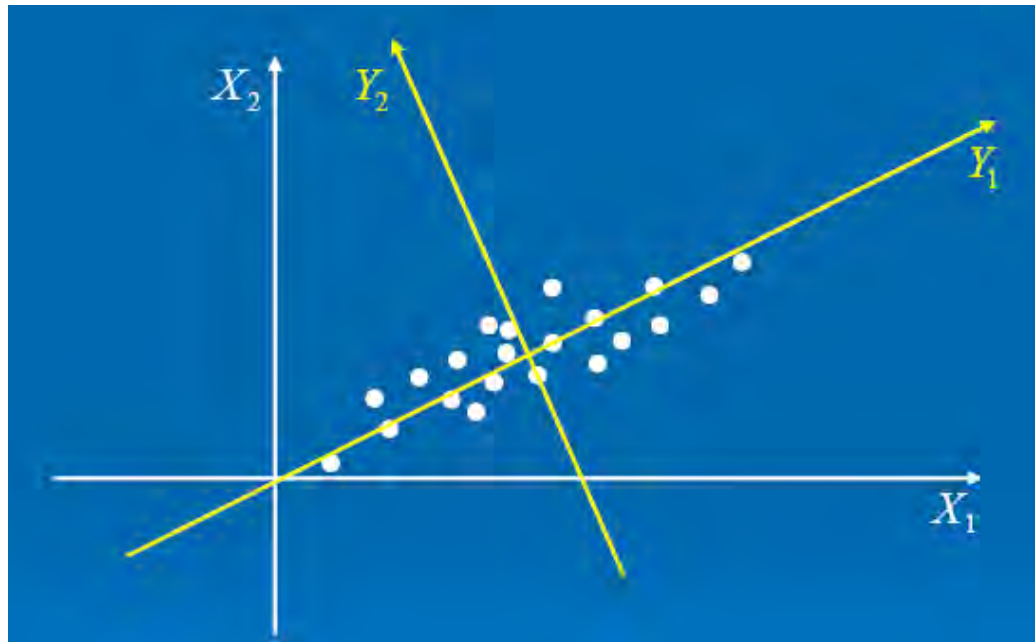


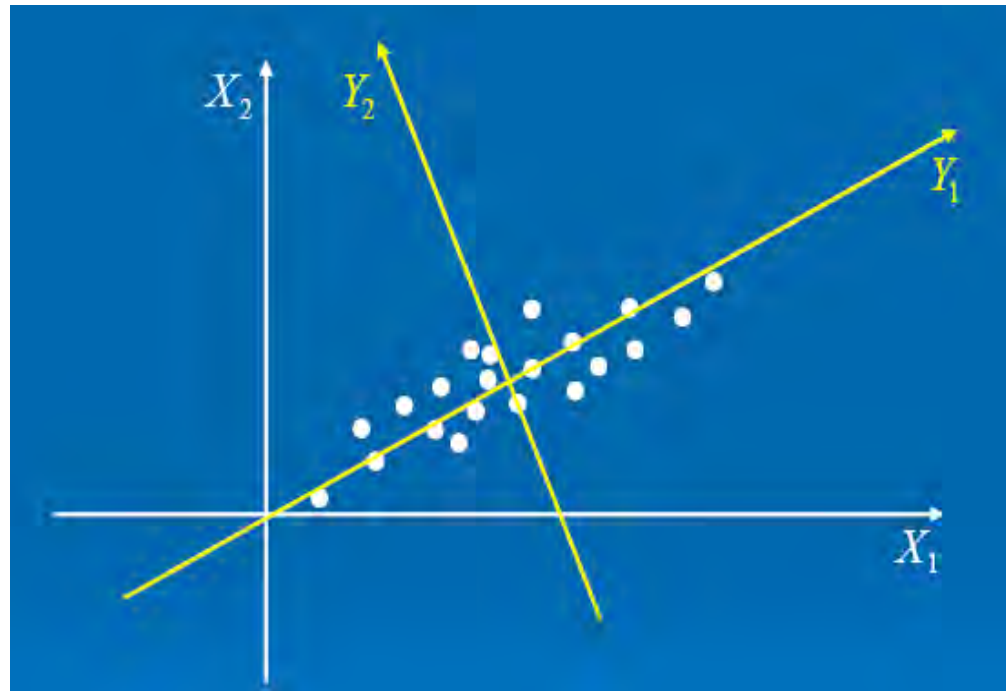
Principal Component Analysis (PCA)

Come evidenziare l'informazione
contenuta nei dati



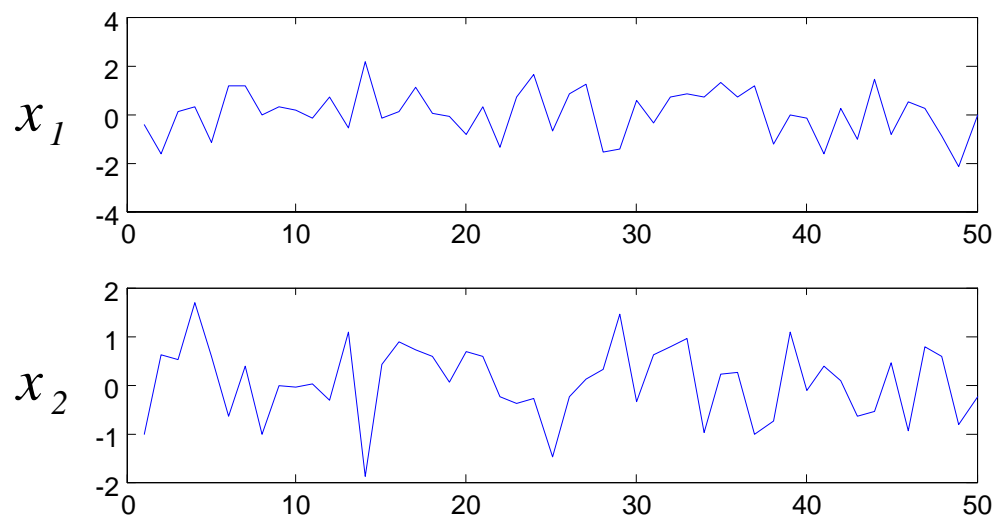
Perche PCA?

- ➡ E' un semplice metodo non-parametrico per estrarre informazione rilevante da un insieme di dati “confuso” (ridondante + rumoroso).
- ➡ Riesce a eliminare la *ridondanza* dell'informazione nei dati, rappresentata dall'*autocorrelazione*
- ➡ Geometricamente l'obiettivo della PCA è presentare i dati nel riferimento che evidenzia maggiormente la loro struttura (Cambio di riferimento)



Correlazione e ridondanza di informazione

☞ Consideriamo una serie di dati bidimensionali, come in figura



Matrice di covarianza

$$\mathbf{C} = \begin{bmatrix} \sigma_1^2 & \rho_{12} \\ \rho_{21} & \sigma_2^2 \end{bmatrix}$$

Matrice di correlazione

$$\mathbf{R} : r_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i} \times C_{j,j}}}$$

☞ Calcolando \mathbf{R} per questi dati si ha

$$\mathbf{R} = \begin{bmatrix} 1.0000 & -0.2074 \\ -0.2074 & 1.0000 \end{bmatrix}$$

← La correlazione fra x_1 e x_2 è circa del 20%. Ciò significa che il 20% dell'informazione di x_1 è contenuta anche in x_2

Ridondanza

Standardizzazione dei dati

- ☞ Generalmente si preferisce svolgere la PCA su dati standardizzati

$$\begin{array}{l} \text{Media nulla} \\ \text{Varianza unitaria} \end{array} \quad \begin{cases} E(x) = \bar{x} = 0 \\ \sigma^2(x) = 1 \end{cases}$$

- ☞ I dati standardizzati si ottengono come

$$z = \frac{x - \bar{x}}{\sigma}$$

- ☞ Ovviamente per i dati standardizzati la matrice di Covarianza coincide con la matrice di correlazione

$$\begin{array}{l} \mathbf{C}(x) \neq \mathbf{C}(z) \\ \mathbf{R}(x) = \mathbf{R}(z) = \mathbf{C}(z) \end{array}$$

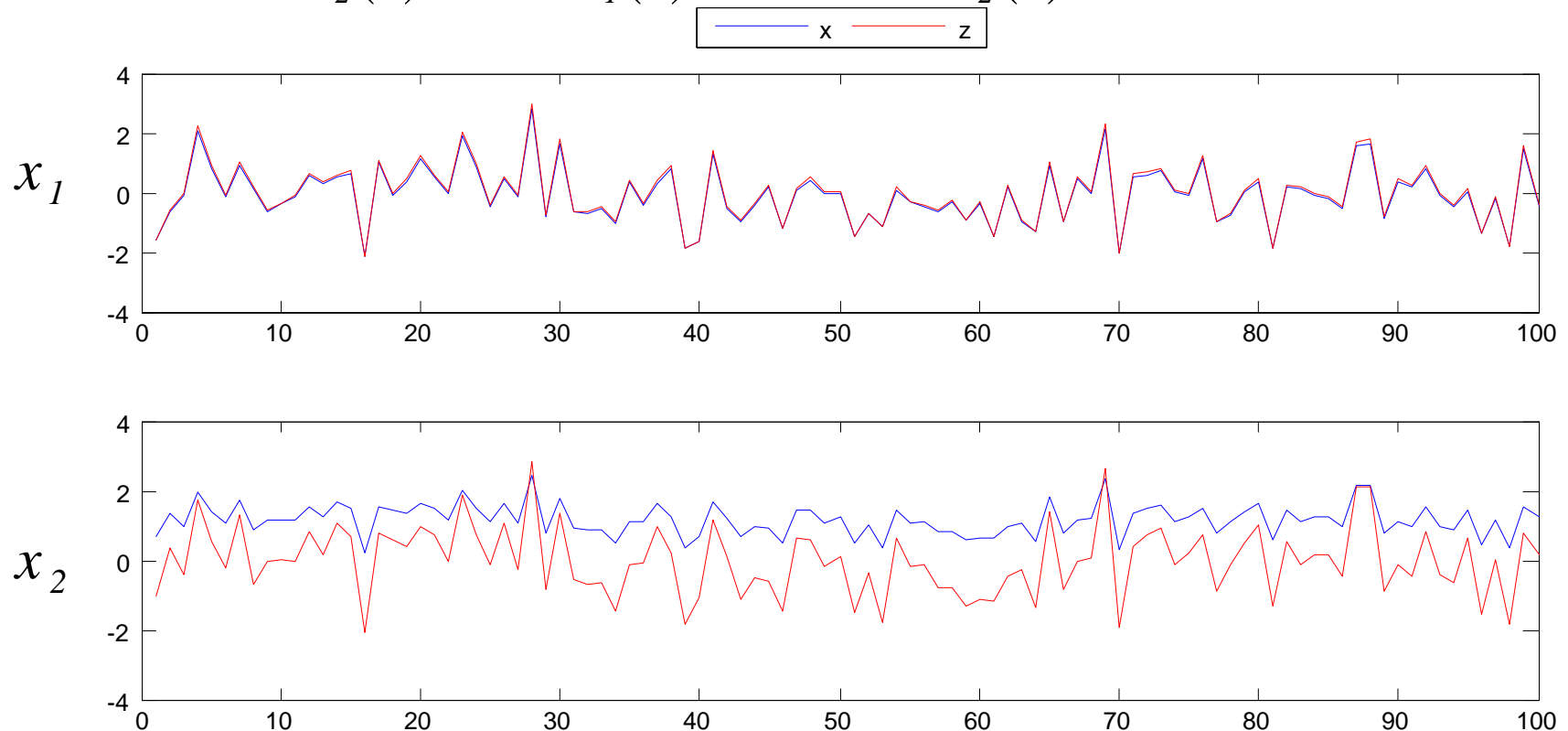
Esempio di dati correlati

☞ Consideriamo il sistema di due variabili *dependenti* (a parte il rumore ε)

$$x_1(k) = \varepsilon_1(k)$$

$$x_2(k) = 0.4 \times x_1(k) + 1.2 + 0.2 \times \varepsilon_2(k)$$

$$\langle \varepsilon_1(k), \varepsilon_2(k) \rangle = 0$$



Matrici di covarianza e correlazione

Dati originali (x)

$$\mathbf{C}_x = \begin{bmatrix} 0.9261 & 0.4000 \\ 0.4000 & 0.2050 \end{bmatrix}$$

$$\mathbf{R}_x = \begin{bmatrix} 1.0000 & 0.9180 \\ 0.9180 & 1.0000 \end{bmatrix}$$

Dati standardizzati (z)

$$\mathbf{C}_z = \begin{bmatrix} 1.0000 & 0.9180 \\ 0.9180 & 1.0000 \end{bmatrix}$$

$$\mathbf{R}_z = \begin{bmatrix} 1.0000 & 0.9180 \\ 0.9180 & 1.0000 \end{bmatrix}$$

