ESTIMATION THEORY

By "estimate" we mean a reasonable evaluation of unaccessible variables from directly accessible variables.
By doing it is important to take into account:
- the relation between unaccessible and accessible variables
- noise affecting this relation and a priori information on the noise itself.

We may distinguish:
- <u>deterministic estimate</u>, in which we have a deterministic relation between accessible and unaccessible variables

- probabilistic (or stochastic) estimate, in which we have a deterministic relation between accessible and unaccessible variable and we also use a priori information on the noise (for example, its density).

To start with, we may formulate our problem as follows:

let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$

$m > n$, $d \in \mathbb{R}^m$,

$$\boxed{y = Ax} + \boxed{d}$$

noise.

deterministic relation between y and x (model)

y represents the accessible variables, while x represents the unaccessible variables.

Our problem is to give a "reasonable" evaluation $\hat{x}$ of x, starting from y.
It is convenient to establish some optimal criteria to evaluate how "good" is an estimate.
Both deterministic and stochastic estimates can be different according to the selected optimal criteria.

In the case of deterministic estimate, we consider d as a "uncertainty" in the deterministic relation between y and x. Usually, an "admissible" set $D \subset \mathbb{R}^n$ is specified in such a way that our estimate $\hat{x}$ of x is "admissible" if $\hat{x} \in D$. A very simple optimal criterium to start with is to minimize in "some sense" the error (due to the error $x - \hat{x}$) $\varepsilon \triangleq y - A\hat{x}$. If we define

$$\| v \| \triangleq \sqrt{v^T W v}, \quad v \in \mathbb{R}^m,$$

W symmetric and positive definite, we say that $\hat{x}$ is optimal if

$$\hat{x} = \underset{x \in \mathcal{D}}{\operatorname{argmin}} \| y - Ax \|^2 .$$

Clearly, $\hat{x}$ must satisfy

$$\frac{\partial}{\partial x} \| y - Ax \|^2 \Big|_{x = \hat{x}} = 0$$

or

$$-2A^T W (y - A\hat{x}) = 0 .$$

If A has full rank r, [illegible]

$$\hat{x} = \underset{x \in \mathcal{D}}{\operatorname{argmin}} \| y - Ax \|^2$$

$$= A_W^+ y, \quad A_W^+ \triangleq (A^T W A)^{-1} A^T W$$

as long as $\hat{x} \in \mathcal{D}$.
Such estimate is known as WEIGHTED LEAST SQUARE estimate.
Notice that $A_W^+ A = I$.
If $W = I$, the estimate $\hat{x}$ is known as CLASSICAL LEAST SQUARE estimate and $\| \varepsilon \| \equiv \| \varepsilon \|_2$ (eucleaden norm).
We can also interpret this estimate as suitable orthogonal projection. Consider the space $\mathcal{H} = \mathbb{R}^m$ and define the scalar product

$$\boxed{\begin{array}{l} \langle y_1, y_2 \rangle_{\mathcal{H}} = y_1^T W y_2, \\ y_1, y_2 \in \mathcal{H}. \end{array}}$$

Moreover, let $\mathcal{M} \triangleq \mathrm{Im}\{A\}$, the subspace of $\mathbb{R}^m$ generated by the columns of A.

The unique vector $\hat{y} \in \mathcal{M}$ such that,

$$\| y - \hat{y} \|_{\mathcal{H}} \leq \| y - Ax \|_{\mathcal{H}}$$

$$\forall x \in \mathbb{R}^n,$$

is the orthogonal projection of y on $\mathcal{I}m\{A\}$.

