3. COMPARISON CRITERIA FOR ESTIMATES. CRAMER-RAO LOWER BOUND

In this chapter, we will study criteria for comparing different (optimal) estimates. In general, we resort to the mean and covariance of the estimates. First, we see the following result.

FACT. If X, Y are random vectors and $E\{YY^T\}$ is nonsingular,

$$E\{XX^T\} - E\{XY^T\}\left(E\{YY^T\}\right)^{-1}E\{YX^T\} \geq 0$$

Proof. We have for all K:

$$E\{(X-KY)(X-KY)^T\} \geq 0,$$

or, which is the same,

$$E\{XX^T\} - KE\{YX^T\} - E\{XY^T\}K^T + KE\{YY^T\}K \geq 0$$

Choosing $K = K^* \triangleq E\{XY^T\}\left(E\{YY^T\}\right)^{-1}$:

$$0 \leq E\{XX^T\} - K^* E\{YX^T\} - E\{XY^T\} K^{*T} + K^* E\{YY^T\} K^* =$$

$$= E\{XX^T\} - E\{XY^T\}\left(E\{YY^T\}\right)^{-1} E\{YX^T\}$$

If X is the vectr to be estimated and Y the measurement vector, assume to have the density $p_{Y|X}$. Moreover, we will denote by $E_{Y|X}$ the expectation evaluated with $p_{Y|X}$.

THEOREM (CRAMER-RAO). For any estimate $\tilde{X} = f(Y)$,

$$R(x) \triangleq E_{Y|X}\{(f(Y) - x)(f(Y) - x)^T\}$$

satisfies

$$R(x) \geq \underbrace{\left(I + \frac{\partial S}{\partial x}(x)\right) \Lambda^{-1}(x) \left(I + \frac{\partial S}{\partial x}(x)\right)^T}_{\triangleq L(x)}$$

where

$$\Lambda(x) \triangleq E_{Y|X}\left\{\left(\frac{\partial}{\partial x}\ln p_{Y|X}(y,x)\right)^T \cdot \left(\frac{\partial}{\partial x}\ln p_{Y|X}(y,x)\right)\right\}$$

and

$$S(x) = E_{Y|X}\{f(Y) - x\},$$

as long as $p_{Y|X}(y,x)$ is differentiable w.r.t. x and $\Lambda(x)$ is non-singular $\forall x$.

PROOF. Since

$$\int_{\mathbb{R}^m} p_{Y|X}(y,x)\,dy = 1$$

then

$$\frac{\partial}{\partial x}\int_{\mathbb{R}^m} p_{Y|X}(y,x)\,dy = 0 .$$

It follows (under some regularity assumption on $p_{Y|X}$) that

$$\int_{\mathbb{R}^m} \frac{\partial}{\partial x} p_{Y|X}(y,x)\,dy = 0 .$$

But

$$\frac{\partial}{\partial \theta} p_{Y|X}(y,x) = \left[\frac{\partial}{\partial \theta} \ln p_{Y|X}(y,x)\right] p_{Y|X}(y,x)$$

$\Rightarrow$

$$\int_{\mathbb{R}^m} \left[\frac{\partial}{\partial x} \ln p_{Y|X}(y,x)\right] p_{Y|X}(y,x)\,dy = 0$$

$\Rightarrow$ premultiplying by $x \in \mathbb{R}^n$:

$$\int_{\mathbb{R}^m} x \left[\frac{\partial}{\partial x} \ln p_{Y|X}(y,x)\right] p_{Y|X}(y,x)\,dy = 0$$

$$\Rightarrow E_{Y|X}\left\{x \frac{\partial}{\partial x} \ln p_{Y|X}(Y,x)\right\} = 0 \quad (*)$$