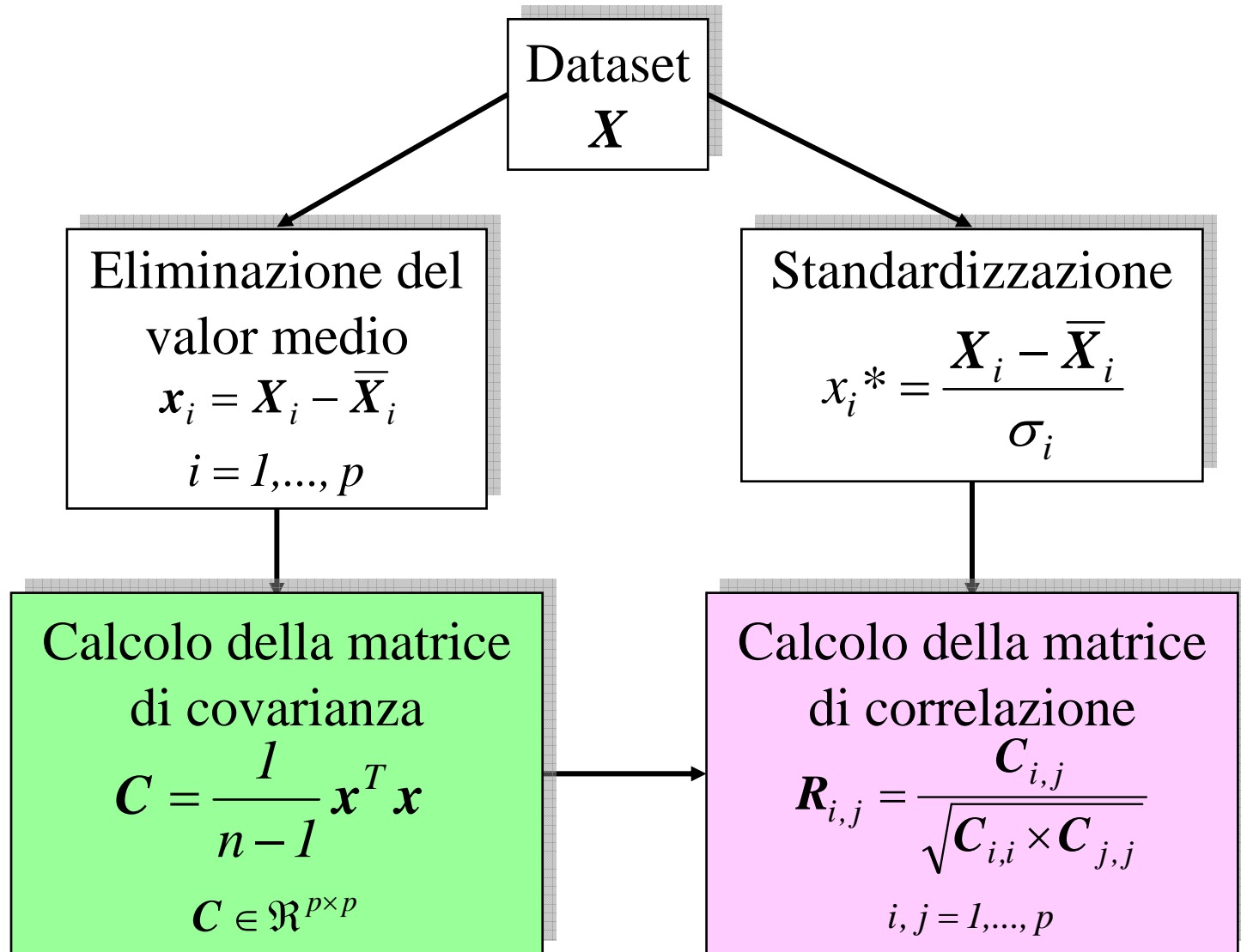
Autovettori di \mathbf{C}_x Autovalori di \mathbf{C}_x

$$\mathbf{W}_c = \begin{bmatrix} 0.4065 & -0.9137 \\ -0.9137 & 0.4065 \end{bmatrix} \quad \begin{matrix} \lambda_1 = 0.0270 \\ \lambda_2 = 1.1041 \end{matrix}$$

Autovettori di $\mathbf{R}_x \mathbf{R}_z$ Autovalori di $\mathbf{R}_x \mathbf{R}_z$

$$\mathbf{W}_r = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \quad \begin{matrix} \lambda_1 = 0.0820 \\ \lambda_2 = 1.9180 \end{matrix}$$

Riassumendo: su quali dati lavoriamo....



Nota sulla standardizzazione

- ☞ La PCA viene normalmente eseguita sulla matrice di Covarianza
 - ⇒ I dati sono depurati dalla media (PCA su dati a media nulla)
- ☞ Se le componenti dei dati hanno ordini di grandezza molto diversi si può ricorrere alla standardizzazione
 - ⇒ PCA su dati a media nulla e varianza unitaria
 - ⇒ La matrice di Covarianza coincide con quella di Correlazione $C = R$
- ☞ Le PCA eseguite su C o su R sono radicalmente diverse perché i rispettivi autovalori e autovalori sono diversi e non ottenibili mediante trasformazione ortonormale
 - ⇒ *Infatti la standardizzazione non è una trasformazione ortogonale*
- ☞ *Conclusione: Se le componenti di x sono molto diverse è conveniente la PCA su R , tenendo comunque presente che essa sarà diversa da quella ottenuta su C*

Rappresentazione grafica della covarianza

- ☞ Dato un insieme di dati $\mathbf{X} \in \mathbb{R}^{n \times 2}$ si calcola la matrice di covarianza \mathbf{C}

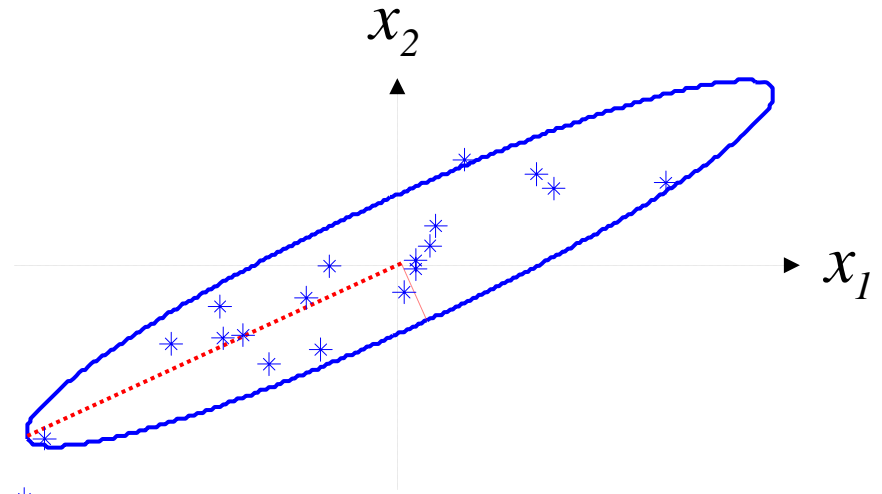
$$\mathbf{x} = \mathbf{X} - \bar{\mathbf{X}}$$

$$\mathbf{C} = \frac{1}{n-1} \mathbf{x}^T \mathbf{x}$$

- ☞ Si calcolano gli autovettori \mathbf{w} e gli autovalori λ
- ☞ Fra di essi valgono le relazioni di similitudine

$$\mathbf{C} = \mathbf{W} \cdot \mathbf{L} \cdot \mathbf{W}^T \quad \text{con} \quad \mathbf{W} = [\mathbf{w}_1 / \mathbf{w}_2]$$

$$\mathbf{L} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

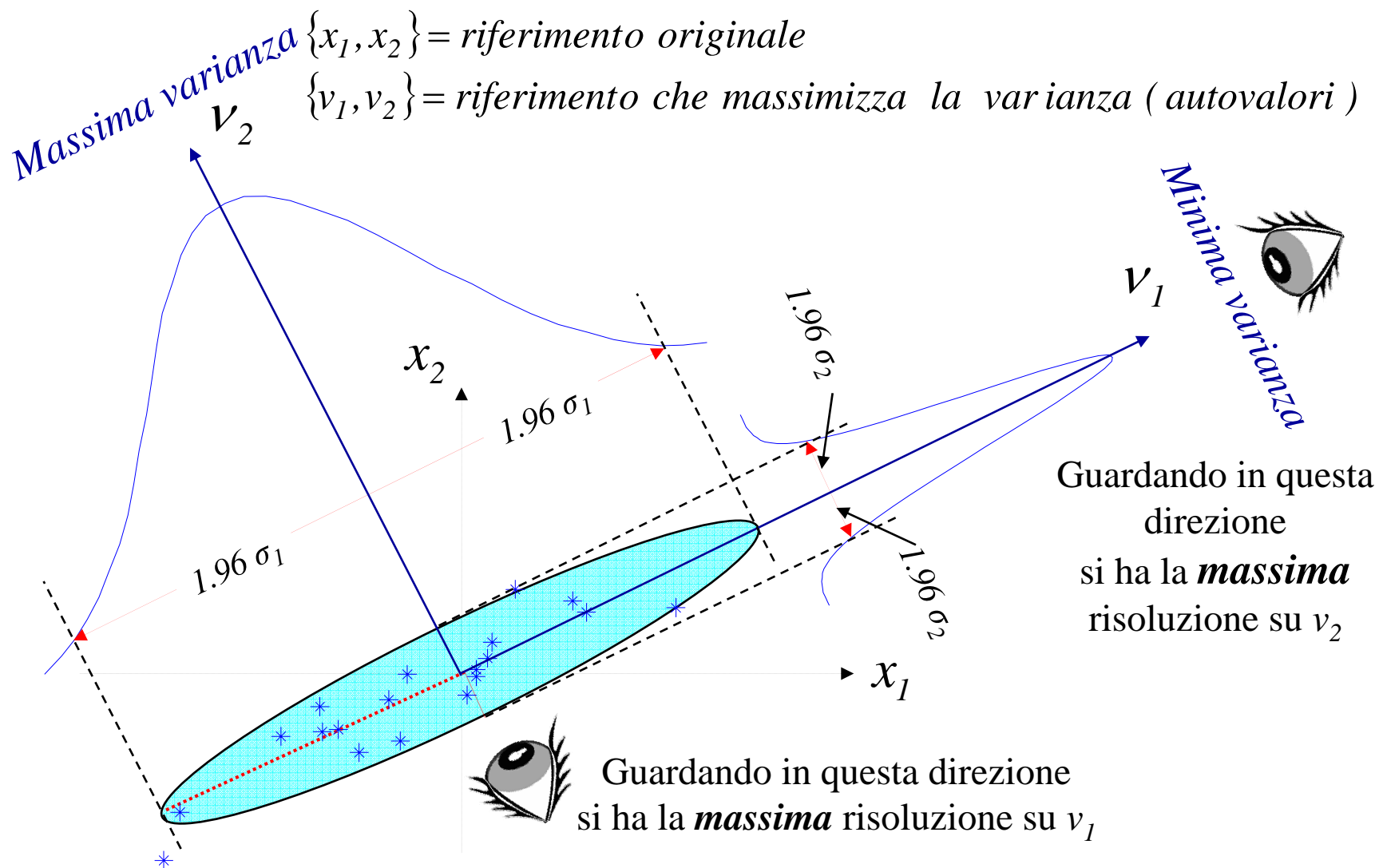


* Mediante autovalori ed autovettori si può scrivere l'equazione parametrica dell'ellisse che racchiude il 95% dei dati.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \cdot \begin{bmatrix} 1.96\sqrt{\lambda_1} & 0 \\ 0 & 1.96\sqrt{\lambda_2} \end{bmatrix} \cdot \begin{bmatrix} \cos \varphi \\ -\sin \varphi \end{bmatrix} + \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} \quad \varphi \in (0, 2\pi)$$

Gli autovettori danno le direzioni degli assi dell'ellisse e gli autovettori la loro lunghezza

Direzioni principali in funzione della covarianza



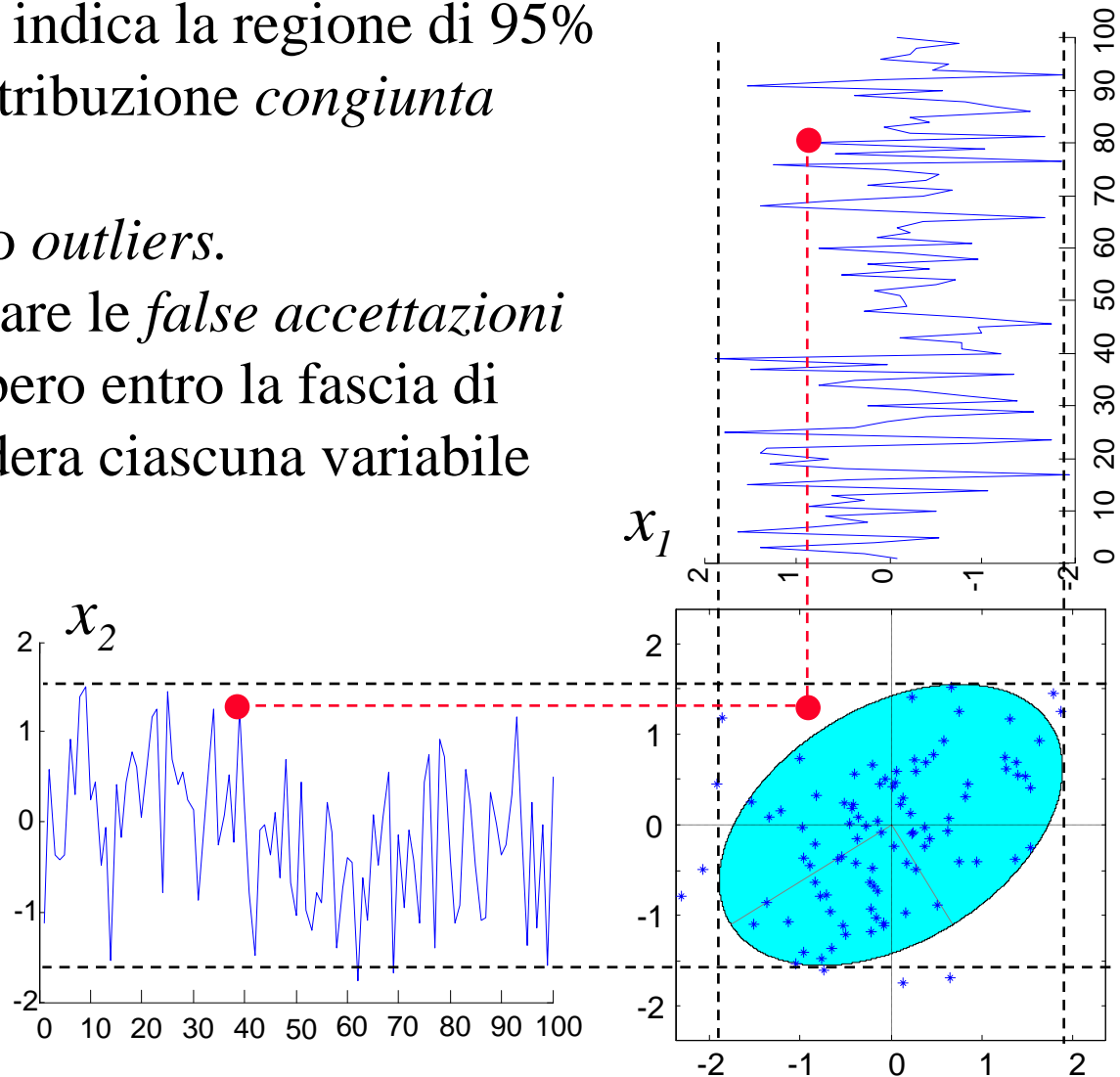
Probabilità congiunta e outliers

L'ellisse di covarianza indica la regione di 95% di confidenza della distribuzione *congiunta* delle due variabili.

I campioni esterni sono *outliers*.