By the PROJECTION THEOREM

$$\langle y - \hat{y}, Ax \rangle_{\mathcal{H}} = 0 \quad \forall x \in \mathbb{R}^n.$$

Therefore, since $\hat{y} \in \mathcal{M} = \mathcal{I}m\{A\}$ and writing $\hat{y} = A\hat{x}$ for some $\hat{x} \in \mathbb{R}^n$,

$$x^T A^T W (y - A\hat{x}) = 0 \quad \forall x \in \mathbb{R}^n.$$

It follows

$$A^T W (y - A\hat{x}) = 0$$

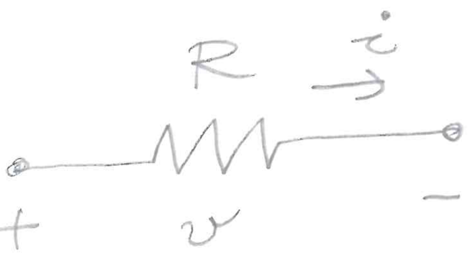
$$\Rightarrow \hat{x} = A_W^+ y$$

$$\Rightarrow \hat{y} = AA_W^+ y \quad \text{and}$$

AA_W^+ is the orthogonal projection matrix on $\mathcal{I}m\{A\}$.

EXAMPLE. Assume that we want to estimate the resistance $R > 0$ of an electrical component using the measurements of current i and voltage v on the component itself. The deterministic relation between i and v is:

$$v = Ri :$$



If we take m measurements of (v, i) we have

$$v_1 = R i_1 + d_1$$

$$\vdots$$

$$v_m = R i_m + d_m$$

where $d_1, \dots, d_m$ are possible disturbances or uncertainty introduced by the measurent device (sensor).

If $i^{(m)} = \begin{pmatrix} i_1 \\ \vdots \\ i_m \end{pmatrix}$ and $v^{(m)} = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}$

and we assign the same level of "confidence" to each measurement (v_i, i_i), $i = 1, \dots, m$, in the sense that we consider each measurement (v_i, i_i) to have the same "precision", then we can implement a least square estimate of R with $W = I$. In other words, we weight all the measurements (v_i, i_i) in the same way. Therefore, our estimate $\hat{R}$ of R is, as long as $i^{(m)} \neq 0$,

$$\boxed{\hat{R} = i^{(m)\#} v^{(m)}}$$

where $i^{(m)\#} = \left(i^{(m)T} i^{(m)}\right)^{-1} i^{(m)T}$.

Notice that

$$\hat{R} = i^{(m)\#} v^{(m)} = \\ = \frac{\sum_{j=1}^{m} i_j v_j}{\sum_{j=1}^{m} i_j^2}$$

Moreover, it is important to repeat the measurements (v_i, i_i) for a number m of times such that $i^{(m)} \neq 0$.

The "precision" of a measurement (v_i, i_i) is characterised by the error introduced by the disturbance d_i. Smaller d_i corresponds to a more precise measurement. In this case, it should be more convenient to implement a weighted least square estimate $\hat{R}$ of R, i.e. with $W \neq I$.

With the stochastic approach to the estimation process we take the noise directly into account and the variables Y, X, D characterising some deterministic relation

$$Y = AX + D \,, \quad \begin{cases} X \in \mathbb{R}^n \\ Y \in \mathbb{R}^m \\ D \in \mathbb{R}^m, \; m > n \end{cases}$$

are random vectors, Y being the measurements, X the variables to be estimated and D the observation noise. The variable D is known through its density $p_D(d)$. However, X can be a deterministic variable as well, according to the a priori information we have on it. As a random vector, it is usually known through its density $p_X(x)$.

A first important step in the estimation process of X is to evaluate the density $p_Y(y)$ (if X is deterministic) or $p_{Y|X}(y,x)$ (if X is a random vector).

CASE A (X deterministic)

In this case the density $p_Y(y)$ is a function of the values x of X and more precisely we denote it by $p_Y(y,x)$. Moreover, the model equation is

$$Y = AX + D$$

and D has density $p_D(d)$. We use now the following result:

FACT. Given two random vectors $D \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$, with a measurable $f: \mathbb{R}^m \to \mathbb{R}^m$, f is invertible and differentiable over its domain, $p_D(d)$ and $p_Y(y)$ the densities of D and Y, respectevely, if

$$Y = f(D)$$

then

$$\boxed{p_Y(y) = p_D(f^{-1}(y)) \left| \det \frac{\partial f^{-1}}{\partial y}(y) \right|}.$$

Notice that since X is deterministic:

$$Y = AX + D \triangleq f(D)$$

with f invertible and differentiable.

