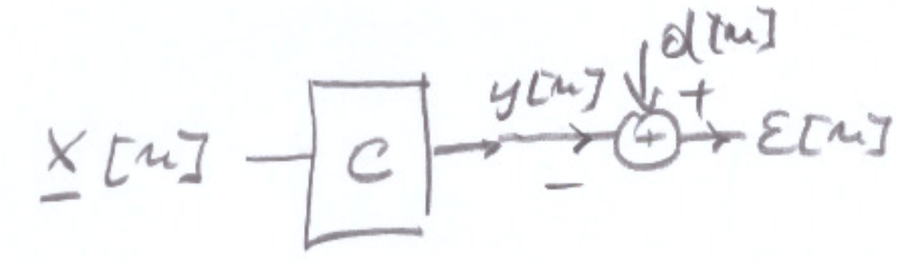
FILTRI A MINIMO ERRORE QUADRATICO (cont.)

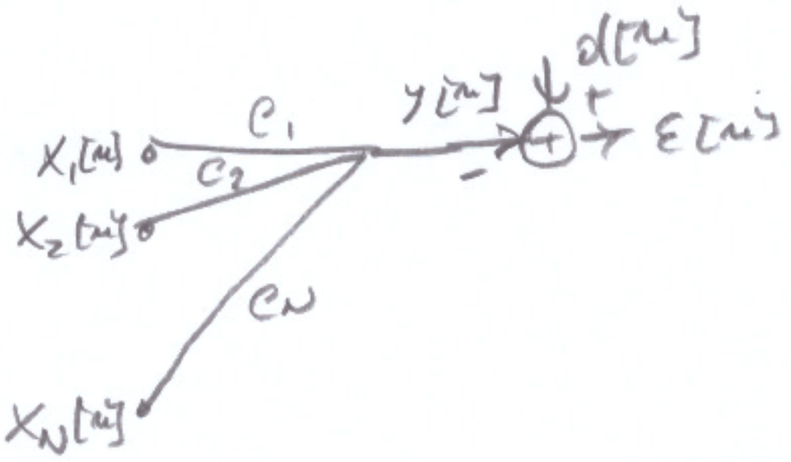
Lezioni di Telecomunicazioni.
Prof. FRANCESCO A.N. PALMIERI

CORSO DI "SIGNAL PROCESSING AND DATA FUSION"
AA 2022-23

CASO VETTORIALE GENERICO

Il paradigma non cambia se il vettore $\underline{x}[n]$ proviene da un processo o sequenza vettoriale (e quindi non necessariamente una finestra che scorre su un segnale)





la soluzione è ovviamente

$$\begin{cases} \underline{c}^o = \underline{\underline{R}}_x^{-1} \underline{r}_{dx} \\ \mathcal{E}(c^o) = \mathcal{E}_d - \underline{r}_{dx}^T \underline{\underline{R}}_x^{-1} \underline{r}_{dx} \end{cases} \qquad \begin{cases} \underline{\underline{R}}_x = E[\underline{x}[n]\underline{x}^T[n]] \\ \underline{r}_{dx} = E[d[n]\underline{x}[n]] \\ \mathcal{E}_d = E[d^2[n]] \end{cases}$$

TRAINING SET

$\{(d[n], \underline{x}[n]),\ n = 1, \dots, L\}$

$\Downarrow$

Ottieni stime di $\underline{\underline{R}}_x$ e $\underline{r}_{dx}$

$$\left(\begin{array}{l} \underline{\underline{R}}_x = \frac{1}{L}\sum_{n=1}^{L} \underline{x}[n]\underline{x}^T[n] \\ \underline{r}_{dx} = \frac{1}{L}\sum_{n=1}^{L} \underline{x}[n]\, d[n] \\ \mathcal{E}_d = \frac{1}{L}\sum_{n=1}^{L} d[n] \end{array} \right)$$

$\Downarrow$

Risolvi $\underline{c}^o = \underline{\underline{R}}_x^{-1} \underline{r}_{dx}$

(*) $\mathcal{E}(c^o) = \mathcal{E}_d - \underline{r}_{dx}^T \underline{\underline{R}}_x^{-1} \underline{r}_{dx}$

le prestazioni sul training set sono quelle ottenute stimando $\underline{\underline{R}}_x$ $\underline{r}_{dx}$ e $\mathcal{E}_d$ dal training set.

VALIDATION SET

le prestazioni sul validation set sono ancora date dalla formula (*) ma con $R_x, r_{dx}, \mathcal{E}_d$ stimate dal validation set.

CORRELATIONS OR COVARIANCES AND THE ADDITION OF A BIAS

In solving our LinAspures problems in considering the cost

$$E[(d-y)^2]$$

we have ignored an important aspect of the problem. It may happen that d and/or $\underline{x}$ have non-zero means. Therefore $y = \underline{c}^T\underline{x}$ will have a non-zero mean and the defined cost may be strongly dominated by a search by the filter to match the mean of d ("mean chasing")