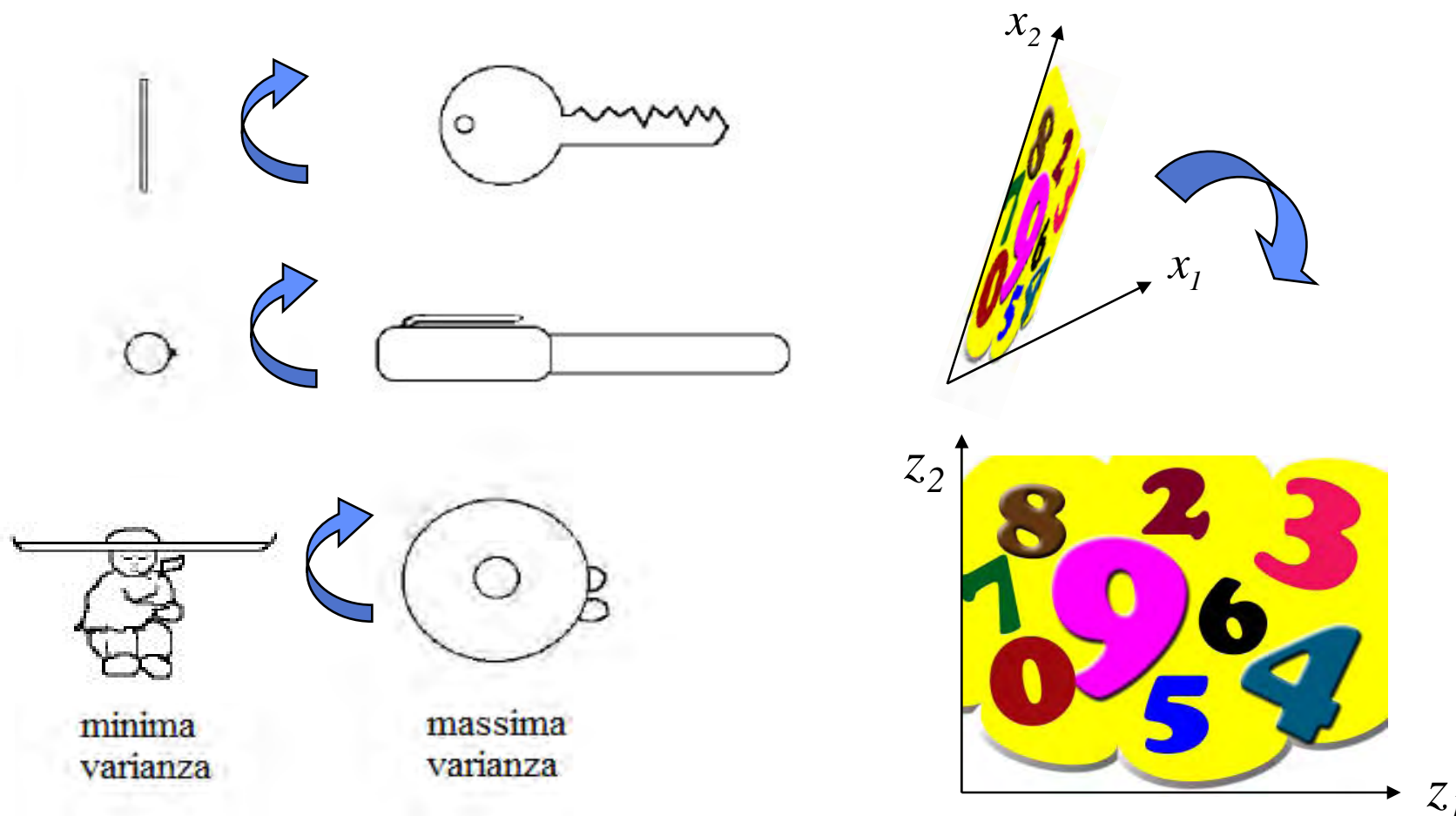
Perciò permette di evitare le *false accettazioni* di campioni che sarebbero entro la fascia di confidenza se si considera ciascuna variabile *separatamente*.

Si vede che il punto rosso ● è fuori solo se si considera la regione di confidenza *bidimensionale*



PCA = migliore visualizzazione

- Il cambio di riferimento può essere visto come un cambio di punto di vista che massimizza l'informazione “visibile” nei dati



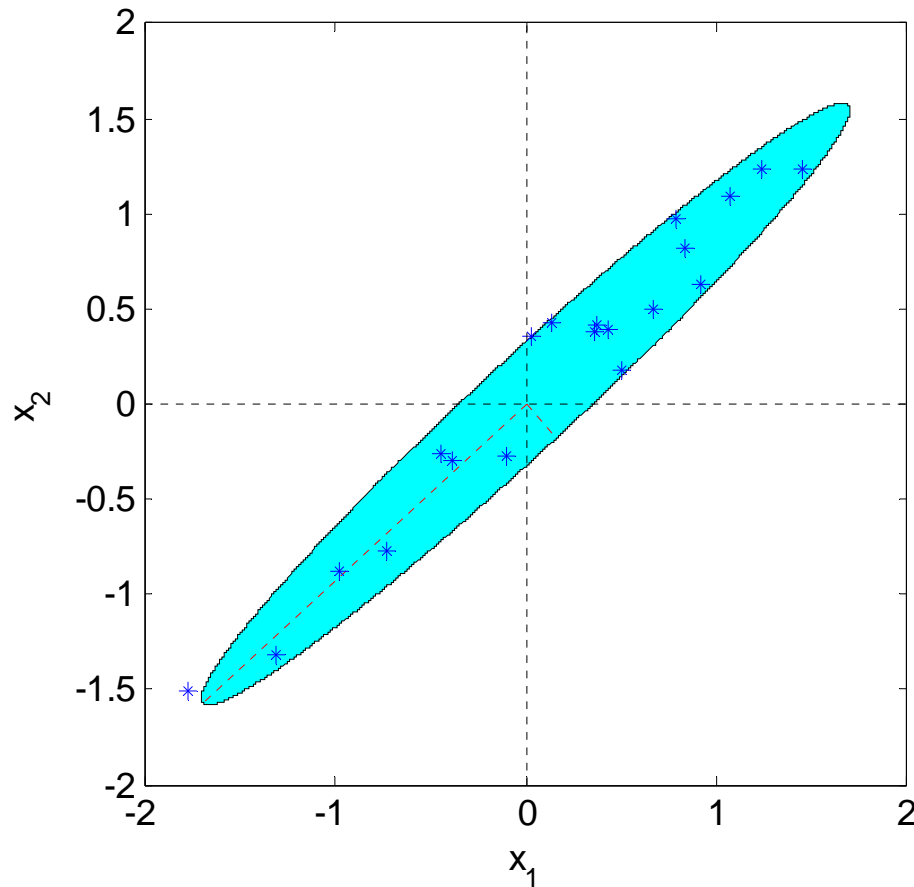
Vantaggi della PCA

- ☞ Le PCA forniscono una *spiegazione alternativa della variabilità osservata* con il pregio di descrivere il fenomeno oggetto di studio *mediante dimensioni fra loro non correlate e ordinate* in termini della loro importanza nella spiegazione
- ☞ Questo permette (con maggiore o minore successo nei vari casi) di :
 - ⇒ *interpretare* il fenomeno attraverso il nuovo significato assunto dalle componenti principali che non sono state scartate
 - ⇒ *ridurre il numero di variabili da considerare*, scartando le ultime componenti principali, che contribuiscono poco alla variabilità osservata

Covarianza fra i dati

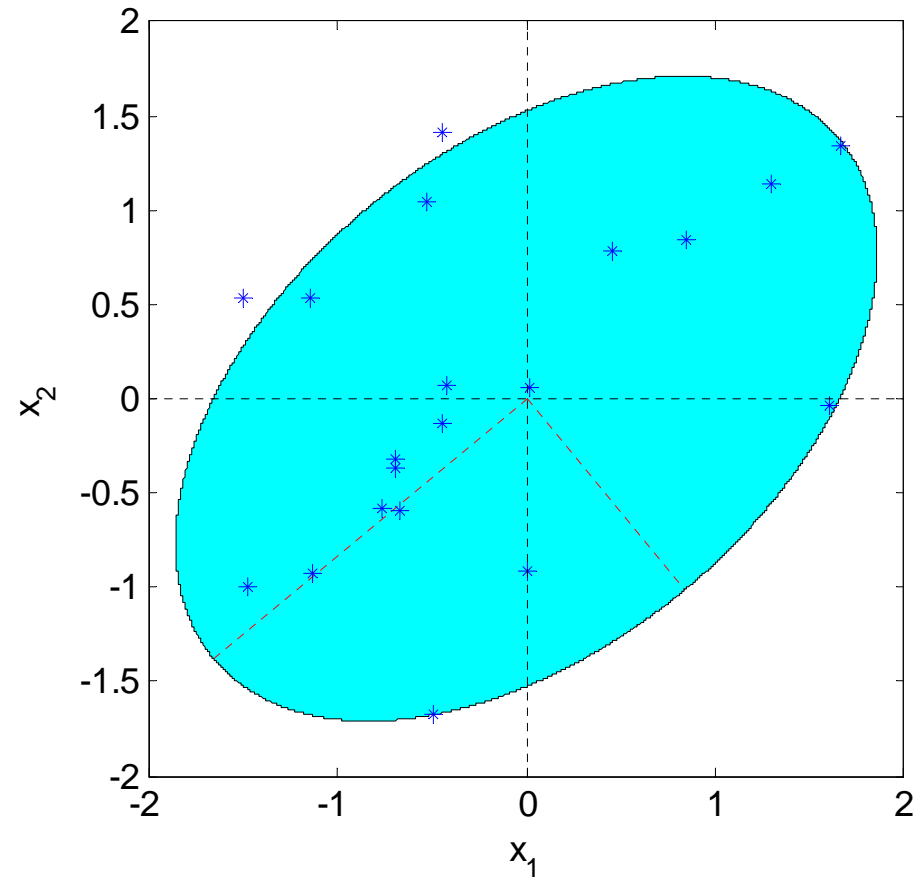
Alta covarianza

$$\begin{cases} x_1 = \varepsilon_1 \\ x_2 = 0.995x_1 + 0.25\varepsilon_2 \end{cases} \quad \langle \varepsilon_1, \varepsilon_2 \rangle = 0$$



Bassa covarianza

$$\begin{cases} x_1 = \varepsilon_1 \\ x_2 = 0.15x_1 + 0.95\varepsilon_2 \end{cases} \quad \langle \varepsilon_1, \varepsilon_2 \rangle = 0$$



Rappresentazione grafica della correlazione

- ☞ Si può ricavare la matrice di correlazione normalizzando le varianze o calcolando la matrice di covarianza sui dati standardizzati

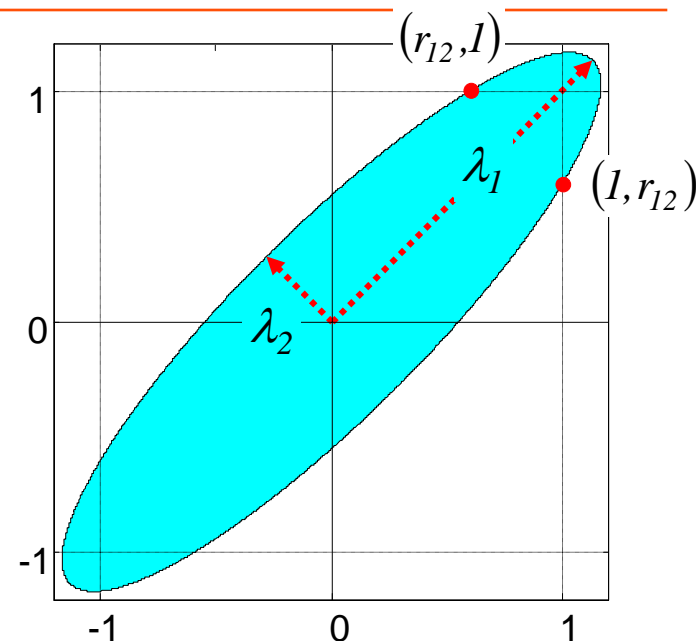
$$R_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i} \times C_{j,j}}}$$

- ☞ Si ha la matrice simmetrica che nel caso 2x2 è del tipo

$$R = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

- ☞ Con autovettori e autovalori

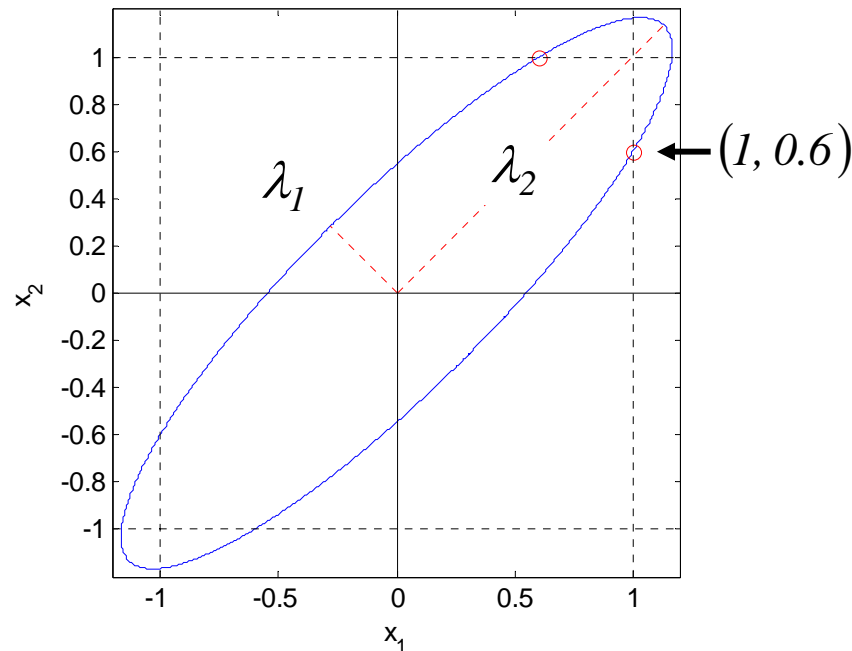
$$W = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \quad \begin{matrix} \lambda_1 = 1 + r_{12} \\ \lambda_2 = 1 - r_{12} \end{matrix}$$



L'equazione dell'ellisse di correlazione, centrata nell'origine è

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} \cos \varphi \\ -\sin \varphi \end{bmatrix}$$