Therefore

$$p_Y(y) = p_D(f^{-1}(y)) \left| \det \frac{\partial f^{-1}}{\partial y}(y) \right|$$

$$= p_D(y - Ax) \left| \det I \right|$$

$$= \underbrace{p_D(y - Ax)}_{\text{a priori information}}$$

since $f^{-1}(Y) = Y - AX$ and x are the values of X.

CASE B. (X random)

In this case, the conditional density $p_{Y|X}(y,x)$ is a function of the values x of X. In a similar way as in case A, we obtain

$$\underbrace{p_{Y|X}(y,x) = \overbrace{p_D}^{\text{a priori information}}(y-Ax)}_{\text{a posteriori information}}$$

Next, by using the Bayes formulas we also obtain $p_{X|Y}(x,y)$ as:

$$p_{X|Y}(x,y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

where

$$p_{X,Y}(x,y) = \underbrace{p_X(x)}_{\text{a priori information}} \overbrace{p_{Y|X}(y,x)}^{\text{a posteriori information}}$$

and

$$p_Y(y) = \int_{\mathbb{R}^m} p_{X,Y}(x,y)\,dx$$

$$= \int_{\mathbb{R}^m} p_X(x)\, p_{Y|X}(y,x)\,dx .$$

With $p_Y(y, x)$ (for deterministic X) and $p_{X/Y}(x, y)$ (for random X) at hand, we proceed in the cases A or B as follows:

CASE A (X deterministic).

From

$$p_Y(y, x) = p_D(y - Ax)$$

we try to maximize its value by choosing a suitable $x = \hat{x}$. The choice of $\hat{x}$ must be done in some admissibility set $\mathcal{D} \subseteq \mathbb{R}^n$. Therefore,

$$\hat{x} = \underset{x \in \mathcal{D} \subseteq \mathbb{R}^n}{\operatorname{argmax}} \; p_Y(y, x)$$

where y are meant as the numerical values of Y, the measurements vector. This $\hat{x}$ is known as MAXIMUM LIKELIHOOD estimate of X. This estimates takes its name by the fact that it is the value of x which maximizes the probability of $Y(\omega)$ being equal to y.

CASE B. (X random)

In this case, a common practice is to choose $\hat{X}$ in such a way to minimize the variance of the estimation error $X - \hat{X}$. Notice that this variance can be calculated from

$$p_{X,Y}(x,y) = p_X(x)\, p_{Y|X}(y,x)$$

and considering that $\hat{X}$ is equal to $f(Y)$ for some measurable f:

$$\sigma^2_{X-\hat{X}} = \int_{\mathbb{R}^n} (x - f(y))^T (x - f(y))\, p_{X,Y}(x,y)\, dx\, dy$$

This function $f(y)$ gives the estimate of X with MINIMUM ERROR VARIANCE and, as will be shown next, it is

$$\boxed{\begin{aligned} \hat{x} &= f(y) = E\{X|Y\}\Big|_{Y=y} \\ &= \int_{\mathbb{R}^n} x\, p_{X|Y}(x,y)\, dx \end{aligned}}$$

1. ESTIMATES WITH MINIMUM (ERROR) VARIANCE

We want to characterize estimates $\hat{X}$ of a random vector X which minimize the variance of $X-\hat{X}$:

$$\boxed{J(\hat{X}) = E\{\|X(\omega)-\hat{X}(\omega)\|^2\}}$$

It is good practice to consider only "centered" candidates for $\hat{X}(\omega)$, which is

$$\boxed{E\{\hat{X}(\omega)\} = E\{X(\omega)\}}$$

so that $J(\hat{X})$ is the variance of the estimation error $\mathcal{E}(\omega) = X(\omega)-\hat{X}(\omega)$.

Indeed, if $\hat{X}(\omega)$ is not centered the new estimate $\hat{X}'(\omega) = \hat{X}(\omega)+\gamma$, $\gamma \triangleq E\{X(\omega)-\hat{X}(\omega)\}$, is centered

and

$$J(\hat{X}') = J(\hat{X}) - \gamma^T \gamma \leq J(\hat{X})$$

so that $\hat{X}'$ is better than $\hat{X}$ and it is centered.