Also in considering dependence among variables we should use covariances

$$E[(\underline{x}-\mu_x)(y-\mu_y)]$$ and not correlations

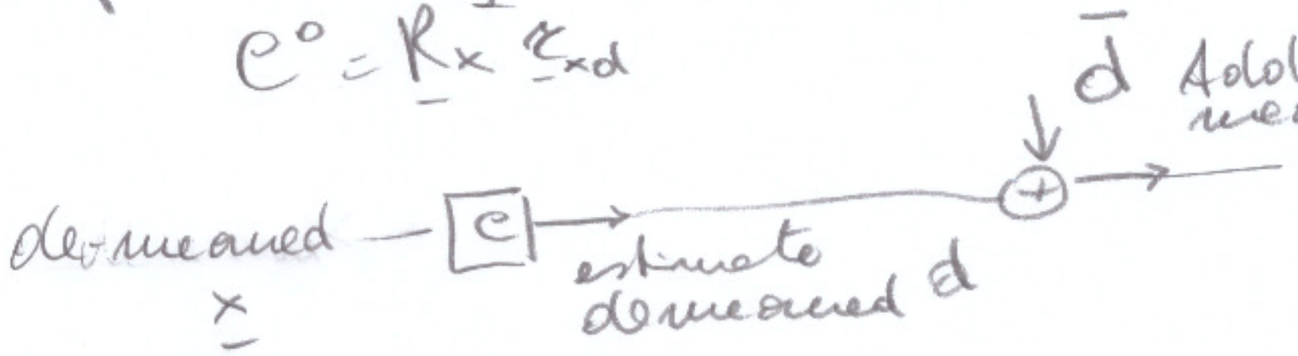
$E[\underline{x}y]$. In practice this is avoided removing the means before posing the problem

$$d \leftarrow d - E[d]$$
$$\underline{x} \leftarrow \underline{x} - E[\underline{x}]$$

The mean of d can be added at the end after solving the problem using correlations

i.e. if $\underline{x}$ and d are "de-meaned"

$$\underline{c}^o = \underline{R}_x^{-1}\underline{r}_{xd}$$



To avoid the issue of the means it is very common to generalize our linear systems into <u>affine systems</u> with the addition of a bias

$$y[n] = \underline{c}^T \underline{x}[n] + b = [c_1 c_2 \dots c_N] \begin{bmatrix} x_1[n] \\ \vdots \\ x_N[n] \end{bmatrix} + b$$

(Note that an affine system is not linear because it does not satisfy the superposition principle

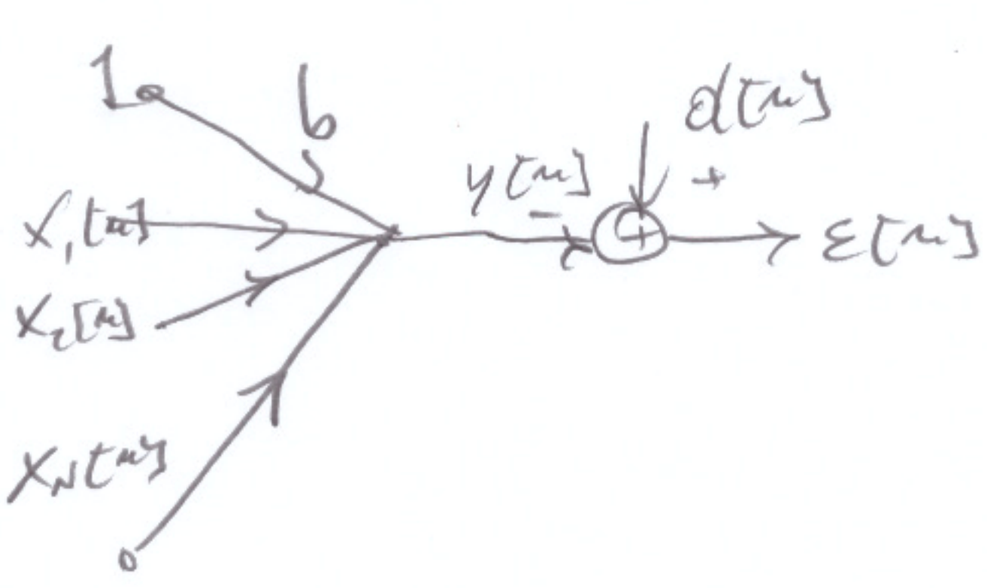
$y_1 = c_1 x_1 + b \quad y_2 = c_2 x_2 + b$

$x_1 + x_2 \rightarrow c(x_1 + x_2) + b \neq y_1 + y_2$)

The constant b can be easily included in the linear system by augmenting the vector $\underline{x}[n]$ with a constant (= 1).

$$y[n] = [c_1 \dots c_N] \begin{bmatrix} x_1[n] \\ x_N[n] \end{bmatrix} + b = [b\ c_1 \dots c_N] \underbrace{\begin{bmatrix} 1 \\ x_1[n] \\ \vdots \\ x_N[n] \end{bmatrix}}_{\underline{x}_a[n]}$$

$\underline{x}_a[n]$ is an <u>augmented vector</u> and the problem becomes linear



$$\begin{cases} \underline{c}^{o} = \underline{\underline{R}}_{x_a}^{-1} \underline{r}_{dx_a} \\ \mathcal{E}(e^{o}) = \mathcal{E}_d - \underline{r}_{dx_a}^{T} \underline{\underline{R}}_{x_a}^{-1} \underline{r}_{dx_a} \end{cases}$$

$$\underline{\underline{R}}_{x_a} = E\left[\underline{x}_a[n]\,\underline{x}_a^{T}[n]\right] = E\left[\begin{bmatrix} 1 \\ x_1[n] \\ \vdots \\ x_N[n] \end{bmatrix} \begin{bmatrix} 1 & x_1[n] & \cdots & x_N[n] \end{bmatrix}\right]$$

$$= \begin{bmatrix} 1 & E[x_1[n]] & \cdots & E[x_N[n]] \\ E[x_1[n]] & & & \\ \vdots & & E[\underline{x}[n]\underline{x}^{T}[n]] & \\ E[x_N[n]] & & & \end{bmatrix}$$