By definition of $S(x)$:

$$\frac{\partial}{\partial x} E_{Y|X}\{f(Y)\} = \frac{\partial}{\partial x}\left(x + S(x)\right) \quad (**)$$

$$= I + \frac{\partial S(x)}{\partial x}$$

Also by differentiation of

$$E_{Y|X}\{f(Y)\} = \int_{\mathbb{R}^m} f(y)\, p_{Y|X}(y,x)\,dy$$

w.r.t. x, we obtain

$$\frac{\partial}{\partial x} E_{Y|X}\{f(Y)\} = \qquad (***)$$

$$= \int_{\mathbb{R}^m} f(y)\left[\frac{\partial}{\partial x} \ln p_{Y|X}(y,x)\right] p_{Y|X}(y,x)\,dy$$

$$= E_{Y|X}\left\{f(Y)\frac{\partial}{\partial x} \ln p_{Y|X}(Y,x)\right\}.$$

Equating $(**)$ and $(***)$, using $(*)$:

$$I + \frac{\partial S}{\partial x}(x) = E_{Y|X}\left\{f(Y)\frac{\partial}{\partial x} \ln p_{Y|X}(Y,x)\right\}$$

$$= E_{Y|X}\left\{\left[f(Y)-x\right]\frac{\partial}{\partial x} \ln p_{Y|X}(Y,x)\right\} \qquad (****)$$

Define

$$V \triangleq f(Y) - x$$

$$W \triangleq \left[\frac{\partial}{\partial x} \ln p_{Y|X}(Y,x)\right]^T$$

From FACT we have that

$$E_{Y|X}\{VV^T\} \geq E_{Y|X}\{VW^T\}\left(E_{Y|X}\{WW^T\}\right)^{-1} \cdot E_{Y|X}\{WV^T\}$$

(since $E_{Y|X}\{WW^T\} = \Lambda(x)$ and it is nonsingular).

This gives exactly the result of the theorem, since from (****)

$$E_{Y|X}\{VW^T\} = I + \frac{\partial S}{\partial x}(x)$$

The matrix $L(x)$ is the CRAMER-RAO LOWER BOUND, $S(x)$ is the POLARIZATION of the estimate and $\Lambda(x)$ is the FISHER INFORMATION MATRIX.

REMARK. $p_{Y|X}(y,x)$ can be replaced by $p_Y(y,x)$ where x is a deterministic parameter.

4. PROPERTIES OF ESTIMATES.

<u>DEFINITION</u>. An estimate is said to be EFFICIENT if

$$R(x) = \left(I + \frac{\partial S}{\partial x}(x)\right)\Lambda^{-1}(x)\left(I + \frac{\partial S}{\partial x}(x)\right)^T$$

$\forall$ x (admissible).

<u>DEFINITION</u>. An estimate is said to be CENTERED if for all x (admissible) the polarization

$$S(x) = 0 .$$

An estimate which is not centered is said policised.

<u>FACT</u>. Necessary condition for an estimate to be efficient is that it is also centered

Proof. If an estimate is efficient and again with

$$Y \triangleq f(Y) - x \quad , \quad W \triangleq \left[\frac{\partial}{\partial x} \ln p_{Y/X}(Y, x)\right]^T$$

one has:

$$E_{Y|X}\{VV^T\} = E_{Y|X}\{VW^T\}\left(E_{Y|X}\{WW^T\}\right)^{-1} \cdot$$

$$\cdot E_{Y|X}\{WV^T\}.$$

With $K^* \triangleq E_{Y|X}\{VW^T\}\left(E_{Y|X}\{WW^T\}\right)^{-1}$:

$$E_{Y|X}\{(V-K^*W)(V-K^*W)^T\} = 0$$

This is a positive semidefinite matrix and it is zero iff

$$V - K^*W = 0$$

which is

$$V = f(Y) - x = E_{Y|X}\{VW^T\}\left(E_{Y|X}\{WW^T\}\right)^{-1} \cdot W$$

Therefore, by applying $E_{Y|X}$ and since $E_{Y|X}\{W\} = 0$:

$$E_{Y|X}\{V\} = E_{Y|X}\{f(Y) - x\} = 0$$

$$\Rightarrow \quad E_{Y|X}\{f(Y)\} = x$$

In many applications the vector Y gets larger in dimension, since it contains an increasing number of measurements. It is expected in these cases that the estimate becomes more precise as the number of measurements increases.