Any estimate $\hat{X}(\omega)$ is considered as the result of a measurable function $\hat{h}(\cdot)$ of the measurements vector $Y(\omega)$:

$$\hat{X}(\omega) = \hat{h}(Y(\omega)) .$$

Our optimal problem is formulated as follows:

$$\hat{X}(\omega) = \hat{h}(Y(\omega)) = \underset{\substack{h:\mathbb{R}^m \to \mathbb{R}^n \\ \text{measurable}}}{\operatorname{argmin}} J(h(Y(\omega)))$$

THEOREM

$$\boxed{\underset{\substack{h:\mathbb{R}^m\to\mathbb{R}^n \\ \text{measurable}}}{\operatorname{argmin}} J(h(Y(\omega))) = E\{X(\omega) \mid \mathcal{F}^Y\}}$$

Proof. Rewrite $J(\tilde{X})$, $\tilde{X} = h(Y)$, as

$$J(\tilde{X}) = E\{\|X - \tilde{X}\|^2\} = E\{\|X - \hat{X} + \hat{X} - \tilde{X}\|^2\},$$

where $\hat{X} \triangleq E\{X \mid \mathcal{F}^Y\}$. Recall that

$$< X - \hat{X}, Z >_{\mathcal{L}_2} = 0$$

$\forall$ $\mathcal{F}^Y$-measurable Z, by the projection theorem. But

$$\hat{X} - \tilde{X} = E\{X \mid \mathcal{F}^Y\} - h(Y)$$
$$= f(Y) - h(Y)$$

for some measurable $f : \mathbb{R}^m \to \mathbb{R}^n$, so that $\hat{X} - \tilde{X}$ is $\mathcal{F}^Y$-measurable.

Therefore

$$J(\tilde{X}) = E\{\|X - \hat{X} + \hat{X} - \tilde{X}\|^2\}$$
$$= E\{\|X - \hat{X}\|^2\} + E\{\|\hat{X} - \tilde{X}\|^2\}$$
$$\geqslant E\{\|X - \hat{X}\|^2\} = J(\hat{X})$$

which proves the main result. $\blacktriangle$

We want to see now that $\hat{X} \triangleq E\{X | \mathcal{Y}\}$ minimizes also the error covariance $\Psi_{\hat{\varepsilon}}$, $\hat{\varepsilon} \triangleq X - \hat{X}$. We have

$$\Psi_{\tilde{\varepsilon}} \triangleq E\{(X - \tilde{X})(X - \tilde{X})^T\}$$
$$= E\{(X - \hat{X} + \hat{X} - \tilde{X})(X - \hat{X} + \hat{X} - \tilde{X})^T\}.$$

If $\Psi' \triangleq E\{(\hat{X} - \tilde{X})(\hat{X} - \tilde{X})^T\}$

then $\Psi_{\tilde{\varepsilon}} = \Psi_{\hat{\varepsilon}} + \Psi' + E\{(X - \hat{X})(\hat{X} - \tilde{X})^T\}$
$$+ E\{(\hat{X} - \tilde{X})(X - \hat{X})^T\}$$

But

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

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

$$= E\{E\{X(\hat{X}-\tilde{X})^T \mid \mathcal{F}^Y\}\} - E\{\hat{X}(\hat{X}-\tilde{X})^T\}$$

$$= E\{E\{X \mid \mathcal{F}^Y\}(\hat{X}-\tilde{X})^T\} - E\{\hat{X}(\hat{X}-\tilde{X})^T\}$$

since $\hat{X}-\tilde{X}$ is $\mathcal{F}^Y$-measurable.