↑ Note that the means are automatically included

$$\underline{r}_{dx_a} = E\left[d[n]\,\underline{x}_a[n]\right] = E\left[d[n]\begin{bmatrix} 1 \\ x_1[n] \\ \vdots \\ x_N[n] \end{bmatrix}\right] = \begin{bmatrix} E[d[n]] \\ E[d[n]x[n]] \\ \vdots \\ E[d[n]x[n]] \end{bmatrix} = \begin{bmatrix} \\ \\ \underline{r}_{dx} \\ \\ \\ \end{bmatrix}$$

Therefore the augmented version takes into account the means implicitly and can be treated without the effect of "closing the mean" explained above, using correlations.

Come verifica ulteriore dopo aver introdotto il bias, ovvero

$$y[n] = \underline{c}^T \underline{x}[n] + b$$

l'errore quadratico diventa

$$\mathcal{E}(\underline{c}, b) = E\left[\left(d[n] - (\underline{c}^T \underline{x}[n] + b)\right)^2\right]$$

Derivando rispetto a b, abbiamo

$$\frac{\partial \mathcal{E}(\underline{c}, b)}{\partial b} = E\left[2\left(d[n] - \underline{c}^T \underline{x}[n] - b\right)\right] = 0$$

Il valore ottimo di b è

$$b^o = E[d[n]] - \underline{c}^T E[\underline{x}[n]]$$

che riportato nell'espressione dell'errore medio fornisce:

$$\mathcal{E}(\underline{c}, b^o) = E\left[\left(d[n] - \underline{c}^T \underline{x}[n] - E[d[n]] + \underline{c}^T E[\underline{x}[n]]\right)^2\right]$$

$$\mathcal{E}(\underline{c}, b^o) = E\left[\left[\left(d[n] - E[d[n]]\right) - \underline{c}^T\left(\underline{x}[n] - E[\underline{x}[n]]\right)\right]\right]$$

$d[n] - E[d[n]]$: osservazione d

$\underline{x}[n] - E[\underline{x}[n]]$: osservazione $\underline{x}[n]$

CASO SINGOLARE

Quando la matrice $\underline{\underline{R}}_x$ è singolare, è evidente che la soluzione $\underline{c}^o = \underline{\underline{R}}_x^{-1} \underline{r}_{dx}$ non può essere calcolata immediatamente.

Ricordiamo che $\underline{\underline{R}}_x = E[\underline{x}[n]\underline{x}^T[n]]$ è definita (semidefinita) positiva

[Prova Una matrice A è def. positiva se $\underline{\xi}^T \underline{\underline{A}} \underline{\xi} \geq 0 \ \forall \xi$

Nel nostro caso

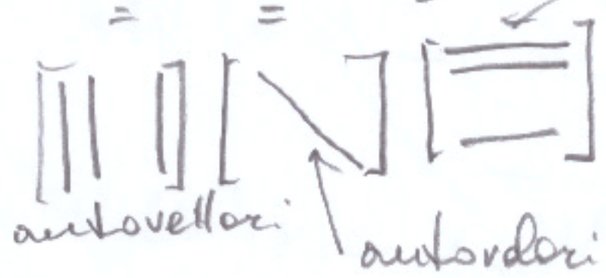
$$\xi^T E[\underline{x}[n]\underline{x}^T[n]]\,\xi$$

$$= E[\xi^T x[n] x^T[n] \xi] = E[(\xi^T x[n])^2] \geq 0$$

$\underline{R}_x$ è anche simmetrica, $R_x^T = E[x[n]x^T[n]] = \underline{\underline{R}}_x$]

Quindi ammette la seguente decomposizione.

$$R_x = \underline{\underline{W}}\ \underline{\underline{\Lambda}}_x\ \underline{\underline{W}}^T$$

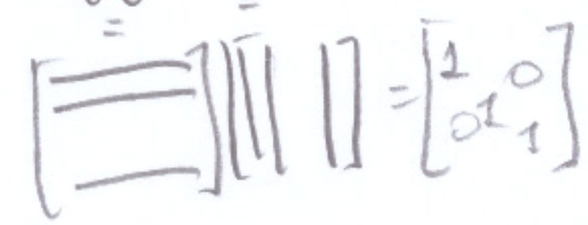


$\underline{\underline{\Lambda}}_x = \text{diag}(\lambda_1 \lambda_2 \ldots \lambda_N)$ Gli autovalori sono tutti non negativi. Supponiamo che essi siano ordinati in ordine decrescente

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_N$$

Gli autovettori sono "ortonormali", pertanto

$$\underline{\underline{W}}^T \underline{\underline{W}} = I_N$$



($\underline{\underline{W}}$ è una matrice unitaria.)