Rappresentazione grafica della correlazione

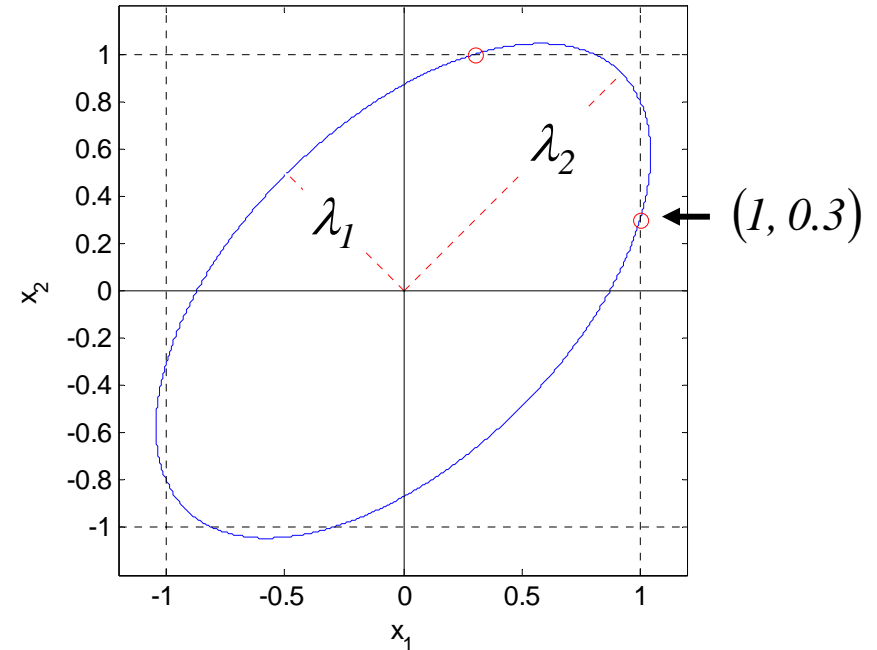


$$R = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$$

$$W = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \xleftrightarrow[\text{autovettori}]{\text{Uguali}}$$

$$\lambda_1 = 0.4 = 1 - r_{12}$$

$$\lambda_2 = 1.6 = 1 + r_{12}$$



$$R = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$$

$$W = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

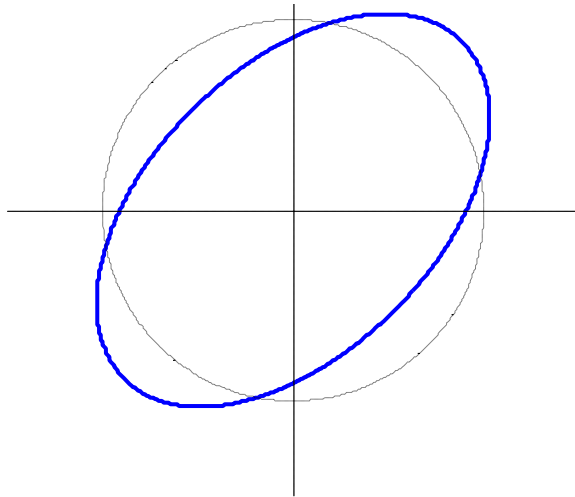
$$\lambda_1 = 0.7 = 1 - r_{12}$$

$$\lambda_2 = 1.3 = 1 + r_{12}$$

Esempi di ellissi di correlazione

Correlazione = Ridondanza di informazione

bassa



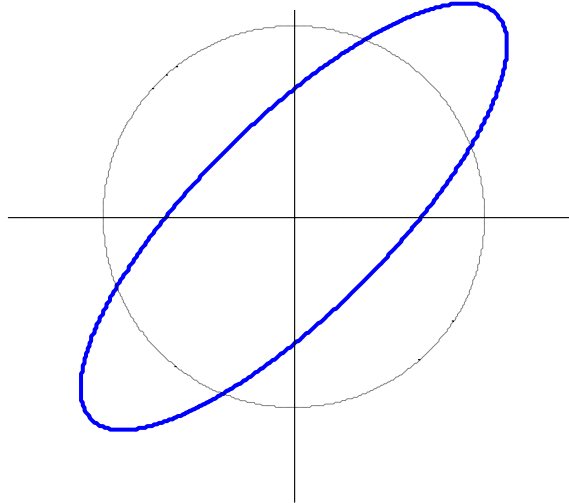
$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.25 \\ 0.25 & 1.0 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

$$\lambda_1 = 1 - r_{12} = 0.75$$

$$\lambda_2 = 1 + r_{12} = 1.25$$

media



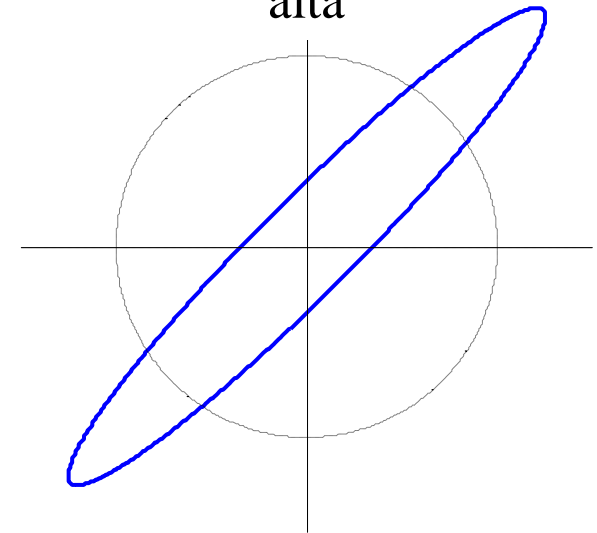
$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

$$\lambda_1 = 1 - r_{12} = 0.5$$

$$\lambda_2 = 1 + r_{12} = 1.5$$

alta



$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

$$\lambda_1 = 1 - r_{12} = 0.25$$

$$\lambda_2 = 1 + r_{12} = 1.75$$

Obiettivi della PCA

- ☞ L'obiettivo primario della PCA è determinare la **base di riferimento** più significativa per rappresentare i dati e filtrare il rumore nella speranza che questa nuova base filtri il **rumore** e riveli **strutture** prima invisibili
- ☞ PCA è una trasformazione lineare dei dati che:
 1. Minimizza la ridondanza misurata dalla covarianza
 2. Massimizza l'informazione, misurata dalla varianza.
- ☞ Le *Principal Components* (PC) sono nuove variabili che hanno le seguenti proprietà:
 1. Ogni PC è una combinazione lineare delle variabili originali
 2. Le PC sono fra di loro ortogonali, ovvero sono mutuamente incorrelate, sopprimendo l'informazione ridondante

Idea base della PCA

👉 **Dataset:** insieme di n misure ciascuna composta da p attributi

$$\mathbf{x} = \begin{matrix} & \overbrace{\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}}^{p \text{ attributi}} \\ \left. \begin{matrix} \\ \\ \\ \end{matrix} \right\} & n \text{ misure } \mathbf{x} \in \mathbb{R}^{n \times p} \end{matrix}$$

👉 L'intuizione di PCA è di trovare una combinazione lineare delle m coordinate dei dati in modo da esprimerli in un nuovo riferimento tale che:

- ⇒ Ogni variabile (attributo) sia indipendente da tutti gli altri
- ⇒ L'insieme degli attributi sia ordinato secondo la loro importanza relativa

Caratteristiche della PCA

PCA è una trasformazione lineare ortonormale dei dati X al fine di ottenere due risultati:

- 👉 **Feature Selection:** classificare le caratteristiche importanti dei dati X , secondo la loro importanza
 - ⇒ PCA evidenzia il contenuto informativo mediante una trasformazione lineare delle coordinate di riferimento dei dati (attributi dei dati)
- 👉 **Dimension Reduction:** quantificare la perdita di informazione derivante dall'eventuale riduzione della dimensionalità dei dati X .
 - ⇒ PCA quantifica la percentuale di informazione nelle varie componenti ordinate per importanza, in modo da conoscere la perdita di informazione per ciascuna componente esclusa dalla riduzione

Risultato fondamentale della PCA

- ☞ Se l'obiettivo primario è l'eliminazione della ridondanza
- ☞ Se la ridondanza è espressa dalle correlazioni

☞ *Allora la PCA consiste nella diagonalizzazione della matrice di covarianza*

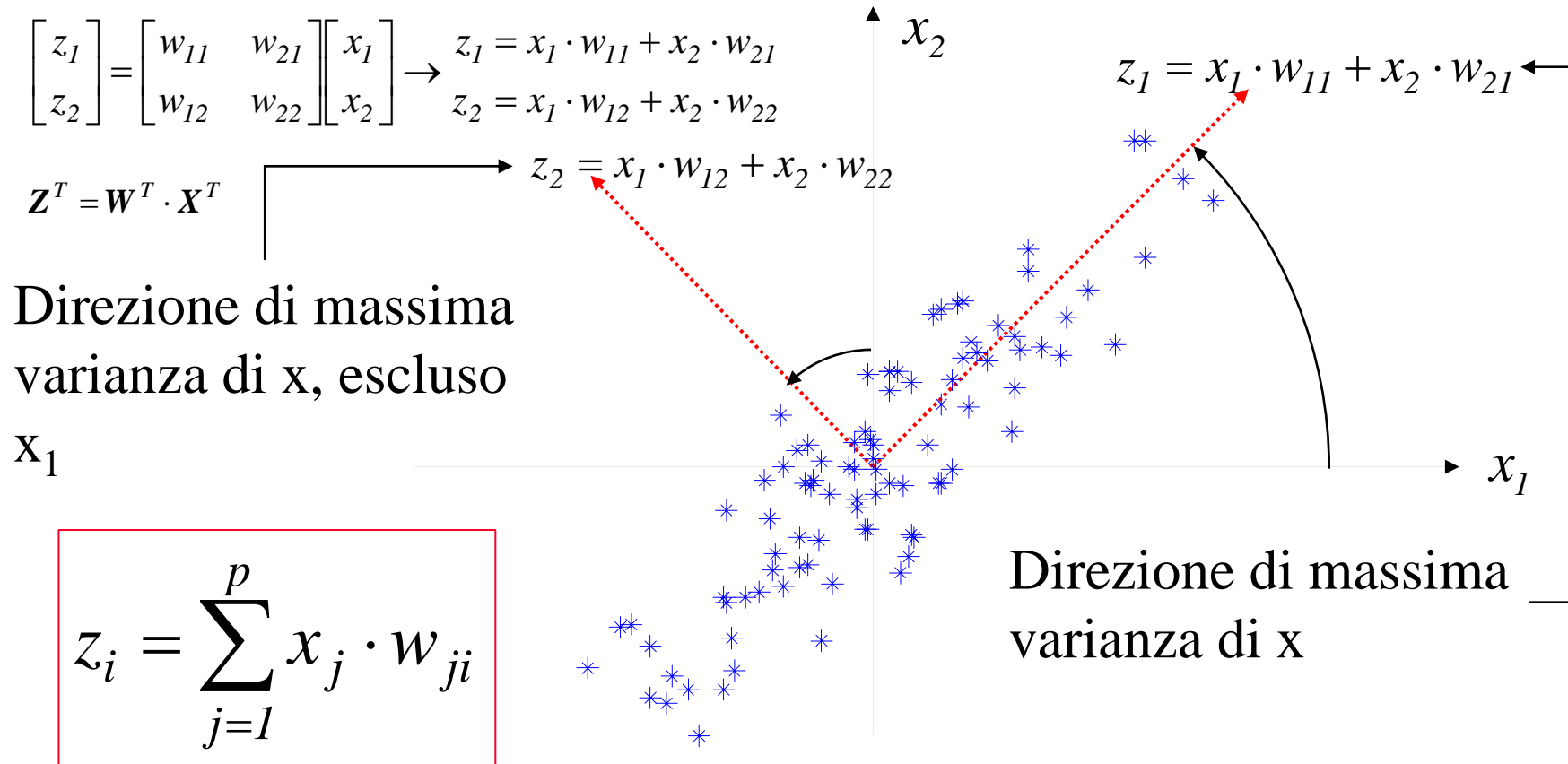
- ☞ PCA consiste dunque in una trasformazione lineare dalle variabili originali ad altre che esprimono la stessa informazione ma sono fra loro incorrelate (Componenti Principali)
- ☞ La trasformazione cercata è la similitudine W fra la matrice di correlazione e la matrice diagonale degli autovalori, tale che

$$L = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2) = W^T \cdot C \cdot W$$

$$Z = X \cdot W$$

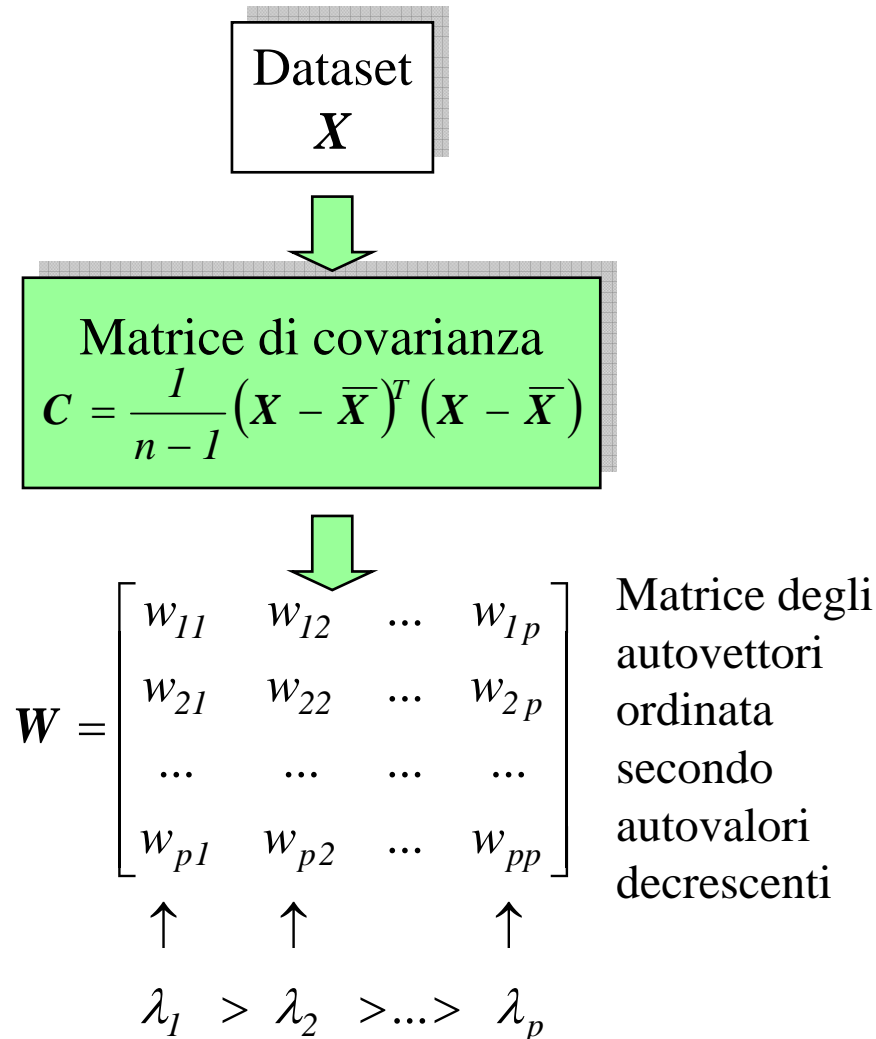
$$W \text{ è } \perp: W^{-1} = W^T$$

PCA come ricerca delle direzioni privilegiate



*PCA consiste dunque nella ricerca di direzioni privilegiate
che massimizzano la variazioni dei dati ed
eliminano le correlazioni*

PCA in sintesi



La matrice W formata dagli autovettori ordinati per autovalori decrescenti indicano le direzioni di massima varianza. La similitudine fra C e L è data da W .

