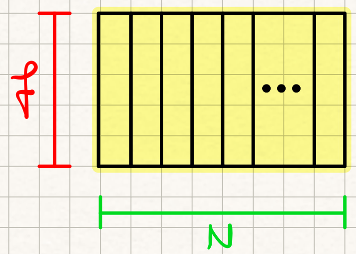
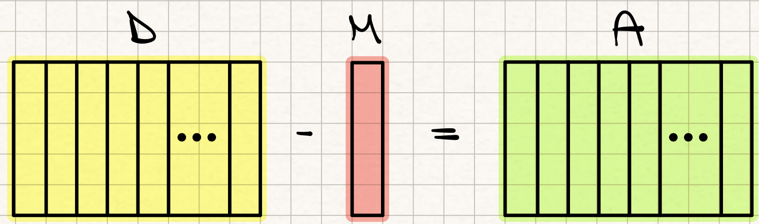
Dati N vettori $\in X^{f}$ → Trovare $K \leq f$ vettori ortogonali

SPAZIO VETTORIALE

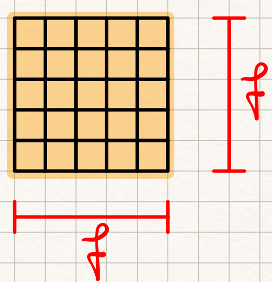


1) Calcolare la matrice A, coi valori sistemati sottraendo al Dataset D la sua media M.



2) Calcoliamo la matrice della covarianze partendo da A.

$$COV(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

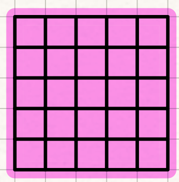
$COV(X,Y) =$  $= C$

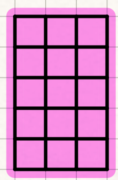
Dove $c_{ij} = cov(i,j)$

$$C = \frac{AA^T}{N-1}$$

Possiamo usare questa formula perché abbiamo normalizzato A, in modo tale che la media delle dimensioni (media(riga_i)) è pari a zero

3) Calcolare gli autovalori di C: $\lambda_1 \dots \lambda_f$ e i rispettivi autovettori. Possiamo scartare gli autovettori con autovalore vicino a $\emptyset$

$S' =$ 

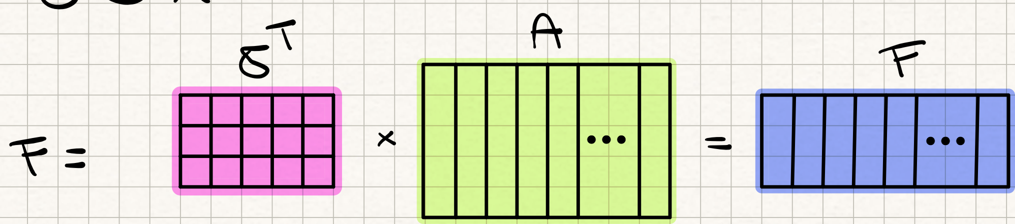
$S'' =$ 

Per esempio eliminiamo gli ultimi due.

4) Trasformiamo il dataset in $F = S^T A$

$S \in X^{5\times 3}$, $A \in X^{5\times n}$

$S^T \in X^{3\times 5}$



PCA

Partendo da un dataset con M feature, l'obiettivo della PCA è quello di trovare K vettori ortogonali di dimensione M che meglio rappresentano il dataset

$$K \leq M$$

Esempio:

Supponiamo di avere il seguente dataset:

altezza	peso	età

Supponiamo che il dataset contiene queste info per 10 persone. Quindi abbiamo dimensione 10×3

STEP 1: Normalizzare i dati

Trasformare ogni attributo in una variabile con media $= 0$ e varianza $= 1$.