Poiché $\underline{\underline{W}}\,\underline{\underline{W}}^T\underline{\underline{W}} = \underline{\underline{W}}$

anche $\underline{\underline{W}}\,\underline{\underline{W}}^T = I_N$

Nel caso singolare, alcuni autovalori sono nulli, pertanto la decomposizione può essere visualizzata come segue ommettendo l'ordine

MSE.7

$$\underline{\underline{R}}_x = \begin{bmatrix} \underline{\underline{W}}_r & \underline{\underline{W}}_{N-r} \end{bmatrix} \begin{bmatrix} \underline{\underline{\Lambda}}_r & 0 \\ 0 & 0_{N-r} \end{bmatrix} \begin{bmatrix} \underline{\underline{W}}_r^T \\ \underline{\underline{W}}_{N-r}^T \end{bmatrix}$$

autovettori della spazio non nullo

autovettori dello spazio nullo

autovettori dello spazio non nullo

autovettori dello spazio nullo

Pertanto $\underline{x}[n]$ può essere proiettato nello spazio non nullo senza perdita di informazioni

$$\underline{x}_r = \underline{\underline{W}}_r^T \underline{x}$$

e lo stimatore lineare può essere applicato a $\underline{x}_r$. Lo scriviamo $\underline{\underline{R}}_{x_r} = E[\underline{x}_r[n]\underline{x}_r^T[n]]$ e $\underline{r}_{d\underline{x}_r} = E[d[n]\underline{x}_r[n]]$

Esplicitiamo

$$\underline{\underline{R}}_{x_r} = E[\underline{x}_r[n]\underline{x}_r^T[n]] = E[\underline{\underline{W}}_r^T \underline{x}[n]\underline{x}^T[n]\underline{\underline{W}}_r] = \underline{\underline{W}}_r^T \underbrace{E[\underline{x}[n]\underline{x}^T[n]]}_{\underline{\underline{R}}_x} \underline{\underline{W}}_r$$

$$= \underline{\underline{W}}_r^T \underline{\underline{W}} \underline{\underline{\Lambda}}_x \underline{\underline{W}}^T \underline{\underline{W}}_r =$$

$$\underline{\underline{W}}_r^T \begin{bmatrix} \underline{\underline{W}}_r & \underline{\underline{W}}_{N-r} \end{bmatrix} \begin{bmatrix} \underline{\underline{\Lambda}}_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{W}}_r^T \\ \underline{\underline{W}}_{N-r}^T \end{bmatrix} \underline{\underline{W}}_r = \begin{bmatrix} \underline{\underline{W}}_r^T \underline{\underline{W}}_r & \underline{\underline{W}}_r^T \underline{\underline{W}}_{N-r} \end{bmatrix} \begin{bmatrix} \underline{\underline{\Lambda}}_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{W}}_r^T \underline{\underline{W}}_r \\ \underline{\underline{W}}_{N-r}^T \underline{\underline{W}}_r \end{bmatrix}$$

$$\begin{bmatrix} \underline{\underline{I}}_r & \underline{\underline{0}}_{N-r} \end{bmatrix} \qquad \underline{\underline{I}}_r \qquad \underline{\underline{0}}_{N-r}$$

$$= \underline{\underline{\Lambda}}_r$$

Diagonale
↳ componenti di $\underline{x}_r$ scorrelate

Il vettore di cross-correlazione diventa MSE. 8

$$\underline{\gamma}_{dx_2} = E[d[n]\underline{x}_2[n]] = E[d[n]\,\underline{\underline{W}}_2^T\underline{x}[n]]$$

$$= E[\underline{W}_2^T\underline{x}[n]\,d[n]] = \underline{W}_2^T E[d[n]\underline{x}[n]]$$

$$= \underline{\underline{W}}_2^T\,\underline{\gamma}_{dx}$$

La soluzione per i coefficienti diventa

$$\underline{c}_2^o = \underline{\underline{R}}_{x_2}^{-1}\,\underline{\gamma}_{dx_2} = \underline{\underline{\Lambda}}_2^{-1}\,\underline{\underline{W}}_2^T\,\underline{\gamma}_{dx}$$

L'errore in condizioni di ottimalità è quindi

$$\mathcal{E}(\underline{c}_2^o) = \mathcal{E}_d - \underline{\gamma}_{dx_2}^T\,\underline{\underline{R}}_{x_2}^{-1}\,\underline{\gamma}_{dx_2} = \mathcal{E}_d - \underline{\gamma}_{dx}^T\,\underline{\underline{W}}_2\,\underline{\underline{\Lambda}}_2^{-1}\,\underline{\underline{W}}_2^T\,\underline{\gamma}_{dx}$$

— o —

THE ORTHOGONALITY PRINCIPLE

We have derived the solution to the minimum mean squared error using the gradient. There is another interesting way of obtaining the solution using orthogonality.

Recall that the equation we solve for optimum linear estimators is

$$\underline{\underline{R}}_x \underline{c}^o = \underline{r}_{dx}$$

$$E[\underline{x}\underline{x}^T]\,\underline{c}^o = E[D\underline{x}]$$

that can be written as

$$E[(D - \underline{c}^{oT}\underline{x})\underline{x}] = \underline{0}$$

In this form (NORMAL EQUATION) we expressed the orthogonality principle

optimal linear mean squared error estimates correspond to orthogonality between data and error.