Nota che essendo ortonormale $W^T = W^{-1}$

$$C = W \cdot L \cdot W^T$$

$$L = W^T \cdot C \cdot W$$

La matrice L (diagonale) degli riporta i valori delle varianze nel nuovo riferimento PCA

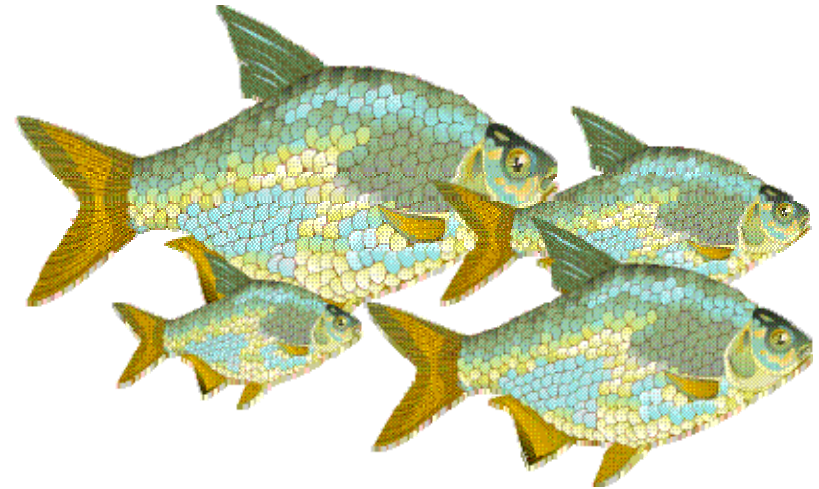
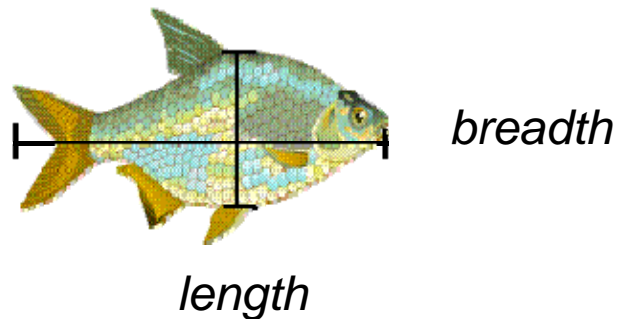
$$L = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$$

La trasformazione dei dati X nelle componenti principali Z è

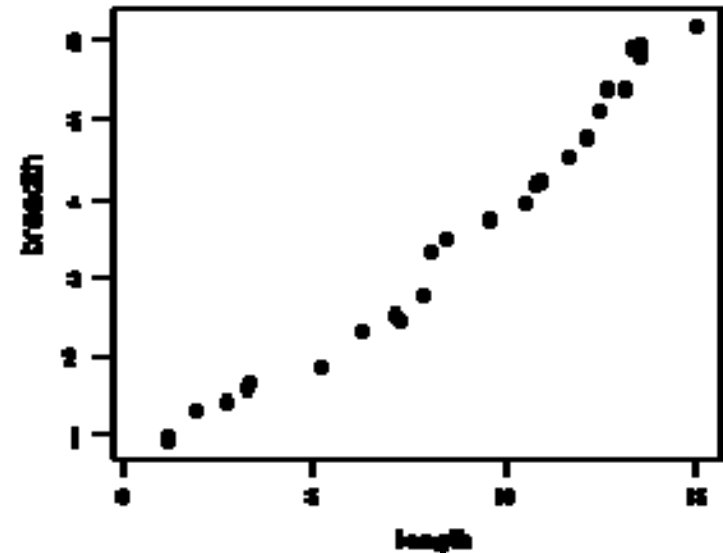
$$Z = X \cdot W \leftrightarrow X = Z \cdot W^T$$

Un semplice esempio

- ➡ Ogni pesce può essere definito dalle sue misure di lunghezza ed larghezza

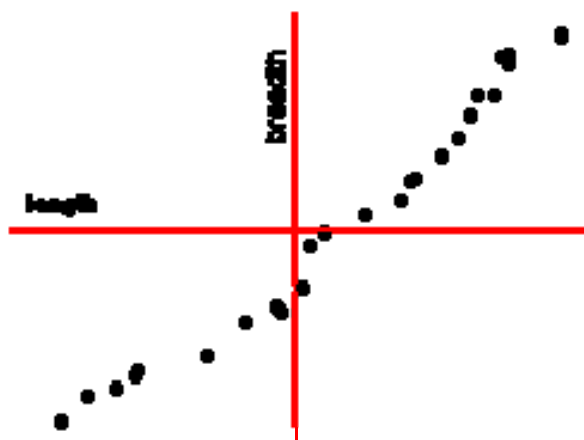


- ➡ Riportando in grafico i dati degli individui del branco di pesci, si ottiene →
- ➡ Domanda 1: Esiste una relazione fra le due misure?
- ➡ Domanda 2: Esiste un *singolo* parametro per definire la taglia di ciascun pesce?

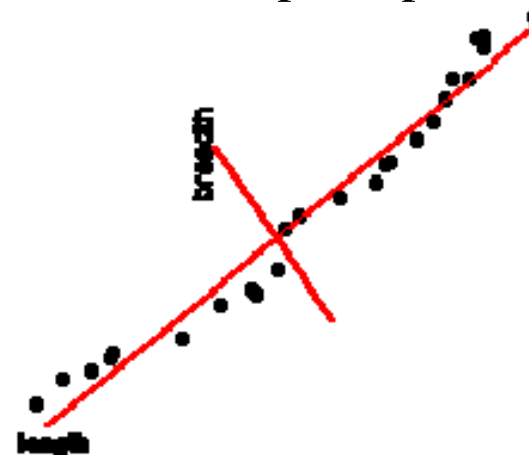


PCA sui dati dei pesci

- ☞ Scegliamo dei nuovi assi centrati nell'insieme dei dati



- ☞ Poi ruotiamo gli assi per disporli lungo la direzione principale dei dati



- ☞ Possiamo allora definire una nuova variabile: **size = length + breadth**
- ☞ Ma dato che **length** e **breadth** non sono ugualmente importanti (vedi grafico) esse dovranno essere “pesate” diversamente, perciò

$$\text{size} = v_1 \text{ length} + v_2 \text{ breadth}$$

- ☞ I pesi v_1 e v_2 sono gli autovettori della matrice di correlazione

Risultato: si è ottenuta una riduzione della dimensione dei dati

Cosa si perde nella riduzione

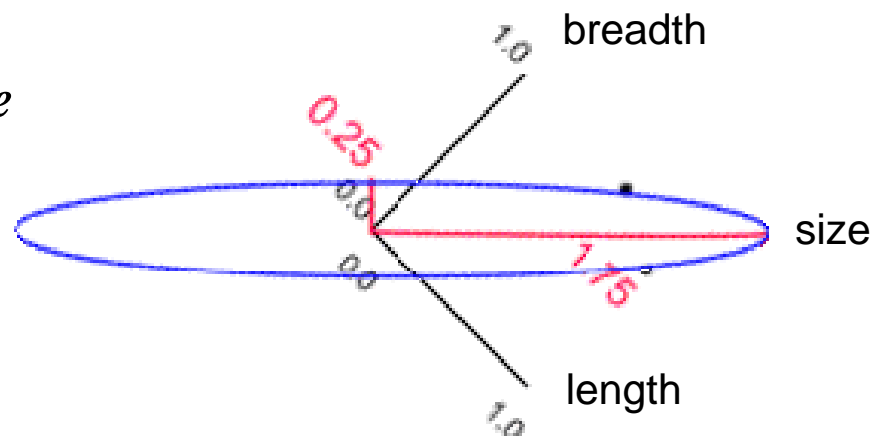
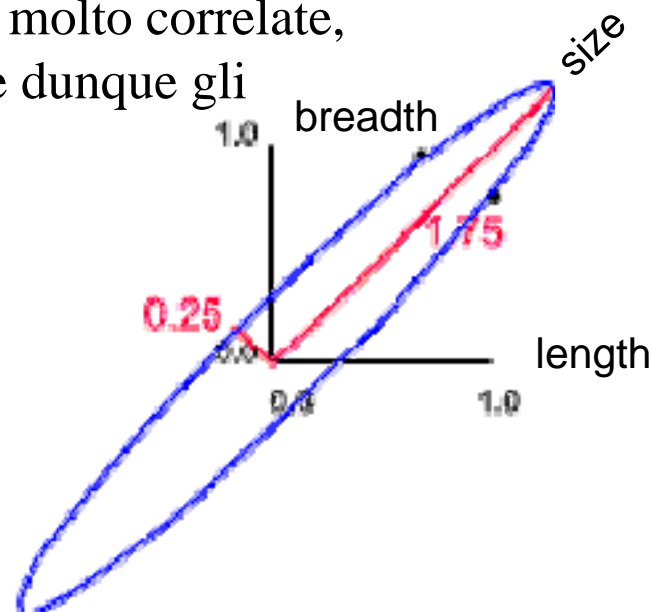
➡ Dato che *length* e *breadth* sono chiaramente molto correlate, la matrice di correlazione è molto allungata e dunque gli autovalori sono molto diversi.

➡ Supponiamo che essi valgano $\lambda_1 = 1.75$ e $\lambda_2 = 0.25$ nel riferimento originale l'ellisse sarà

➡ Dopo la rotazione dovuta al cambio di riferimento sarà data dall'orientamento degli autovettori e dalla grandezza degli autovalori

➡ Se si ritiene solamente la variabile *size* si conserva solamente 87.5% della variabilità originale. Infatti

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.75}{1.75 + 0.25} = 87.5\%$$



PCA in Matlab

- 👉 Le funzioni PCA sono contenute nella Statistics Toolbox
- 👉 Si può effettuare la PCA partendo dai dati (X) o dalla matrice di covarianza (C)
- 👉 dai dati: **[W,score,L] = princomp(X)**
 - **W** è la matrice degli autovettori, detta matrice dei *Loadings*
 - ✍ E' ordinata per autovalori decrescenti
 - **Scores** sono le osservazioni Z trasformate delle X nel riferimento PCA
 - **L** sono gli autovalori, ordinati in ordine decrescente
- 👉 dalla matrice di covarianza: **[W,L,expl] = pcacov(C)**
 - **W** è la matrice degli autovettori
 - ✍ E' ordinata per autovalori decrescenti
 - **L** sono gli autovalori, ordinati in ordine decrescente
 - **expl** è un vettore che contiene la percentuale di varianza spiegata da ciascuna componente principale (la somma fa 100)

Un esempio: Employee Satisfaction

☞ Un sondaggio fra 9147 impiegati di una grande azienda ha rilevato i seguenti parametri di soddisfazione

- Lavoro (SJ)
- Formazione (SJT)
- Condizioni di lavoro (SWC)
- Assicurazione medica (SMC)
- Assicurazione Dentistica (SDC)

☞ Il sondaggio ha prodotto la seguente matrice di correlazione

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.451 & 0.511 & 0.197 & 0.162 \\ 0.451 & 1.000 & 0.445 & 0.252 & 0.238 \\ 0.511 & 0.445 & 1.000 & 0.301 & 0.227 \\ 0.197 & 0.252 & 0.301 & 1.000 & 0.620 \\ 0.162 & 0.238 & 0.227 & 0.620 & 1.000 \end{bmatrix}$$

Un esempio: Employee Satisfaction

- ☞ Si ricava la matrice delle PCA ordinata per autovalori decrescenti

$$W = \begin{bmatrix} 0.442 & 0.443 & 0.301 & -0.716 & 0.074 \\ 0.457 & 0.290 & -0.832 & 0.114 & 0.034 \\ 0.479 & 0.308 & 0.454 & 0.658 & -0.185 \\ 0.443 & -0.531 & 0.095 & 0.060 & 0.714 \\ 0.412 & -0.586 & 0.032 & -0.191 & -0.670 \end{bmatrix}$$

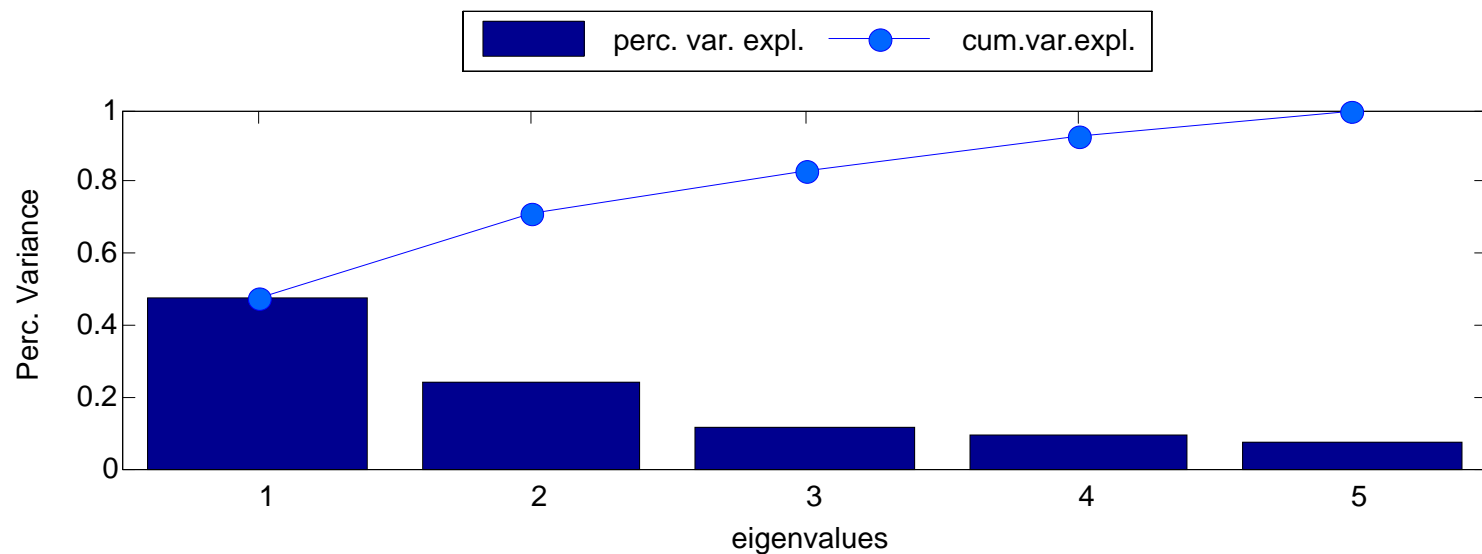
Verifica: la somma dei quadrati degli elementi di ciascuna colonna somma a 1