DEFINITION. A sequence of estimates $\{\hat{X}_N\}$, $N=1,2,\dots$, of X is CONSISTENT if $\hat{X}_N \triangleq f_N(Y)$ converges in probability

$\mathcal{P}_{Y|X}$:

$$\hat{X}_N \triangleq f_N(Y) \xrightarrow{\mathcal{P}_{Y|X}} X$$

or, which is the same, $\forall \epsilon > 0$:

$$\lim_{N\to+\infty} \mathcal{P}_{Y|X}\{\| f_N(Y(\omega)) - X(\omega)\| > \epsilon\} = 0$$

The index N may represent a temporal index or a discretization index.

EXAMPLE 1. (MAXIMUM LIKELIHOOD)

Consider the problem of estimating the mean m and variance σ^2 of a gaussian random variable using a certain number N of experiments Y_i, $i=1,..,N$, assumed to be independent each other. In this case, the vector to be estimated is $X = \begin{pmatrix} m \\ \sigma \end{pmatrix}$ and it is determinictic. We will implement a maximum likelihood estimate. The density $p_{Y_i}(y_i;x)$ is (gaussian)

$$p_{Y_i}(y_i;x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y_i-m)^2/2\sigma^2}$$

$$= \frac{1}{\sqrt{2\pi}\,x_2} e^{-(y_i-x_1)^2/2x_2^2}$$

where $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \triangleq \begin{pmatrix} m \\ \sigma \end{pmatrix}$.

Since Y_i, $i=1,..,N$, are independent

$$p_{Y_1 \cdots Y_N}(y_1,..y_N;x) = \prod_{i=1}^{N} p_{Y_i}(y_i;x)$$

$$= \frac{1}{(2\pi)^{N/2}} \cdot \frac{1}{x_2^N} e^{-\frac{1}{2x_2^2}\sum_{i=1}^{N}(y_i-x_1)^2}$$

Our estimates is given as:

$$\hat{X} = \underset{x:\, x_2>0}{\operatorname{argmax}}\ p_{Y_1 \dots Y_N}(y_1, \dots, y_N; x).$$

In our case it is more convenient to maximize $\ln p_{Y_1 \dots Y_N}(y_1, \dots y_N; x)$:

$$\hat{X} = \underset{x:\, x_2>0}{\operatorname{argmax}}\ p_{Y_1 \dots Y_N}(y_1, \dots, y_N; x)$$
$$= \underset{x:\, x_2>0}{\operatorname{argmax}}\ \ln p_{Y_1 \dots Y_N}(y_1, \dots, y_N; x)$$

Extremals for this problem are obtained from

$$0 = \frac{\partial}{\partial x}\left(\ln p_{Y_1 \dots Y_N}(y_1, \dots, y_N; x)\right)\Big|_{x=\hat{x}}$$

Therefore,

$$0 = \frac{\partial}{\partial x_1}\left(\ln p_{Y_1 \dots Y_N}(y_1, \dots, y_N; x)\right)\Big|_{x=\hat{x}}$$
$$= \frac{1}{\hat{x}_2^2} \sum_{i=1}^{N} (y_i - \hat{x}_1) \Rightarrow \hat{x}_1 = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$= \frac{\partial}{\partial x_2}\left(\ln p_{Y_1 \cdots Y_N}(y_1, \dots, y_N; x)\right)\Big|_{x=\hat{x}}$$

$$= -\frac{N}{\hat{x}_2^2} + \frac{1}{\hat{x}_2^3} \sum_{i=1}^{N} (y_i - \hat{x}_1)^2$$

$$\Rightarrow \hat{x}_2^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{x}_1)^2$$