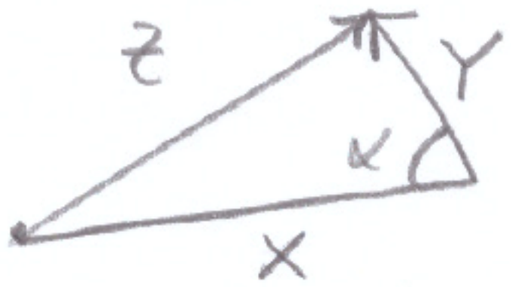
$$E[\varepsilon^o \underline{x}] = \underline{0}$$

We have also that the correlation between the error and the desired in optimal conditions is the minimum error

$$E[D(D - \underline{c}^{oT}\underline{x})] = E[D^2] - \underline{c}^{oT} E[D\underline{x}] = \varepsilon^o$$

VECTOR-SPACE INTERPRETATION

Random variables can be associated to vectors in an abstract vector space. The inner product is the correlation $E[XY]$. Orthogonality between X and Y is expressed as $E[XY]=0$. To see what this means, consider two vectors (random variables) X and Z. The vector difference $Y = Z - X$ is depicted as

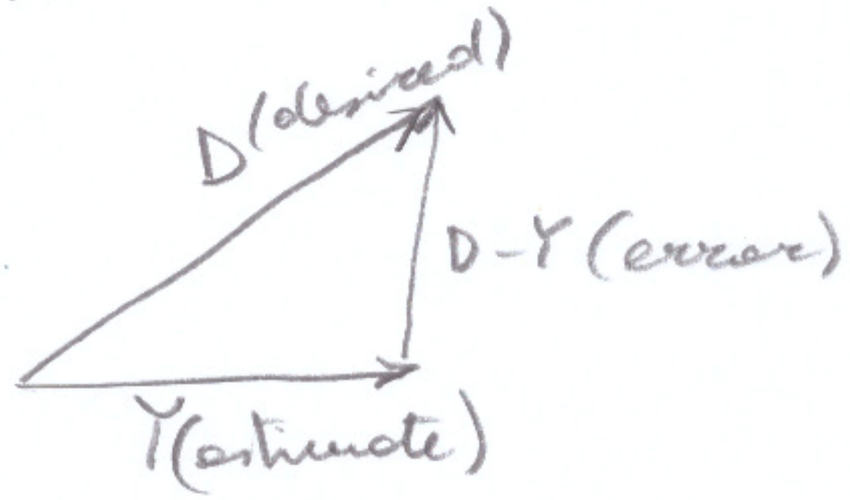


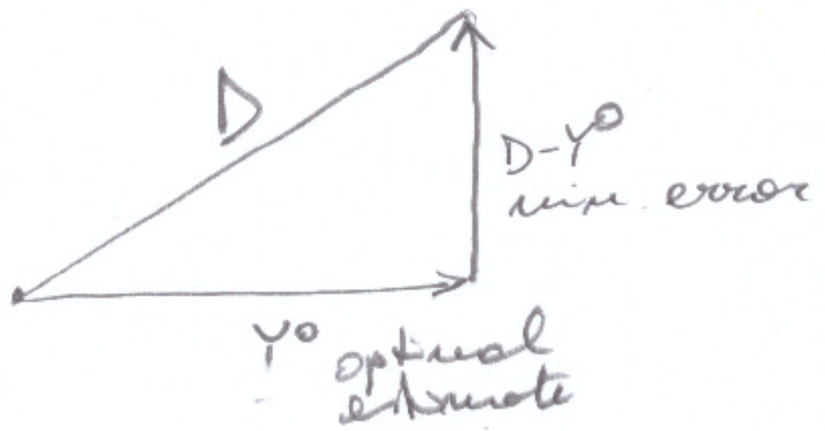
Now the pitagorean theorem applies if $\alpha = 90^0$, i.e. X and Y are ortogonal.