$$L = [2.370 \quad 1.202 \quad 0.573 \quad 0.484 \quad 0.373]$$

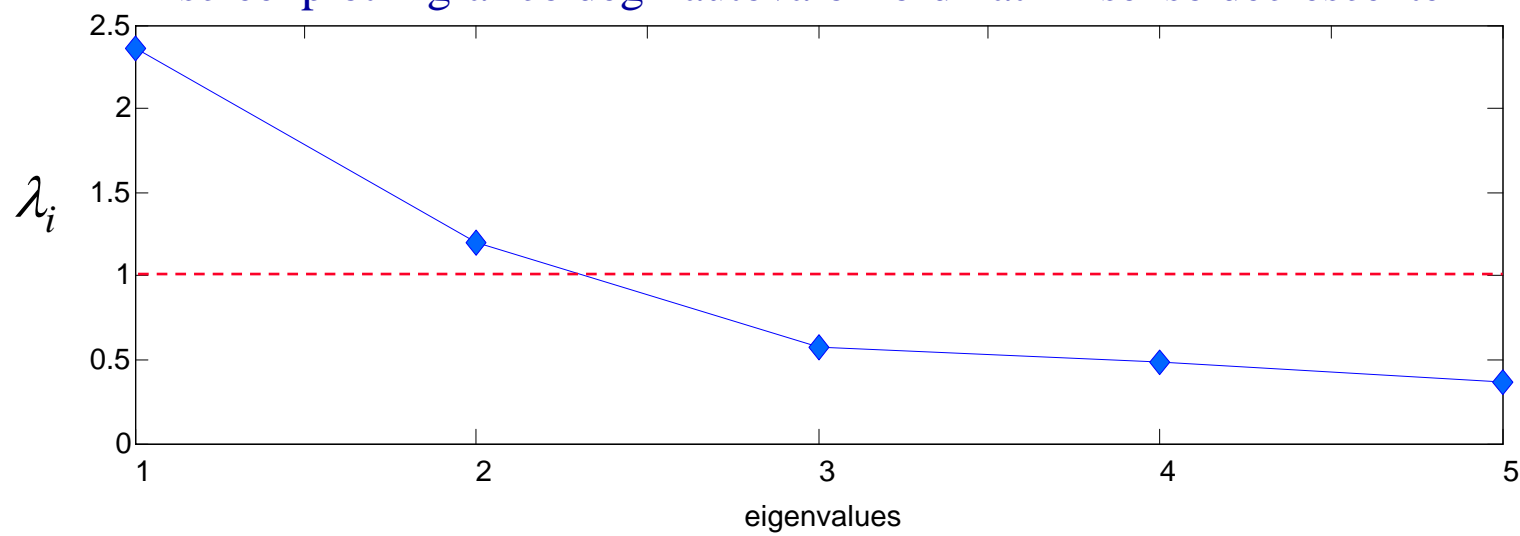
- ☞ La prima PC spiega il 47.3% della varianza totale

$$\frac{\lambda_1}{\sum_{i=1}^5 \lambda_i} = \frac{2.370}{2.370 + 1.202 + 0.573 + 0.484 + 0.373} = \frac{2.370}{5} = 0.473$$

Varianza spiegata e scree plot



scree plot = grafico degli autovalori ordinati in senso decrescente



Un esempio: Employee Satisfaction

☞ La prima PC z_1 si ottiene come

$$z_1 = 0.442 \cdot SJ + 0.457 \cdot SJT + 0.479 \cdot SWC + 0.443 \cdot SMC + 0.412 \cdot SDC$$

essendo i pesi tutti positivi e dello stesso ordine, z_1 si può interpretare come un indice di soddisfazione generale

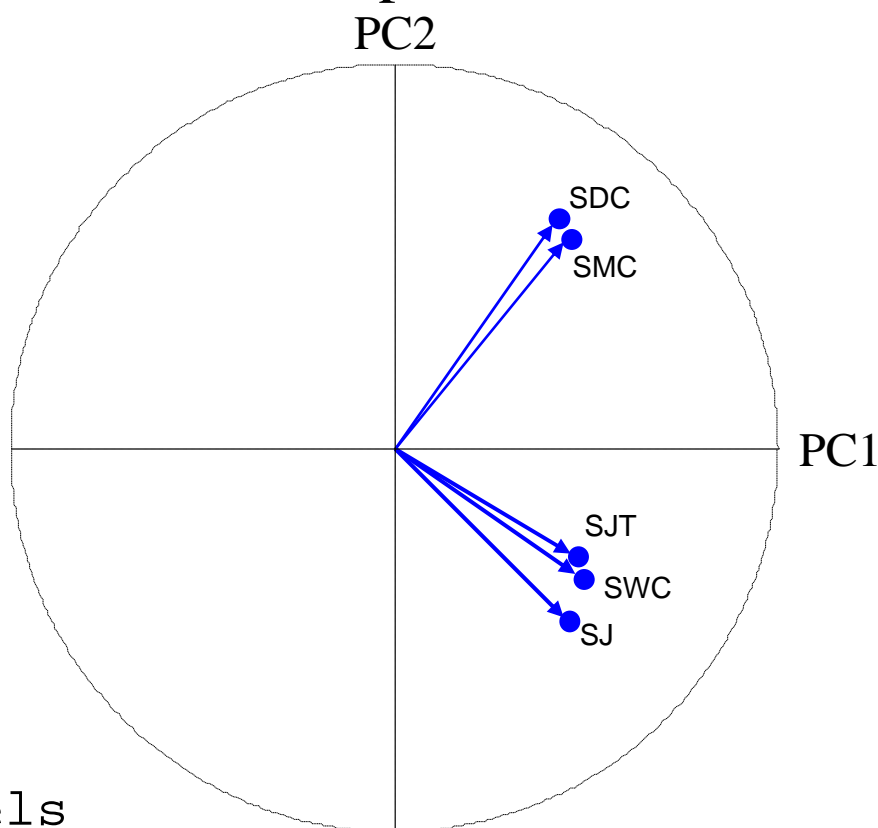
☞ La seconda PC z_2 è

$$z_2 = 0.443 \cdot SJ + 0.290 \cdot SJT + 0.308 \cdot SWC - 0.531 \cdot SMC - 0.586 \cdot SDC$$

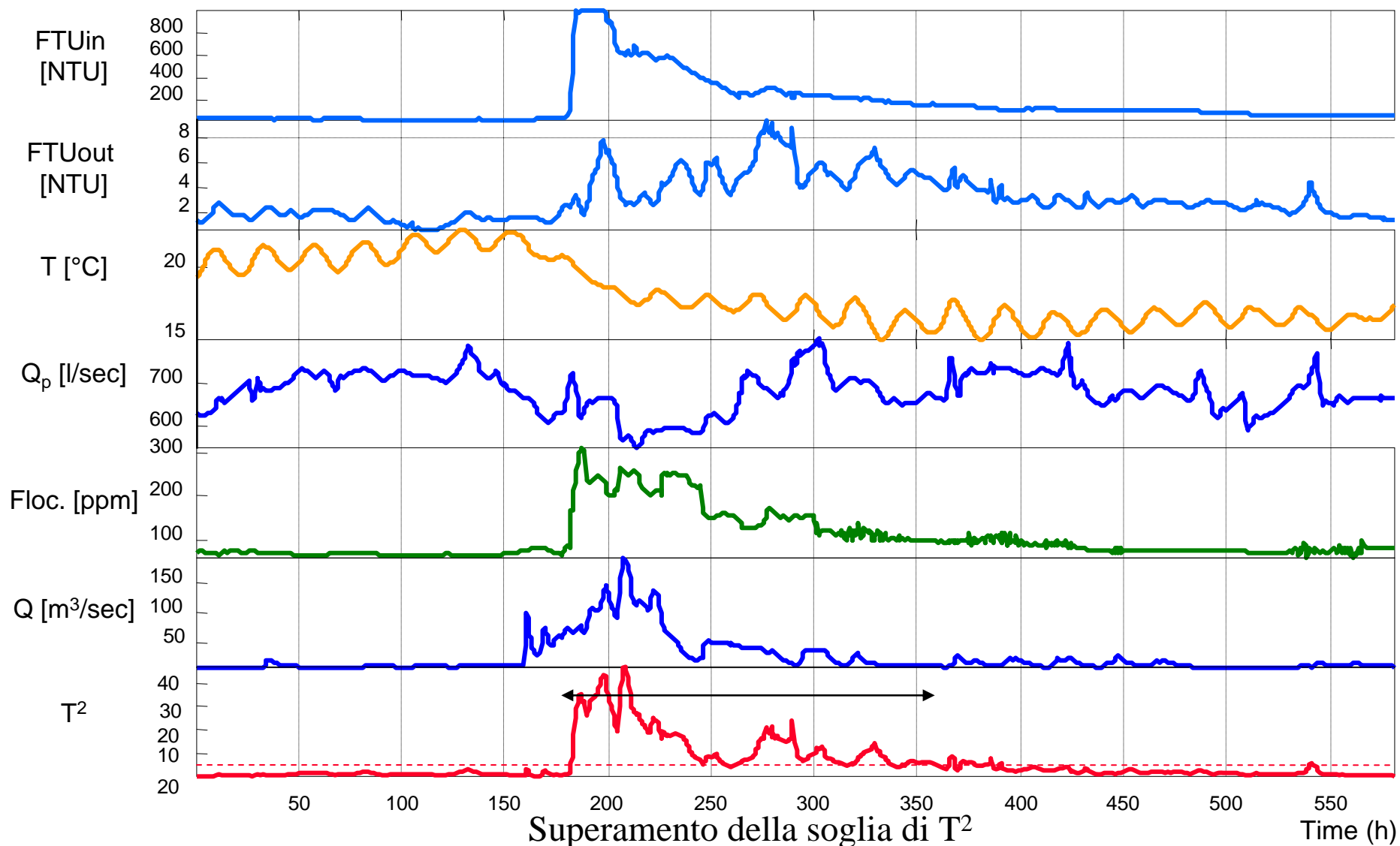
dato che le prime tre variabili sono associate positivamente a fattori di soddisfazione del lavoro, mentre gli ultimi due sono associati all'assistenza medica e sono negativi, z_2 può essere vista come un contrasto fra la soddisfazione del lavoro e l'insoddisfazione dell'assistenza medica.

Il Biplot come visualizzazione dei contributi

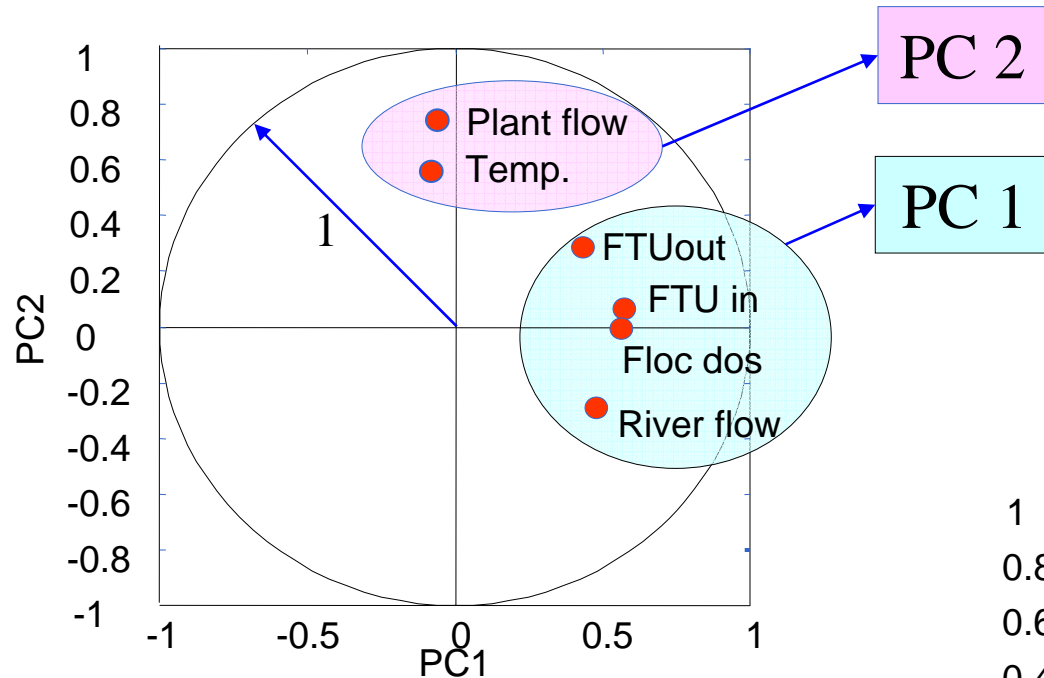
- ➡ Per visualizzare in che misura ogni variabile originaria contribuisce alle PC, si plottano le componenti dei *loadings* delle prime 2 colonne della matrice W , corrispondenti alle prime PC, le più importanti
- ➡ Si possono anche avere biplot tridimensionali
- ➡ La sintassi del comando Matlab è
`biplot(coefs, ..., 'Scores', scores, 'ObsLabels', obslabs)`
dove `coefs` sono le prime 2 o 3 colonne di W , `Scores` sono i dati trasformati (se presenti) e `ObsLabels` sono le etichette che identificano ciascuna variabile



Analisi di un episodio di torbidità



Biplot come zone di influenza delle PC

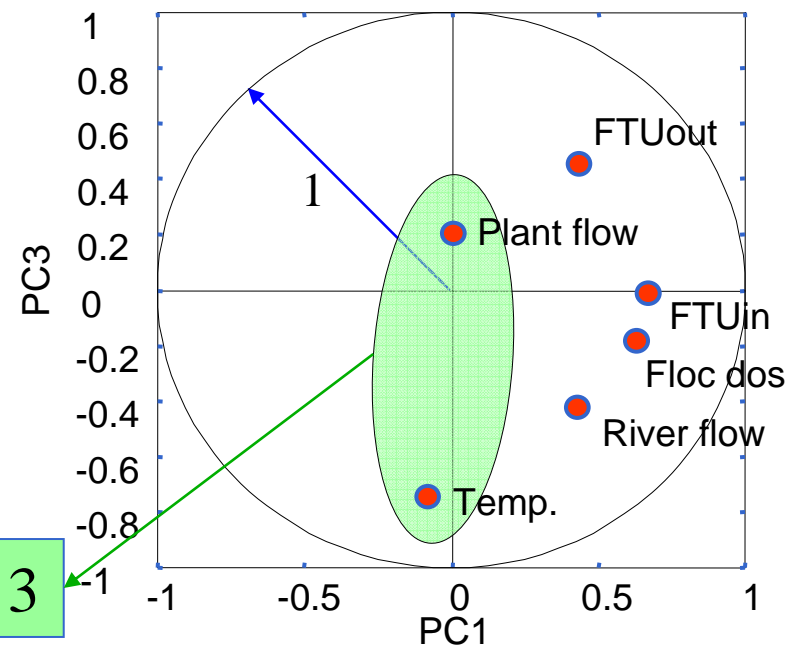


In questo caso si nota che le componenti *FTUin* e *Floc dos* contribuiscono fortemente a PC1 e quasi niente a PC2.