Since

$$\ln p_{Y_1 \cdots Y_N}(y_1, \dots, y_N; x) = -N \log x_2 - \frac{1}{2x_2^2} \sum_{i=1}^{N} (y_i - x_1)^2$$

and

$$\lim_{\|x\| \to +\infty} \ln p_{Y_1 \cdots Y_N}(y_1, \dots, y_N; x) = -\infty$$

then

$$\hat{x} = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{N} \sum_{i=1}^{N} y_i \\ \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{x}_1)^2} \end{pmatrix}$$

is the maximum of $\ln p_{Y_1 \cdots Y_N}(y_1, \dots, y_N; x)$, with admissibility condition

$$\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{x}_1)^2 > 0 .$$

The estimate

$$\hat{x}_1 = \frac{1}{N} \sum_{i=1}^{N} y_i$$

is centered:

$$E\{\hat{x}_1\} = \frac{1}{N} \sum_{i=1}^{N} E\{y_i\} = \frac{1}{N} N x_1 .$$

The estimate of σ^2

$$\hat{x}_2^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{x}_1)^2$$

is not centered:

$$E\{\hat{x}_2^2\} = \frac{1}{N} \sum_{i=1}^{N} E\left\{\left(y_i - \frac{1}{N}\sum_{j=1}^{N} y_j\right)^2\right\}$$

$$= \frac{1}{N} \sum_{i=1}^{N} E\left\{\left(y_i - x_1 - \frac{1}{N}\sum_{j=1}^{N} (y_j - x_1)\right)^2\right\}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \Bigg(E\{(y_i - x_1)^2\} + \frac{1}{N} E\left\{\left(\sum_{j=1}^{N}(y_j - x_1)\right)^2\right\}$$
$$- \frac{2}{N} E\left\{(y_i - x_1) \sum_{j=1}^{N} (y_j - x_1)\right\} \Bigg)$$

But

$$E\{(y_i - x_1)(y_j - x_1)\} = 0$$

$\forall\, i \neq j$ (since Y_i, Y_j are independent for $i \neq j$)

Therefore,

$$E\{\hat{x}_2^2\} = \frac{1}{N}\sum_{i=1}^{N}\left(\sigma^2 + \frac{1}{N^2}N\sigma^2 - \frac{2}{N}\sigma^2\right)$$

$$= \frac{N-1}{N}\sigma^2$$

The "modified" estimate

$$\overline{x_2} \triangleq \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{x}_1)^2 = \frac{N}{N-1}\hat{x_2}$$

will be centered, but not a maximum likelihood estimate. ◀

<u>EXAMPLE 2.</u> (Fisher matrix and Cramer–Rao lower bound).

Consider

$$Y(\omega) = Ax + N(\omega)$$

where $N \in \mathcal{N}(0, \Psi_N)$, x deterministic (unknown). Clearly

$$Y \in \mathcal{N}(\underbrace{Ax}_{m_Y}, \underbrace{\Psi_N}_{\Psi_Y}).$$

Therefore

$$p_Y(y;x) = \frac{1}{(2\pi)^{m/2} (\det \Psi_N)^{1/2}} \cdot e^{-\frac{1}{2}(y-Ax)^T \Psi_N^{-1} (y-Ax)}$$

and a maximum likelihood estimate is

$$\hat{x} = \underset{x}{\operatorname{argmax}} \ln p_Y(y;x)$$

$$= \underset{x}{\operatorname{argmin}} \underbrace{\frac{1}{2}(y-Ax)^T \Psi_N^{-1} (y-Ax)}_{J(y;x)}$$

Extremals are obtained from:

$$0 = \frac{\partial}{\partial x}\left(\frac{1}{2}(y-Ax)^T \Psi_N^{-1} (y-Ax)\right)\Big|_{x=\hat{x}}$$

$$= -(y-Ax)^T \Psi_N^{-1} A \Big|_{x=\hat{x}}$$

$\Rightarrow$ (if A has full column rank)

$$\hat{x} = \underbrace{(A^T \Psi_N^{-1} A)^{-1} A^T \Psi_N^{-1}}_{A^{\#}_{\Psi_N}} y$$

The maximum likelihood estimate is

$$\hat{x} = A^{\#}_{\Psi_N} y$$

We have

$$E\{\hat{x}\} = E\{A^{\#}_{\Psi_N} y\} = A^{\#}_{\Psi_N} E\{y\}$$

$$= A^{\#}_{\Psi_N} A x = x$$

$\Rightarrow \hat{x}$ is centered.