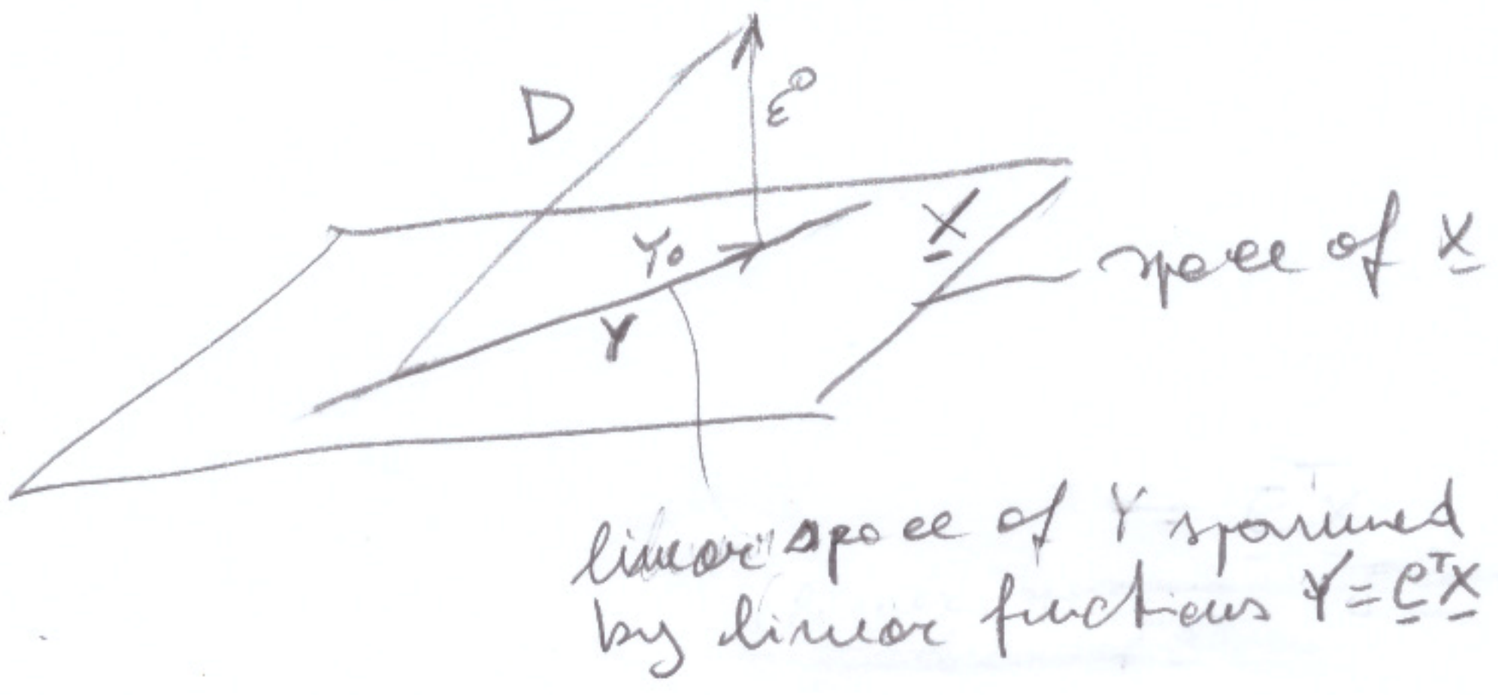
In fact. $E[Z^2] = E[(X+Y)^2] = E[X^2] + E[Y^2] + 2E[XY]$

only if $E[XY]=0$. The smallest difference corresponds to the vector orthogonal to X.

In the estimation case

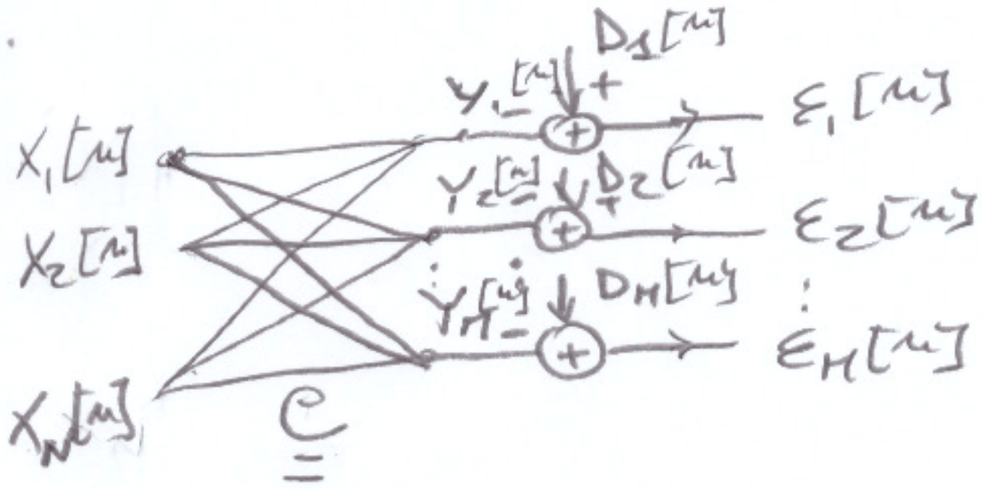






THE MATRIX CASE (PARALLEL FILTERS)

Using the same framework presented before, we could consider a bank of M filters (linear combiners) to be optimized in parallel.



$$\left[\underline{\underline{c}}_1\ \underline{\underline{c}}_2 \cdots \underline{\underline{c}}_M\right]$$

The cost function to be minimized can be taken to be the sum of all mean squared values

$$\mathcal{E}(\underline{\underline{C}}) = E\left[(\underline{D}-\underline{Y})^T(\underline{D}-\underline{Y})\right] = \sum_{i=1}^{M} E\left[(D_i-Y_i)^2\right] \quad (1)$$

Therefore we can optimize separately each filter to satisfy the normal equations

$$\underline{\underline{R}}_x \underline{c}_i^o = r_{dix} \qquad i = 1, \cdots, M$$

---

(1) Sometimes when more emphasis is given to one output with respect to the others, a more general cost could be used $E[(D-Y)^T\underline{\underline{Q}}(D-Y)]$ where $\underline{\underline{Q}}$ is symmetric and typically diagonal. We disregard this generality for now and take it to be $\underline{\underline{Q}} = \underline{\underline{I}}_M$.