Al contrario *Plant flow* e *Temp* sono molto rappresentate in PC2 e PC3

Biplot indica quanto ciascuna variabile contribuisca a ciascuna PC. Si ottiene plottando i *loadings* (componenti degli autovettori) nel piano di due PC si ha un'idea di

Il **Biplot** più comune è quello PC1/PC2



PCA con Covarianza o Correlazione?

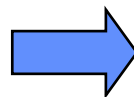
Le PCA ottenute nei due casi sono molto diverse e non facilmente riconducibili l'una all'altra perché i loro autovalori e autovettori non sono legati da una semplice relazione

Covarianza

- ☞ Riflette le reali proporzioni fra variabili
- ☞ E' sensibile alle unità di misura, enfatizzando l'importanza delle variabili più grandi

Esempio: Supponendo di effettuare una PCA su misure di lunghezza (x_1) e peso (x_2), a seconda che x_1 sia espresso in cm (a) o in mm (b), mentre x_2 è sempre espresso in grammi, si ha nei due casi una PCA molto diversa. Nel secondo caso la dominanza di x_1 è totale

$$C_a = \begin{bmatrix} 80 & 44 \\ 44 & 80 \end{bmatrix} \quad C_b = \begin{bmatrix} 8000 & 44 \\ 44 & 80 \end{bmatrix}$$



$$z_1^{(a)} = 0.707 \cdot x_1 + 0.707 \cdot x_2$$
$$z_1^{(b)} = 0.998 \cdot x_1 + 0.055 \cdot x_2$$



Come ridurre l'ordine della PCA

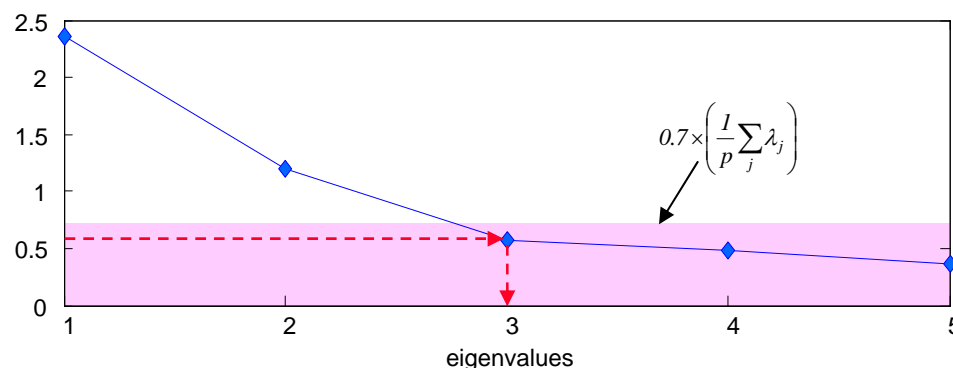
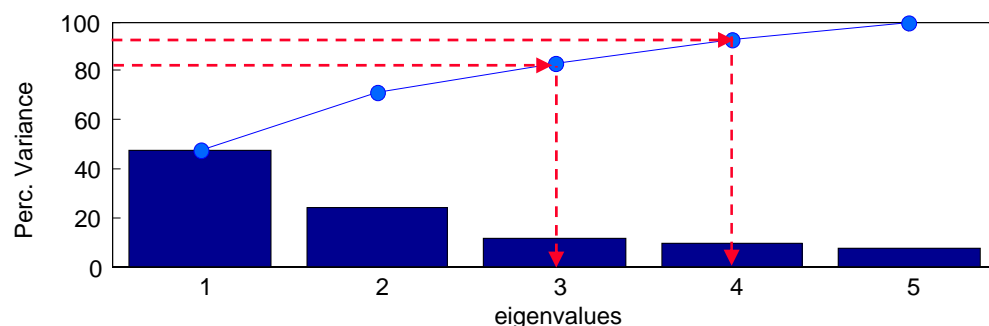
☞ Ci si basa sulla quantità di varianza spiegata: si trattengono le componenti che forniscono una varianza totale spiegata fra il 70% e il 90%

⇒ Ovviamente si arrotonda all'intero più vicino

☞ Alternativamente (o insieme) si “taglia” all'autovalore un po' inferiore ad 1

⇒ Nel caso di matrice di correlazione si taglia intorno a $\lambda \cong 0.7$

⇒ Nel caso di matrice di covarianza si taglia intorno a circa 0.7 della media degli autovalori



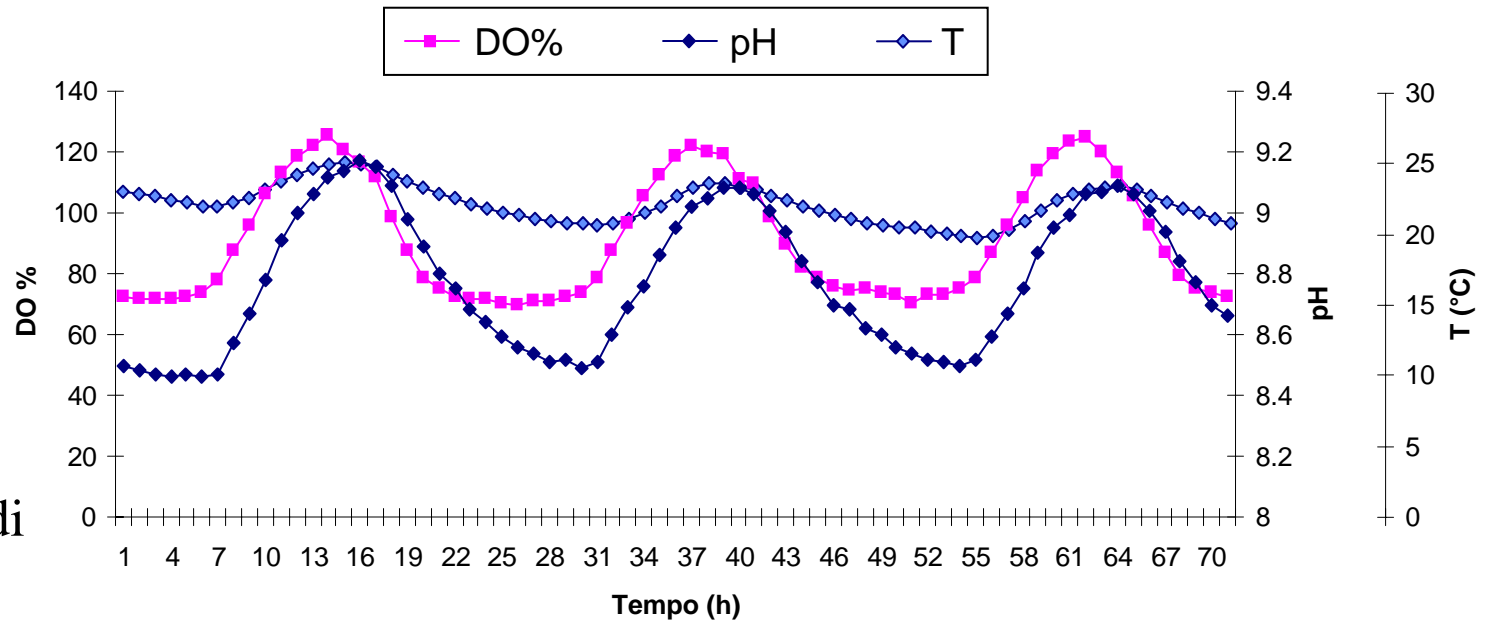
$$\lambda_{min} \cong 0.7 \times \left(\frac{1}{p} \sum_j \lambda_j \right) = 0.7 \times \bar{\lambda}$$

Esempio di dati correlati

Dati di
qualità
dell'acqua
del fiume
Arno

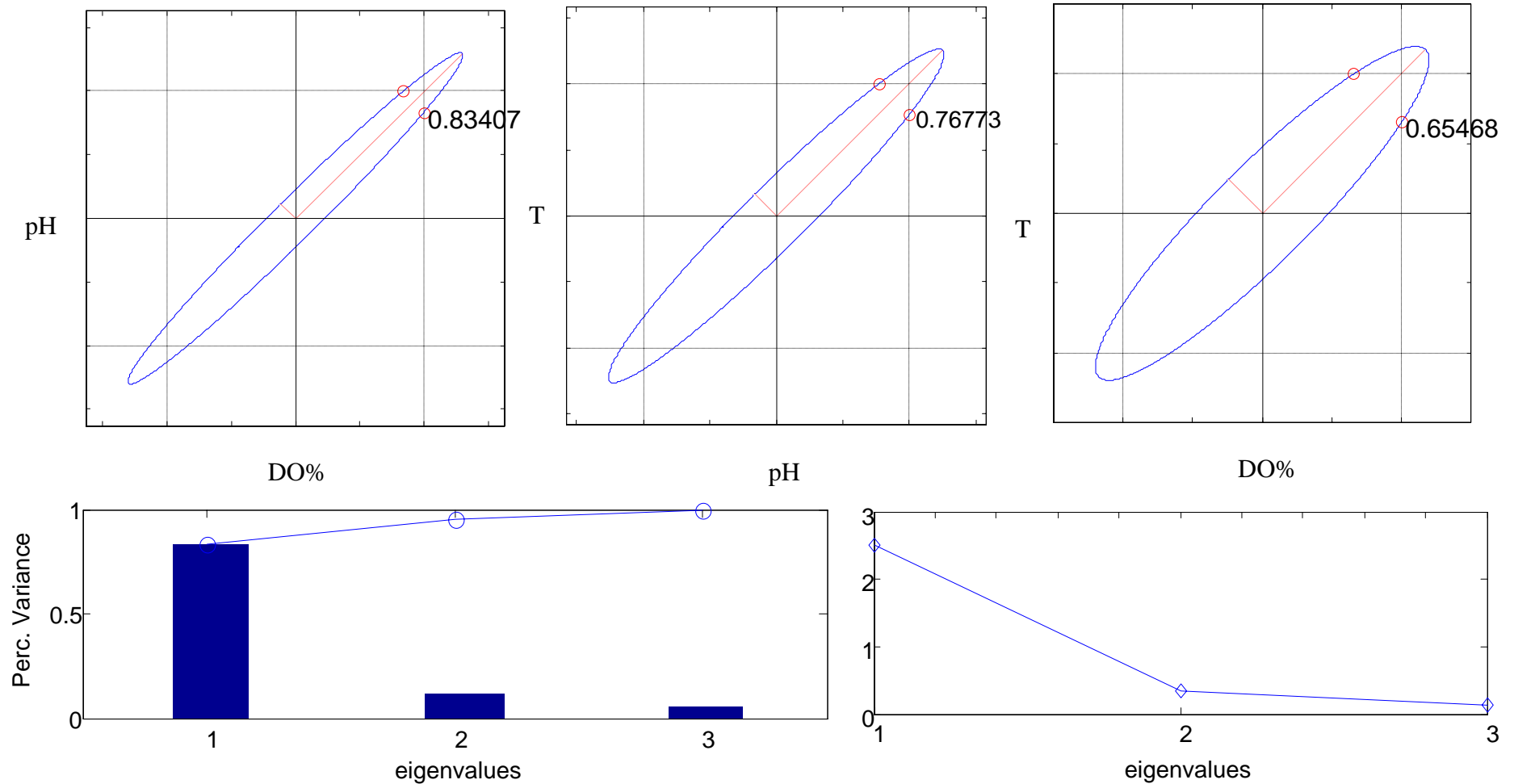
Centralina di
Rosano

5 -7
Settembre
2004



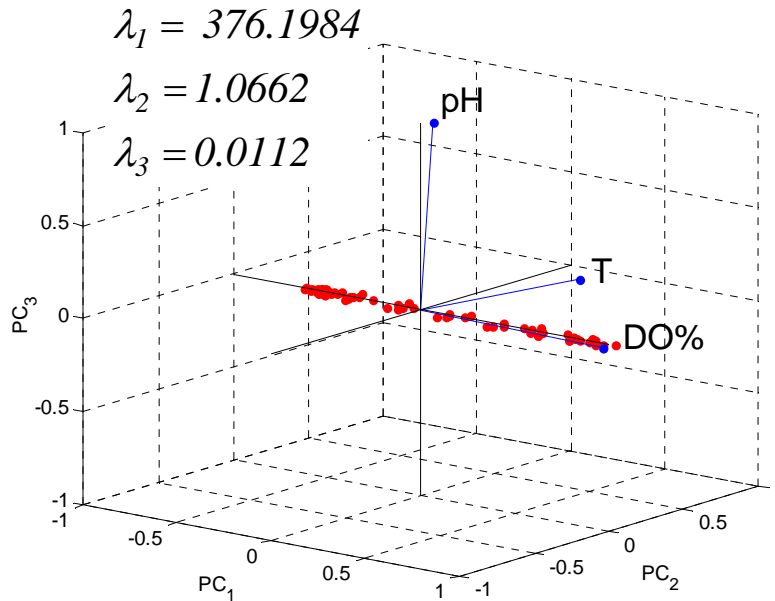
- ☐ Le tre variabili seguono tutte il ritmo circadiano, pilotato dalla luce solare. La correlazione fra di esse è evidente.
- ☐ L'ossigeno disciolto è il più pronto, seguito dalla temperatura a 2 h e dal pH a 3 h di ritardo.
- ☐ L'aumento del pH è dovuto al consumo di CO_2 a seguito della fotosintesi

Correlazioni e autovalori



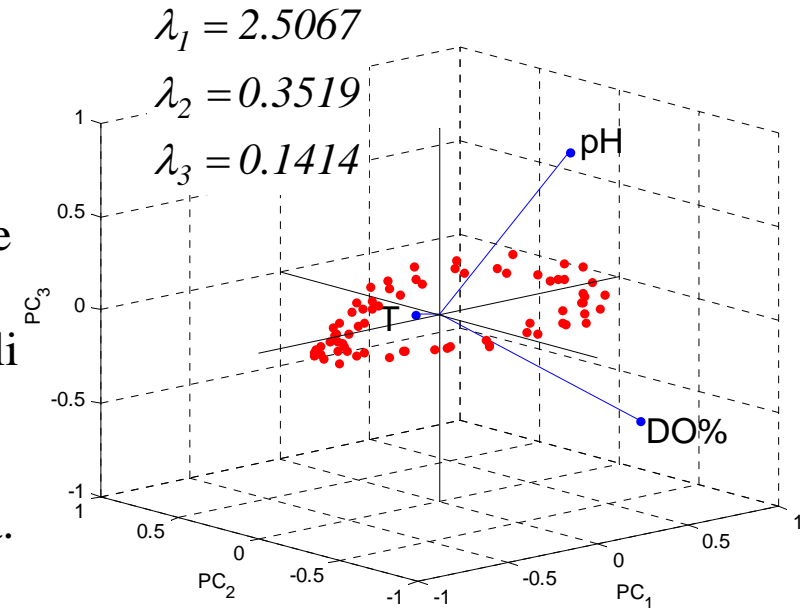
La forte correlazione fra le prime due variabili, fa sì che PC1 spieghi circa 83% della variabilità totale. Analogamente il primo autovalore è dominante rispetto agli altri due.