Moreover,

$$E\{(x-\hat{x})(x-\hat{x})^T\} =$$

$$= E\{A^{\#}_{\Psi_N}(y-Ax)(y-Ax)^T (A^{\#}_{\Psi_N})^T\}$$

$$= (A^T \Psi_N^{-1} A)^{-1}$$

The CRAMER-RAO lower bound is

$$(I + \frac{\partial S}{\partial x}(x))^T \Lambda^{-1}(x) (I + \frac{\partial S}{\partial x}(x))$$

where

$$\Lambda(x) \triangleq E_Y \left\{ \left(\frac{\partial}{\partial x} \ln p_Y(Y;x) \right)^T \left(\frac{\partial}{\partial x} \ln p_Y(Y;x) \right) \right\}$$

$$S(x) \triangleq E_Y \{\hat{x} - x\}$$

In our case

$$\frac{\partial}{\partial x} \ln p_Y(Y;x) = A^T \Psi_N^{-1}(Y - Ax)$$

so that

$$\begin{cases} S(x) = 0 \\ \Lambda(x) = E\{A^T \Psi_N^{-1}(Y-Ax)(Y-Ax)^T \Psi_N^{-1} A\} \\ \quad = A^T \Psi_N^{-1} A \end{cases}$$

$\Rightarrow \hat{x}$ is efficient.

EXAMPLE 3. (Markov estimators)

Consider the problem of estimating a deterministic x from M independent measurements:

$$y_i = x + n_i \quad , \quad i = 1, .., M,$$

with $E\{n_i\} = 0$, $E\{n_i^2\} = \sigma_{n_i}^2$.

We can collect all the measurements $y_1 \dots y_M$ into a single vector

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix}$$

and also

$$N = \begin{pmatrix} n_1 \\ \vdots \\ n_M \end{pmatrix} \Rightarrow$$

$$Y = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} x + N$$

with $\Psi_N = \text{diag}\{\sigma_{N_1}^2, \dots, \sigma_{N_M}^2\}$

$$= \begin{pmatrix} \sigma_{N_1}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{N_M}^2 \end{pmatrix}.$$

We can implement in this case a weighted least square estimate $\hat{x}$ of x ; selecting the weights matrix

$$W = \Psi_N^{-1} = \text{diag}\left\{\frac{1}{\sigma_{N_1}^2}, \dots, \frac{1}{\sigma_{N_M}^2}\right\}.$$

We obtain

$$\hat{x} = A_W^{\#} y = \frac{\sum_{i=1}^{M} \frac{y_i}{\sigma_{N_i}^2}}{\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}}$$

The weighted least squares estimators with $W = \Psi_N^{-1}$ are called MARKOV estimators.

It should be noticed that this estimate coincides with the maximum likelihood estimate (since the measurement equation is linear in x and N is gaussian).

Also, if $\sigma_{N_i}^2 = \sigma_{N_j}^2 = \sigma^2 \; \forall i,j \Rightarrow$

$$\hat{x} = \frac{\sum_{i=1}^{M} \frac{y_i}{\sigma_{N_i}^2}}{\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}} = \frac{1}{M} \sum_{i=1}^{M} y_i$$

which is the classical least square estimate $\hat{x}$ of x (with $W = I$).