Finally

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

$$= E\{E\{X \mid \mathcal{F}^Y\}(\hat{X}-\tilde{X})^T\} - E\{\hat{X}(\hat{X}-\tilde{X})^T\}$$

$$= E\{\hat{X}(\hat{X}-\tilde{X})^T\} - E\{\hat{X}(\hat{X}-\tilde{X})^T\} = 0 .$$

and

$$\Psi_{\tilde{\varepsilon}} = \Psi_{\hat{\varepsilon}} + \Psi' .$$

But $\Psi' \geq 0$ so that

$$\Psi_{\tilde{\varepsilon}} \geq \Psi_{\hat{\varepsilon}}$$

(where $A \geq B$ means $A - B \geq 0$) ◀

REMARK.

$$J(\tilde{X}) = \mathrm{Tr}\,\Psi_{\tilde{\varepsilon}}$$
$$= \mathrm{Tr}\,E\{(X-\tilde{X})(X-\tilde{X})\}$$
$$= E\{\|X-\tilde{X}\|^2\}$$

REMARK. If Z is another random vector such that $\mathcal{F}^Z = \mathcal{F}^Y$, the optimal estimate is

$$\hat{X} = E\{X|\mathcal{F}^Y\} = E\{X|\mathcal{F}^Z\}.$$

If $\mathcal{F}^X$ and $\mathcal{F}^Y$ are independent then

$$\hat{X} = E\{X|\mathcal{F}^Y\} = E\{X\}.$$

If no observations are available, i.e. $\mathcal{F}^Y = \mathcal{F}_m \triangleq \{\phi, \Omega\}$, then

$$\hat{X} = E\{X|\mathcal{F}^Y\} = E\{X\}$$

2. CALCULUS OF ESTIMATES WITH MINIMUM VARIANCE UNDER GAUSSIAN NOISE

Let $Z = (X^T \; Y^T)^T$ be a gaussian vector, $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. We know that also X and Y are gaussian vectors. Assume $E\{Z\} \triangleq m_Z = 0$ (otherwise redefine Z as $Z - m_Z \triangleq \tilde{Z}$).

We have

$$E\{Z\} = \begin{pmatrix} E\{X\} \\ E\{Y\} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Psi_Z = E\left\{ \begin{pmatrix} X \\ Y \end{pmatrix} (X^T Y^T) \right\}$$

$$= \begin{pmatrix} \Psi_X & \Psi_{XY} \\ \Psi_{YX} & \Psi_Y \end{pmatrix}.$$

We know that the estimate $\hat{X}$ with minimum error variance is

$$\hat{X} = E\{X \mid \mathcal{F}^Y\}$$

By $\hat{x}$, ~~x~~ and y we denote the numerical values of $\hat{X}(\omega)$, $X(\omega)$ and $Y(\omega)$, respectively:

$$\hat{x} = E\{X | \mathcal{F}^Y\}_{Y=y}$$

$$= E\{X | Y=y\}.$$

Recall that

$$\hat{x} = E\{X | Y=y\} = \int_{\mathbb{R}^n} x \, p_{X|Y}(x, y) \, dx$$

$$= f(y)$$

for some measurable $f(\cdot)$. We want to show that $f(y)$ is linear:

$$\boxed{f(y) = Ky \quad \text{for some matrix } K}$$

Using Bayes theorem :

$$\hat{x} = \int_{\mathbb{R}^n} x \underbrace{\boxed{\frac{p_{X,Y}(x,y)}{p_Y(y)}}}_{= p_{X|Y}(x,y)} dx$$

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

$$= \frac{\int_{\mathbb{R}^n} x \, p_{X,Y}(x,y)\, dx}{\int_{\mathbb{R}^n} p_{X,Y}(x,y)\, dx}$$

But $p_{X,Y}(x,y)$ is gaussian :

$$p_{X,Y}(x,y) = \frac{1}{(2\pi)^{\frac{n+m}{2}} (\det \Psi_Z)^{1/2}} e^{-\frac{1}{2}(x^T y^T)\Psi_Z^{-1}\binom{x}{y}}$$

Define

$$\Psi_Z^{-1} \triangleq \begin{pmatrix} \overline{\Psi}_X & \overline{\Psi}_{XY} \\ \overline{\Psi}_{XY}^T & \overline{\Psi}_Y \end{pmatrix}$$

where since $\Psi_Z \Psi_Z^{-1} = 0$:

$$\left.\begin{aligned}
&\bar{\Psi}_X \Psi_X + \bar{\Psi}_{XY} \Psi_{XY}^T = I_n \\
&\bar{\Psi}_X \Psi_{XY} + \bar{\Psi}_{XY} \Psi_Y = 0 \\
&\Psi_{XY}^T \Psi_X + \bar{\Psi}_Y \Psi_{XY}^T = 0 \\
&\bar{\Psi}_{XY}^T \Psi_{XY} + \bar{\Psi}_Y \Psi_Y = I_m
\end{aligned}\right\} \quad (R)$$

