is given by:

$$E[\theta|\, d(1), d(2)] = \underbrace{\bar{\theta} + \Sigma_{\theta 1}\Sigma_{11}^{-1}[d(1) - \bar{d}(1)]}_{E[\theta|\, d(1)]} + \Sigma_{\theta e}\Sigma_{ee}^{-1} e =$$

$$= E[\theta|\, d(1)] + \underbrace{\Sigma_{\theta e}\Sigma_{ee}^{-1} e + \bar{\theta}}_{E[\theta|\, e]} - \bar{\theta} =$$

$$= E[\theta|\, d(1)] + E[\theta|\, e] - \bar{\theta}$$