Notice that from the Markov estimate $\hat{x}$:

$$x - \hat{x} = \frac{\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2} x - \sum_{i=1}^{M} \frac{y_i}{\sigma_{N_i}^2}}{\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}}$$

$$= \frac{\sum_{i=1}^{m} \frac{1}{\sigma_{N_i}^2} (x - y_i)}{\sum_{i=1}^{m} \frac{1}{\sigma_{N_i}^2}}$$

But n_i and n_j are independent for $i \neq j$:

$$\sigma_{\hat{x}}^2 = E\{(x-\hat{x})^2\} = \frac{\sum_{i=1}^{M} \frac{E\{(x-y_i)^2\}}{\sigma_{N_i}^4}}{\left(\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}\right)^2}$$

$$= \frac{1}{\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}}.$$

Instead, for the classical least square estimate $\hat{x}$ of x:

$$\sigma_{\hat{x}}^2 = \frac{1}{M^2} \sum_{i=1}^{M} \sigma_{N_i}^2 \geqslant \frac{1}{\sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}}$$

therefore, the variance of the Markov estimator is less than the variance of a classical least square estimator. On the other hand, in the present case

$$\Lambda(x) = (1 \cdots 1)\, \Psi_N^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \sum_{i=1}^{M} \frac{1}{\sigma_{N_i}^2}$$

EXAMPLE 4. Consider the problem of estimating the resistance R of an electrical device using m voltage/current measurements, affected by noise. Our model is

$$\underbrace{\begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}}_{\substack{\text{voltage}\\ \text{measurements}\\ V}} = \underbrace{\begin{bmatrix} i_1 \\ \vdots \\ i_m \end{bmatrix}}_{\substack{\text{current}\\ \text{measurements}\\ I}} R + \underbrace{\begin{bmatrix} n_1 \\ \vdots \\ n_m \end{bmatrix}}_{\text{noise } N} \qquad (*)$$

with $\Psi_N = \text{diag}\{\sigma_{n_1}^2, \ldots, \sigma_{n_m}^2\}$ and $E\{N\} = 0$ (i.e. uncorrelated noise with zero mean). If in addition we assume $N \in \mathcal{N}(0, \Psi_N)$, a maximum likelihood estimate $\hat{R}$ of R is (see example 2 above)

$$\hat{R} = \frac{\sum_{j=1}^{m} \left(i_j v_j / \sigma_{N_j}^2 \right)}{\sum_{j=1}^{m} \left(i_j^2 / \sigma_{N_j}^2 \right)}$$

Clearly, (*) assumes that the current measurements are not affected by noise, while the voltage measurements are affected by the noise H. If the current measurements are affected by noise (while the voltage measurements not) we use the other model

$$I = V\frac{1}{R} + H$$

noise

admittance

currents voltages

and a maximum likelihood estimate $\hat{Z}$ of $Z \triangleq \frac{1}{R}$ is

$$\hat{Z} = \frac{\displaystyle\sum_{j=1}^{m} \frac{i_j v_j}{\sigma_{H_j}^2}}{\displaystyle\sum_{j=1}^{m} \frac{v_j^2}{\sigma_{H_j}^2}}$$

If all the variances $\sigma_{H_j}^2$ or $\sigma_{H_j}^2$ are equal

$$\hat{R} = \frac{V^T I}{I^T I}$$

$$\hat{Z} = \frac{V^T I}{V^T V}$$

so that from Cauchy-Schwatz inequality

$$\hat{R}\hat{Z} = \frac{(V^T I)^2}{(I^T I)(V^T V)} = \frac{\langle V, I \rangle^2}{\langle I, I \rangle \langle V, V \rangle} \leq 1$$

EXAMPLE 5. Consider

$$Y = AX + N$$

in which $N \in \mathcal{N}(0, \Psi_N)$ and $X \in \mathcal{N}(m_X, \Psi_X)$. Clearly, Y is also gaussian with

$$\underbrace{E\{Y\}}_{\triangleq m_Y} = E\{AX + N\} = Am_X$$

$$\underbrace{E\{(A(X-m_X)+N)(A(X-m_X)+N)^T\}}_{\triangleq \Psi_Y} =$$

$$= A\Psi_X A^T + \Psi_N$$

in which we also assume that N and X are independent (for simplicity). Moreover,

$$\Psi_{XY} \triangleq E\{(X-m_X)(Y-m_Y)^T\} =$$

$$= E\{(X-m_X)(A(X-m_X)+N)^T\} =$$

$$= \Psi_X A^T.$$

Therefore,

$Z \triangleq \begin{pmatrix} X \\ Y \end{pmatrix} \in \mathcal{N}(m_Z, \Psi_Z)$ where

$$m_Z \triangleq \begin{pmatrix} m_X \\ Am_X \end{pmatrix}$$

$$\Psi_Z \triangleq \left[\begin{array}{c|c} \Psi_X & \Psi_X A^T \\ \hline A\Psi_X & A\Psi_X A^T + \Psi_N \end{array}\right]$$