With these notations

$$(x^T \ y^T)\Psi_Z^{-1}\begin{pmatrix} x \\ y \end{pmatrix} =$$

$$x^T \bar{\Psi}_X x + 2x^T \bar{\Psi}_{XY} y + y^T \bar{\Psi}_Y y$$

$$= (x - My)^T \Psi_1^{-1} (x - My) + y^T \Psi_2^{-1} y$$

where

$$\Psi_1^{-1} \triangleq \bar{\Psi}_X , \quad M \triangleq -\bar{\Psi}_X^{-1} \bar{\Psi}_{XY} ,$$

$$\Psi_2^{-1} \triangleq \bar{\Psi}_Y - M^T \bar{\Psi}_X M \qquad (S)$$

$$= \bar{\Psi}_Y - \bar{\Psi}_{XY}^T \bar{\Psi}_X^{-1} \bar{\Psi}_{XY} .$$

Noticing that the second and fourth relations in (R) are

$$\begin{cases} \Psi_{XY}\Psi_Y^{-1} = -\overline{\Psi}_X^{-1}\overline{\Psi}_{XY} \\ \overline{\Psi}_Y - \overline{\Psi}_{XY}^T\overline{\Psi}_X^{-1}\overline{\Psi}_{XY} = \Psi_Y^{-1} \end{cases}$$

and (S) can be rewritten as

$$\begin{cases} \Psi_1 = \overline{\Psi}_X^{-1} \\ M = \Psi_{XY}\Psi_Y^{-1} \\ \Psi_2 = \Psi_Y . \end{cases}$$

Using these relations

$$f_{X,Y}(x,y) = \frac{1}{(2\pi)^{\frac{n+m}{2}}(\det\Psi_Z)^{1/2}} \cdot e^{-\frac{1}{2}(x-My)^T\overline{\Psi}_X(x-My)}e^{-\frac{1}{2}y^T\Psi_Y^{-1}y}$$

and

$$\hat{x} = \frac{\int_{\mathbb{R}^n} x e^{-\frac{1}{2}(x-My)^T \overline{\Psi}_X (x-My)} dx}{\int_{\mathbb{R}^n} e^{-\frac{1}{2}(x-My)^T \overline{\Psi}_X (x-My)} dx}$$

$$= \frac{\frac{1}{(2\pi)^{n/2}(\det \overline{\Psi}_X^{-1})^{1/2}} \int_{\mathbb{R}^n} x e^{-\frac{1}{2}(x-My)^T \overline{\Psi}_X (x-My)} dx}{\frac{1}{(2\pi)^{n/2}(\det \overline{\Psi}_X^{-1})^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}(x-My)^T \overline{\Psi}_X (x-My)} dx}$$

But

$$\frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{(\det \overline{\Psi}_X^{-1})^{1/2}} e^{-\frac{1}{2}(x-My)^T \overline{\Psi}_X (x-My)}$$

is a gaussian density with mean My and covariance $\overline{\Psi}_X^{-1}$. It follows

$$\boxed{\begin{aligned} \hat{x} &= E\{X \mid Y = y\} = My \\ &= \Psi_{XY} \Psi_Y^{-1} y \end{aligned}}$$

The error covariance is $\overline{\Psi}_X^{-1}$ and from the second in (R) we have $\bar{\Psi}_{XY} = -\bar{\Psi}_X \Psi_{XY} \Psi_Y^{-1}$ and replacing in the first of (R):

$$\boxed{\bar{\Psi}_X^{-1} = \Psi_X - \Psi_{XY} \Psi_Y^{-1} \Psi_{XY}^T}$$

REMARK If $m_X \neq 0$ and $m_Y \neq 0$:

$$\boxed{\hat{x} = m_X + \Psi_{XY} \Psi_Y^{-1} (y - m_Y)}$$

REMARK If X and Y are uncorrelated then $\Psi_{XY} = 0$ and

$$\boxed{\begin{aligned} \hat{x} &= m_X \\ \bar{\Psi}_X^{-1} &= \Psi_X \end{aligned}}$$

Notice that using the correlation between X and Y we obtain a value of the error covariance $\bar{\Psi}_X^{-1}$ which is lower than Ψ_X, since $\Psi_{XY} \Psi_Y^{-1} \Psi_{XY}^T \geq 0$.