The mean squared error for each filter can be written in one of the equivalent forms

$$\mathcal{E}_i^o = \mathcal{E}_{di} - \underline{r}_{dix}^T \underline{\underline{R}}_x^{-1} \underline{r}_{dix}$$

$$= \mathcal{E}_{di} - \underline{c}_i^{oT} \underline{r}_{dix}$$

$$= \mathcal{E}_{di} - \underline{r}_{diy}$$

Reorganizing into matrices

$$\underline{\underline{R}}_x [\underline{c}_1^o \dots \underline{c}_M^o] = [\underline{r}_{d_1x} \; \underline{r}_{d_2x} \dots \underline{r}_{d_Mx}]$$

$$E[\underline{x}\underline{x}^T \underline{c}^o] = E[D_1 \underline{x} \; D_2 \underline{x} \dots D_M \underline{x}]$$

$$E[\underline{x}\underline{x}^T \underline{\underline{C}}^o] = E[\underline{x}\underline{D}^T] \qquad \underline{\underline{R}}_{xD} \triangleq E[\underline{x}\underline{D}^T]$$

$$\boxed{\underline{\underline{R}}_x \underline{\underline{C}}^o = \underline{\underline{R}}_{xD}} \Rightarrow \underline{\underline{C}}^o = \underline{\underline{R}}_x^{-1} \underline{\underline{R}}_{xD}$$

The total squared error in optimal condition is

$$\mathcal{E}^o = \sum_{i=1}^{M} \mathcal{E}_i^o = \sum_{i=1}^{N} \mathcal{E}_{di} - \sum_{i=1}^{M} \underline{r}_{dix}^T \underline{\underline{R}}_x^{-1} \underline{r}_{dix}$$

Note that, even if the optimization for each filters is computed separately, the errors may be dependent also because $D_1 D_2 \dots D_M$ may be dependent and have a correlation $\underline{\underline{R}}_D = E[\underline{D}\underline{D}^T]$

Therefore it may be useful to consider the error correlation matrix

$$\underline{\underline{R}}_\varepsilon = E[\underline{\varepsilon}\,\underline{\varepsilon}^T] = E[(\underline{D} - \underline{\underline{C}}^T \underline{x})(\underline{D} - \underline{\underline{C}}^T \underline{x})^T]$$

M × M M × N N

that in optimal conditions is

$$\underline{\underline{R}}_{\varepsilon^o} = E[\underline{\varepsilon}^o \underline{\varepsilon}^{oT}] = E[(\underline{D} - \underline{\underline{C}}^{oT}\underline{x})(\underline{D} - \underline{\underline{C}}^{oT}\underline{x})^T]$$

$$= E[\underline{D}\underline{D}^T] + \underline{\underline{C}}^{oT} E[\underline{x}\underline{x}^T]\underline{\underline{C}}^o - \underline{\underline{C}}^{oT} E[\underline{x}\underline{D}^T] - E[\underline{D}\underline{x}^T]\underline{\underline{C}}^o$$

$$= \underline{\underline{R}}_D + R_{xD}^T R_x^{-1} \cancel{R_x R_x^{-1}} R_{xD} - R_{xD}^T R_x^{-1} R_{xD} - R_{xD}^T R_x^{-1} R_{xD}$$

$$= \underline{\underline{R}}_D - R_{xD}^T R_x^{-1} \underline{\underline{R}}_{xD}$$

$$= \underline{\underline{R}}_D - \underline{\underline{C}}^{oT} \underline{\underline{R}}_{xD}$$

It is easy to see that the total squared error can be written as

$$\mathcal{E}^o = tr(\underline{\underline{R}}_{\varepsilon^o}) = tr(\underline{\underline{R}}_D - \underline{\underline{R}}_{xD}^T \underline{\underline{R}}_x^{-1} \underline{\underline{R}}_{xD})$$

$$= tr(R_D - \underline{\underline{C}}^{oT} \underline{\underline{R}}_{xD})$$

ADDING LINEAR CONSTRAINTS

If we have constraints on the filter that can be expressed linearly in the coefficients, it is possible to find closed-form solutions.

ONE LINEAR CONSTRAINT

(a linear constraint may be $\sum_{i=1}^{N} c_i = 1$: $\underline{c}^T \underline{e} = 1$)

$$\begin{cases} \min_{\underline{c}} \mathcal{E}(\underline{c}) \\ \underline{s}^T \underline{c} = \mu \end{cases}$$

The Lagrange function is then

$$L(\underline{c}, \lambda) = E[(d - \underline{c}^T \underline{x})^2] + \lambda(\underline{c}^T \underline{s} - \mu)$$

$$= E[d^2] - 2\underline{c}^T \underline{r}_{dx} + \underline{c}^T \underline{\underline{R}}_x \underline{c} + \lambda(\underline{c}^T \underline{s} - \mu)$$

The gradient equation is

$$\nabla_c L(\underline{c}, \lambda) = -2\underline{r}_{dx} + 2\underline{\underline{R}}_x \underline{c} + \lambda \underline{s} = \underline{0} \quad (a)$$

Multiply by R_x^{-1} (on the left)

$$-2\underline{\underline{R}}_x^{-1} \underline{r}_{dx} + 2\underline{c} + \lambda \underline{\underline{R}}_x^{-1} \underline{s} = 0$$

Multiply by s^T both sides.

$$-2\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{r}_{dx} - 2\underbrace{\underline{s}^T \underline{c}}_{\mu} + \lambda \underline{s}^T \underline{\underline{R}}_x^{-1} \underline{s} = 0$$

$$\lambda = 2 \frac{\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{c} - \underline{s}^T \underline{c}}{\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{s}}$$