Recalling that the estimate $\hat{X}$ with minimum variance of X given the measurements Y is

$$\hat{X} = m_X + \Psi_{XY} \Psi_Y^{-1} (Y - m_Y)$$

then we obtain

$$\hat{X} = m_X + \Psi_X A^T (A \Psi_X A^T + \Psi_H)^{-1} (Y - A m_X)$$

This is also the solution of

$$\hat{X} = \arg\min_x J(x, Y)$$

$$J(x,Y) \triangleq (x - m_X)^T \Psi_X^{-1} (x - m_X) + (Y - Ax)^T \Psi_H^{-1} (Y - Ax)$$

We conclude that, since we are minimizing $J(x,Y)$,

(*) $\hat{x}$ is "close" to m_X

(**) $A\hat{x}$ is "close" to y.

a maximum likelihood estimate, is designed to satisfy only (**), since no a priori information is available on X (like mean or covariance).

Notice that

$$\hat{X} = A^{\#}(Y - Am_X) + m_X$$

where

$$A^{\#} = \Psi_X A^T \underbrace{(A \Psi_X A^T + \Psi_N)}^{-1} \quad (\circ)$$

in which we are inverting a matrix of dimensions equal to the dimension of Y. Since the dimension of Y is, in most applications, greater than the dimension of X, it is convenient to use an alternative expression of $A^{\#}$ in which we have to invert a matrix with dimension equal to the dimension of X :

$$A^{\#} = \left(\Psi_X^{-1} + A^T \Psi_N^{-1} A\right)^{-1} A^T \Psi_N^{-1} \quad (\bullet\bullet)$$

The equality between the two formulas $(\bullet)$ and $(\bullet\bullet)$ is guaranteed by the following ~~fact~~ fact.

<u>FACT</u>. if P, Q are nonsingular matrix

$$QD(P + CQD)^{-1} = (DP^{-1}C + Q^{-1})^{-1} DP^{-1}$$

with P, Q, C, D having suitable dimensions.

Proof. From

$$QD = Q(DP^{-1}CQ + I)^{-1}(DP^{-1}CQ + I)D$$

and since $(AB + I)^{-1} = B^{-1}(A + B^{-1})^{-1}$

$\Rightarrow$

$$QD = [DP^{-1}C + Q^{-1}]^{-1} D [I + P^{-1}CQD]$$

$$= [DP^{-1}C + Q^{-1}]^{-1} DP^{-1} [P + CQD] \blacktriangleleft$$

From

$$\hat{X} = A^{\#}(Y - Am_X) + m_X$$

with $A^{\#} = (\Psi_X^{-1} + A^T \Psi_N^{-1} A)^{-1} A^T \Psi_N^{-1}$

if $\Psi_X^{-1} \to 0$ then $\hat{X}$ tends to the maximum likelihood estimate (since $A^{\#} \to A^{\#}_{\Psi_N}$). This case corresponds to consider a random vector to be deterministic and unknown (its distribution is uniform on $\mathbb{R}^n$).