REMARK. By the projection theorem, the unique measurable function $f(\cdot)$ for which $X - f(Y) \perp h(Y)$ for any other measurable function $h(\cdot)$ is $E\{X|Y\}$. It can be directly checked that $E\{X|Y\} = \Psi_{XY}\Psi_Y^{-1}Y$.

Indeed,

$$E\{(X - \Psi_{XY}\Psi_Y^{-1}Y)Y^T\} = E\{XY^T\}$$
$$- \Psi_{XY}\Psi_Y^{-1}E\{YY^T\} = 0$$

and this implies that $X - \Psi_{XY}\Psi_Y^{-1}Y$ and Y are uncorrelated. But Y and $X - \Psi_{XY}\Psi_Y^{-1}Y$ are (jointly) gaussian that they are also independent.

Therefore for any measurable $h(\cdot)$:

$$E\{(X - \Psi_{XY}\Psi_Y^{-1}Y)\,h(Y)\} =$$
$$E\{X - \Psi_{XY}\Psi_Y^{-1}Y\}\,E\{h(Y)\} = 0 \Rightarrow$$
$$X - \Psi_{XY}\Psi_Y^{-1}Y \perp h(Y)$$

3. ESTIMATES WITH MINIMUM VARIANCE UNDER NON-GAUSSIAN NOISE

Consider X, Y non-gaussian, with zero mean. We will look for an estimate of X having the form

$$\tilde{X} = KY, \quad K \in \mathbb{R}^{r \times m}$$

for which the error variance is minimum. Without loss of generality, as we have seen in the previous chapter, we can minimize as well the error covariance

$$\boxed{J(K) = E\{(X-KY)(X-KY)^T\}}$$

Our problem is to find

$$K^* = \underset{K \in \mathbb{R}^{r \times m}}{\arg\min}\, J(K)$$

We have

$$J(K) = E\{XX^T\} - KE\{YX^T\}$$
$$- E\{XY^T\}K^T + KE\{YY^T\}K^T$$
$$= \Psi_X - K\Psi_{YX} - \Psi_{XY}K^T + K\Psi_Y K^T.$$

To obtain necessary conditions for K^*, we will write the Taylor series of $J(K)$ around K^*:

$$J(K^* + \Delta) = \Psi_X - (K^* + \Delta)\Psi_{YX}$$
$$- \Psi_{XY}(K^* + \Delta)^T + (K^* + \Delta)\Psi_Y(K^* + \Delta)^T$$
$$= J(K^*) - \Delta(-\Psi_{YX} + \Psi_Y K^{*T})$$
$$+ (-\Psi_{XY} + K^*\Psi_Y)\Delta^T + O(\|\Delta\|^2)$$

Therefore, K^* must satisfy

$$-\Psi_{XY} + K^*\Psi_Y = 0 \Rightarrow$$
$$K^* = \Psi_{XY}\Psi_Y^{-1}$$

and

$$\left\{\begin{aligned} \tilde{X} &= K^* Y = \Psi_{XY}\Psi_Y^{-1} \\ J(K^*) &= \Psi_X - K^*\Psi_{YX} \\ &= \Psi_X - \Psi_{XY}\Psi_Y^{-1}\Psi_{YX} \end{aligned}\right.$$

The estimate $\tilde{X}$ is the LINEAR estimate which minimises the error variance. Notice that $J(K)$ can be also rewritten as

$$J(K) = J(K^*) + \underbrace{\frac{1}{2}(K-K^*)\Psi_V(K-K^*)^T}_{\left(=0 \text{ iff } K=K^*\right)}$$

$\forall\, K$.