Next substitute back in (a) gives:

$$-\cancel{2}\underline{r}_{dx} + \cancel{2}\underline{\underline{R}}_x \underline{c} + \cancel{\lambda} \frac{\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{c} - \cancel{\underline{s}^T \underline{c}}}{\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{s}} \underline{s} = 0 \qquad \mu$$

$$\boxed{\underline{c}^0 = \underline{\underline{R}}_x^{-1} \underline{r}_{dx} - \frac{\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{s} - \mu}{\underline{s}^T \underline{\underline{R}}_x^{-1} \underline{s}} \underline{\underline{R}}_x^{-1} \underline{s}}$$

M LINEAR CONSTRAINTS

Linear constraints could be introduced, for example, to impose response to a set of given frequencies

$$\begin{cases} \min\limits_{\underline{c}} \mathcal{E}(\underline{c}) \\ \underline{\underline{S}}^T \underline{c} = \underline{\mu} \end{cases}$$

$\underline{\underline{S}}$: N × M; $\underline{c}$: N; $\underline{\mu}$: M

$$L(\underline{c}, \underline{\lambda}) = \underline{c}^T \underline{\underline{R}}_x \underline{c} - 2\underline{r}_{dx}^T \underline{c} + \mathcal{E}_d + (\underline{c}^T \underline{\underline{S}} - \underline{\mu}^T)\underline{\lambda}$$

$\underline{\lambda}$: M Lagrange multipliers; $(\underline{c}^T \underline{\underline{S}} - \underline{\mu}^T)$: M; $\underline{\lambda}$: M

$$\nabla_{\underline{c}} L(\underline{c}, \underline{\lambda}) = 2\underline{\underline{R}}_x \underline{c} - 2\underline{r}_{dx} + \underline{\underline{S}}\,\underline{\lambda} = \underline{0} \qquad (*)$$

Multiply by R_x^{-1} on the left.

$$2\underline{\underline{R}}_x^{-1}\underline{\underline{R}}_x \underline{c} - 2\underline{\underline{R}}_x^{-1}\underline{r}_{dx} + \underline{\underline{R}}_x^{-1}\underline{\underline{S}}\,\underline{\lambda} = \underline{0}$$

($\underline{\underline{R}}_x^{-1}\underline{\underline{R}}_x = \underline{\underline{I}}$)

Multiply both sides by $\underline{\underline{S}}^T$ on the left.

$$2\underline{\underline{S}}^T\underline{c} - 2\underline{\underline{S}}^T\underline{\underline{R}}_x^{-1}\underline{r}_{dx} + \underline{\underline{S}}^T\underline{\underline{R}}_x\underline{\underline{S}}\,\underline{\lambda} = \underline{0}$$

$$\underline{\lambda} = 2\left(\underline{\underline{S}}^T\underline{\underline{R}}_x\underline{\underline{S}}\right)^{-1}\left(\underline{\underline{S}}^T\underline{\underline{R}}_x^{-1}\underline{r}_{dx} - \underline{\underline{S}}^T\underline{c}\right)$$

($\underline{\underline{S}}^T\underline{c} = \underline{\mu}$)

Substituting into (*)

$$2\underline{\underline{R}}_x\underline{c} - 2\underline{r}_{dx} + \underline{\underline{S}}\left(\underline{\underline{S}}^T\underline{\underline{R}}_x\underline{\underline{S}}\right)^{-1}\left(\underline{\underline{S}}^T\underline{\underline{R}}_x^{-1}\underline{r}_{dx} - \underline{\mu}\right) = 0$$

$$\boxed{\underline{c}^o = \underline{\underline{R}}_x^{-1}\underline{r}_{dx} - \underline{\underline{R}}_x^{-1}\underline{\underline{S}}\left(\underline{\underline{S}}^T\underline{\underline{R}}_x\underline{\underline{S}}\right)^{-1}\left(\underline{\underline{S}}^T\underline{\underline{R}}_x^{-1}\underline{r}_{dx} - \underline{\mu}\right)}$$

This expression clearly requires the invertibility of $\underline{\underline{S}}^T\underline{\underline{R}}_x\underline{\underline{S}}$ and must be checked case-by-case.

THE CANONICAL FORM

It is often useful to write the error in the so-called canonical form

$$\boxed{\mathcal{E}(\underline{c}) = \mathcal{E}(\underline{c}^o) + (\underline{c}-\underline{c}^o)^T \underline{\underline{R}}_x (\underline{c}-\underline{c}^o)}$$

Where $\underline{c}^o = \underline{\underline{R}}_x^{-1} \underline{r}_{dx}$ and $\mathcal{E}(c^o) = \mathcal{E}_d - \underline{r}_{dx}^T \underline{\underline{R}}_x^{-1} \underline{r}_{dx}$

The canonical form can be verified by direct substitution:

$$\mathcal{E}(c) = \mathcal{E}_d - r_{dx}^T R_x^{-1} r_{dx} + c^T R_x c + e^{oT} R_x e^o - c^T R_x e^o - e^{oT} R_x c$$

$$= \mathcal{E}_d - r_{dx}^T R_x^{-1} r_{dx} + c^T R_x c + r_{dx}^T R_x^{-1} R_x R_x^{-1} r_{dx} - c^T R_x R_x^{-1} r_{dx} - r_{dx}^T R_x^{-1} R_x c$$

$$= \mathcal{E}_d + \underline{c}^T \underline{\underline{R}}_x \underline{c} - 2\, \underline{r}_{dx}^T \underline{c} \quad \square$$