Moreover, from the formula for the covariance of the estimate $\hat{X}$ with minimum variance

$$\Psi_{\hat{X}} = \Psi_X - \Psi_{XY} \Psi_Y^{-1} \Psi_{XY}^T$$

with the substitutions for Ψ_Y, Ψ_{XY}:

$$\Psi_{\hat{X}} = \Psi_X - \underbrace{\Psi_X A^T (A \Psi_X A^T + \Psi_N)^{-1} A \Psi_X}$$

$$= \underbrace{(I - A^{\#}A)}_{\uparrow} \Psi_X$$

additional information
on correlation between X
and Y

Also we can rewrite

$$\hat{X} = A^{\#}Y + \underbrace{(I - A^{\#}A) m_X}_{\uparrow}$$

additional information
on correlation between
X and Y

5. SIMULTANEOUS ESTIMATION OF RANDOM VARIABLE AND PARAMETERS

Consider the problem of estimating a random vector X from the measurements vector Y, with a joint density $p_{X,Y}(x,y;\vartheta)$ depending on a deterministic parameter vector $\vartheta \in \mathbb{R}^p$. The estimate with minimum variance is

$$\hat{X} = E_{\vartheta}\{X \mid Y\} \triangleq f(Y;\vartheta) .$$

We must find first

$$\underbrace{\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmax}}\ p_Y(y;\vartheta)}_{\text{maximum likelihood}}$$

where $p_Y(y;\vartheta) = \int_{\mathbb{R}^n} p_{X,Y}(x,y;\vartheta)\,dx$.

Finally replace $\hat{\vartheta}$ in

$$\hat{X} = E_{\vartheta}\{X|Y\}\Big|_{\vartheta=\hat{\vartheta}}$$

EXAMPLE. Consider

$$Y = AX + B\vartheta + N$$

with random $X \in \mathbb{R}^n$, $N \in \mathbb{R}^m$ and deterministic parameter vector $\vartheta \in \mathbb{R}^p$. The mean of X is m_X and its covariance Ψ_X; while the mean of N is 0 and its covariance Ψ_N (a priori information). Moreover, let X and N be independent and gaussian. also Y is gaussian as linear combination of gaussian vectors. We have as estimate of X with minimum variance

$$\hat{X} = m_X + \Psi_{XY}(\vartheta)\Psi_Y^{-1}(\vartheta)(Y - m_Y(\vartheta))$$

dependence on ϑ (*)

Moreover

$$m_Y(\vartheta) = Am_X + B\vartheta$$
$$\Psi_{XY}(\vartheta) = \Psi_X A^T$$
$$\Psi_Y(\vartheta) = A\Psi_X A^T + \Psi_H$$

$\Rightarrow$

$$\hat{X} = m_X + \Psi_X A^T (A\Psi_X A^T + \Psi_H)^{-1}(Y - Am_X - B\vartheta)$$

Also, we have, since Y is gaussian,

$$p_Y(y,\vartheta) = \frac{1}{(2\pi)^{m/2}(\det \Psi_Y)^{1/2}} \cdot e^{-\frac{1}{2}(y - m_Y(\vartheta))^T \Psi_Y^{-1}(y - m_Y(\vartheta))}$$

Moreover

$$\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmax}}\, p_Y(y,\vartheta) = \underset{\vartheta}{\operatorname{argmax}} \ln p_Y(y,\vartheta)$$

and

$$\ln p_Y(y;\vartheta) = \ln \frac{1}{(2\pi)^{m/2}(\det \Psi_Y)^{1/2}} - \frac{1}{2}(y - Am_X - B\vartheta)^T \Psi_Y^{-1}(y - Am_X - B\vartheta)$$

We obtain (assuming rank $B = p$)

$$\hat{\vartheta} = (B^T \Psi_Y^{-1} B)^{-1} B^T \Psi_Y^{-1} (y - A m_X)$$

and replacing in (*) ϑ with $\hat{\vartheta}$:

$$\hat{X} = m_X + \Psi_X A^T \cdot \Psi_Y^{-1} \cdot$$

$$\cdot (I - B(B^T \Psi_Y^{-1} B)^{-1} B^T \Psi_Y^{-1})(y - A m_X)$$

$$= m_X + \Psi_X A^T \cdot$$

$$\cdot \{\Psi_Y^{-1} - \Psi_Y^{-1} B (B^T \Psi_Y^{-1} B)^{-1} B^T \Psi_Y^{-1}\}(y - A m_X)$$