	Altezza	Peso	Età
1	1.73	62	27
2	1.68	65	31
3	1.75	70	24
4	1.70	68	28
5	1.80	73	32
6	1.85	75	26
7	1.62	55	30
8	1.76	69	27
9	1.72	71	29
10	1.78	74	33

Altezza Media $= 1,74$
Peso Medio $= 68,20$
Età Media $= 28,70$

Deviazione Standard Altezza $= 0,0816$
Deviazione Standard Peso $= 6,035$
Deviazione Standard Età $= 2,297$

Per ogni valore calcoliamo: $\vec{VAL} - \vec{MEDIA}$

Il dataset dopo la trasformazione appare così:

$A =$

	Altezza	Peso	Età
1	-0,01	-6,20	-1,70
2	-0,06	-3,20	2,30
3	0,01	1,80	-4,70
4	-0,04	-0,20	-0,70
5	0,06	4,80	3,30
6	0,11	6,80	-2,70
7	-0,12	-13,20	1,30
8	0,02	0,80	-1,70
9	-0,02	2,80	0,30
10	0,04	5,80	4,30

STEP 2: Calcolare la matrice di covarianza

Partendo dal dataset trasformato, calcoliamo la matrice di covarianza.
Considerando la formula della covarianza fra due attributi X e Y

$$\mathrm{COV}(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})}{N-1}$$

La matrice di covarianza si calcola $C_{ij} = \mathrm{Cov}(i,j)$ dove i e j rappresentano i vari attributi.
Sulla diagonale della matrice ci saranno i valori $\mathrm{COV}(i,i)$.

$C =$

	Altezza	Peso	Età
Altezza	0	0,32	-0,05
Peso	0,32	0	-0,58
Età	-0,05	-0,58	0

STEP 3: Calcolare gli autovalori di C

$$\lambda_1 = -0,6422$$
$$\lambda_2 = -0,0422$$
$$\lambda_3 = 0,6844$$

Se qualche autovalore è 0 o molto vicino a 0 possiamo scartare questi autovalori e i corrispondenti autovettori

STEP 4: Trasformare il dataset con la nuova base

Chiamiamo σ la matrice degli autovettori:

$$\sigma = \begin{array}{c} \begin{pmatrix} 0,31 & 0,88 & -0,37 \\ -0,72 & -0,04 & -0,70 \\ -0,62 & 0,48 & -0,62 \end{pmatrix} \\ \begin{array}{ccc} v_1 \quad & \quad v_2 \quad & \quad v_3 \end{array} \end{array}$$

$$\lambda_1 = -0,6422$$
$$\lambda_2 = -0,0422$$
$$\lambda_3 = 0,6844$$

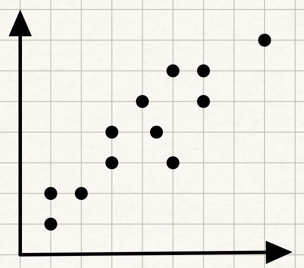
La trasformazione sarà: $F = \sigma^T A$

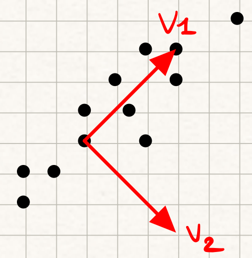
La matrice A va presa capovolta, con le tuple sulle colonne e gli attributi sulle righe.

$$F = \begin{bmatrix} 5.5073 & 0.8425 & 1.6446 & 0.5680 & -5.4857 & -3.1609 & 8.6245 & 0.4930 & -2.2028 & -6.8336 \\ -0.5741 & 1.1821 & -2.3222 & -0.3632 & 1.4433 & -1.4765 & 1.0546 & -0.8316 & 0.0131 & 1.8656 \\ 3.2666 & 3.6642 & -4.1513 & -0.2775 & -1.3258 & -6.4320 & 10.0225 & -1.6111 & -1.7544 & -1.3975 \end{bmatrix}$$

F è il dataset trasformato, MA NON RIDOTTO.
Non si è ridotto perché non ho eliminato nessun v_i

Una volta trovato F, posso scegliere di eliminare le righe, corrispondendi agli attributi i cui autovalori sono bassi.
Difatti la PCA tende a trasformare il punto di vista del dataset in modo tale che la varianza rispetto a un dato autovettore sia la minima. Questo lo si evince dagli autovalori che danno l'idea di quanto una feature è rappresentata bene dalla direzione del autovettore





dove v_1 avra λ_1 vicino a 1, $\lambda_2 \simeq \emptyset$