Meglio l'analisi di correlazione



L'analisi di covarianza tende ad enfatizzare le dipendenze fra le variabili ed a mantenere solo la prima.

L'analisi di correlazione evidenzia il legame delle tre, pur mantenendo il carattere dominante di DO%



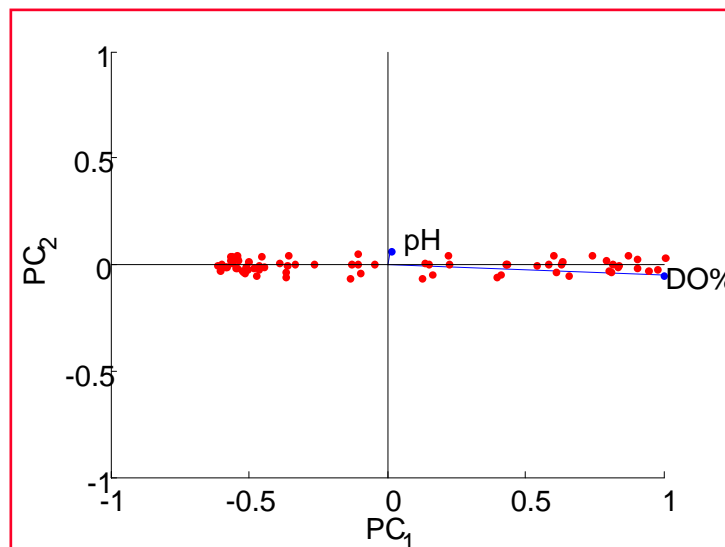
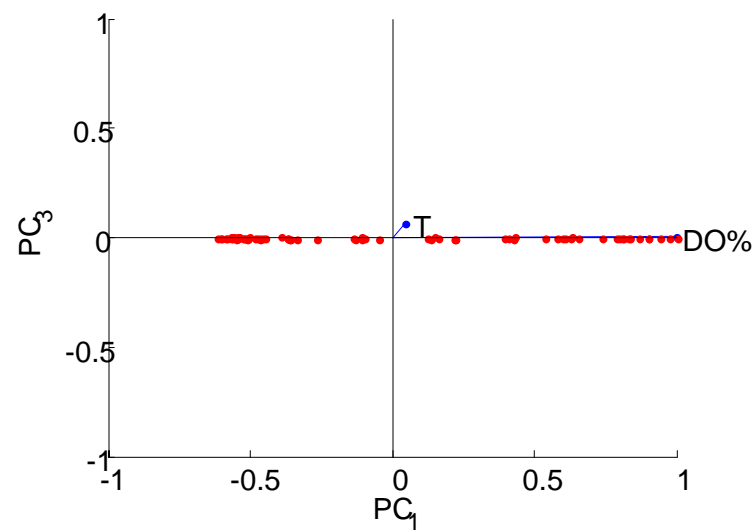
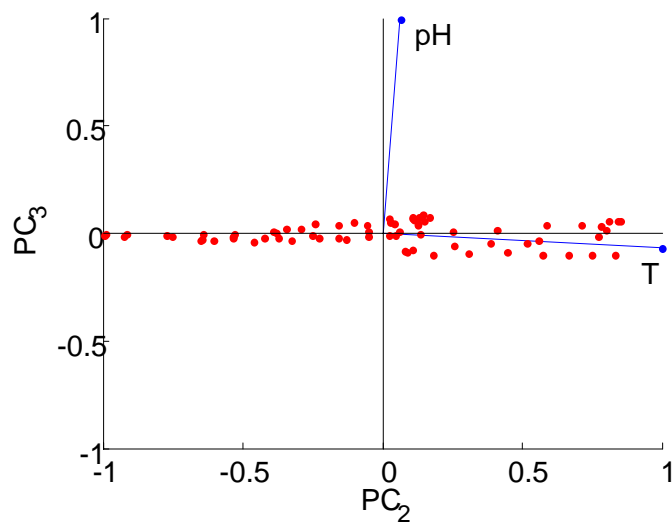
$$C = \begin{bmatrix} 375.3621 & 3.6624 & 17.3091 \\ 3.6624 & 0.0514 & 0.2374 \\ 17.3091 & 0.2374 & 1.8623 \end{bmatrix}$$

$$PC_{dati} = \begin{bmatrix} 0.9989 & 0.0467 & -0.0067 \\ 0.0098 & -0.0646 & 0.9979 \\ 0.0462 & -0.9968 & -0.0650 \end{bmatrix}$$

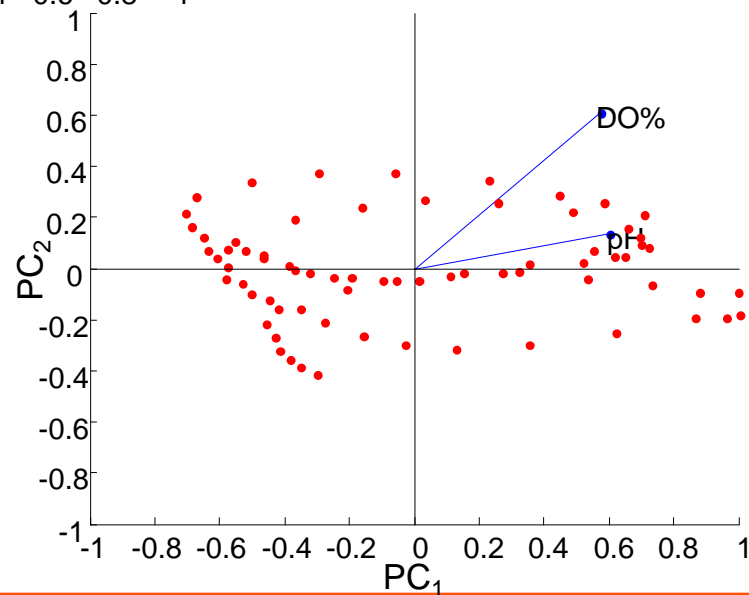
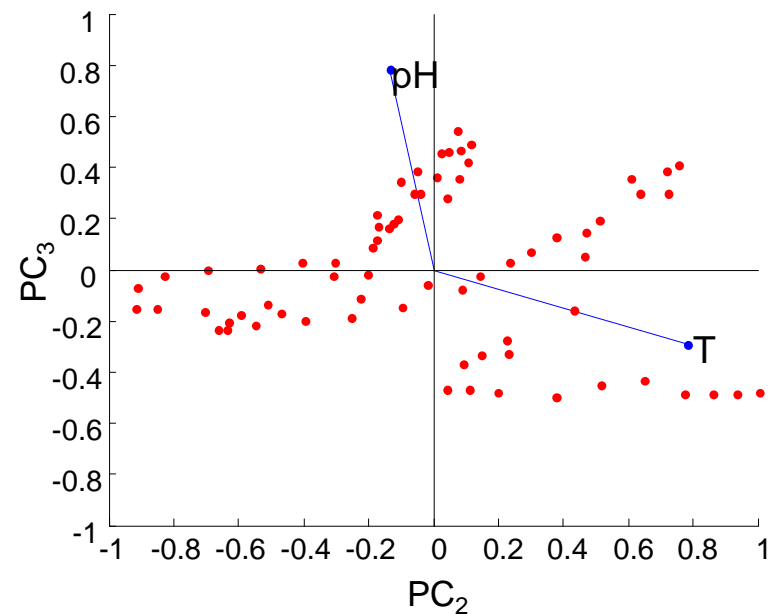
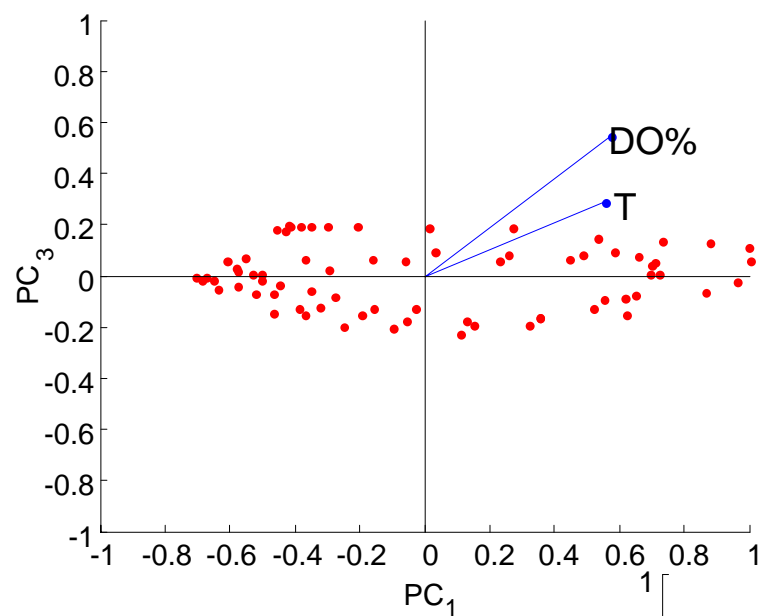
$$R = \begin{bmatrix} 1.0000 & 0.8341 & 0.6547 \\ 0.8341 & 1.0000 & 0.7677 \\ 0.6547 & 0.7677 & 1.0000 \end{bmatrix}$$

$$PC_R = \begin{bmatrix} 0.5743 & 0.6111 & -0.5448 \\ 0.6011 & 0.1371 & 0.7873 \\ 0.5558 & -0.7796 & -0.2886 \end{bmatrix}$$

Biplot proiettato sulle tre PC (cov)



Biplot proiettato sulle tre PC (corr)



Test di consistenza

👉 Le componenti principali non avendo un significato immediato devono essere spiegate con statistiche di sintesi:

👉 *Hotelling's T^2*

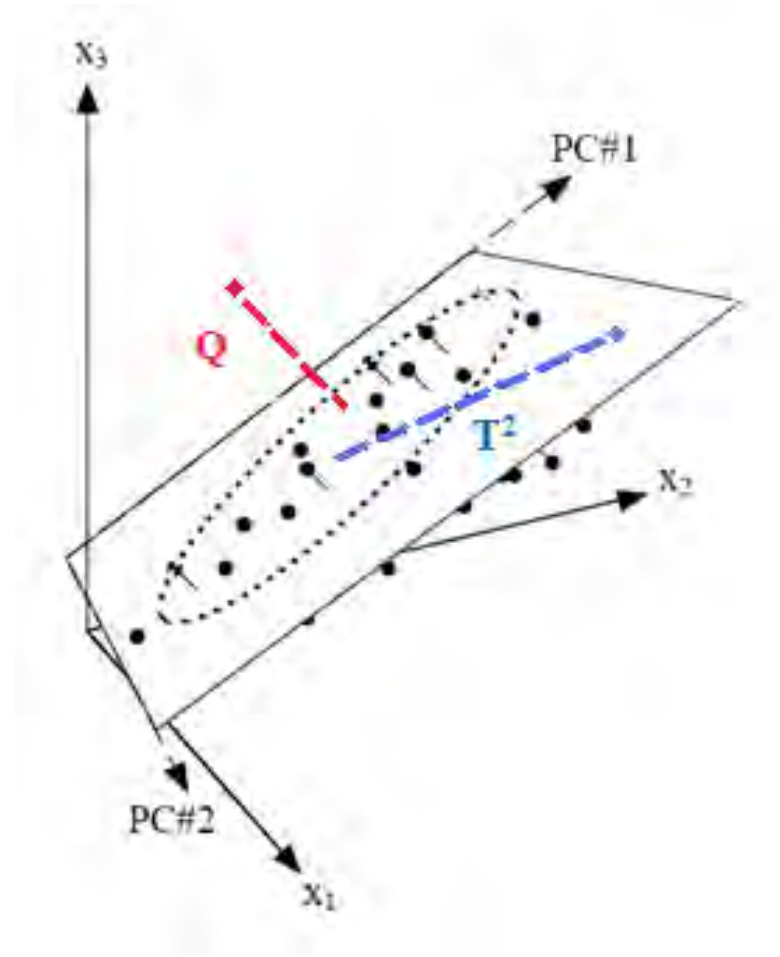
$$T^2 = \mathbf{Z}^T \mathbf{L}^{-1} \mathbf{Z} = \mathbf{X}^T \mathbf{W} \mathbf{L}^{-1} \mathbf{W}^T \mathbf{X}$$

coglie la variazione delle componenti all'interno del modello di riferimento (spazio delle PC)

👉 *Statistica Q*

$$Q = \mathbf{X}^T (\mathbf{I} - \mathbf{W} \mathbf{W}^T) \mathbf{X}$$

misura la variazione non considerata dal modello (spazio ortogonale alle PC)



La statistica Hotelling T^2

- ☞ E' un'estensione multivariabile della statistica Student t .
- ☞ Data una matrice di misure $\mathbf{X} \in \mathbb{R}^{n \times p}$ e \mathbf{W} la loro matrice di covarianza, la statistica Hotelling T^2 è data da

$$T^2 = \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}$$

- ☞ Inoltre, volendola usare come limite di accettabilità (hypothesis testing) si può definire un valore limite di T^2 in funzione della statistica F di cui T^2 rappresenta una realizzazione

$$T_{lim}^2 = \frac{p(n-1)}{n-p} F_{n-p, p}^{\alpha}$$

- ➡ Dove n è il numero di misure e p le variabili di ciascuna misura, mentre α è il limite di confidenza fissato (generalmente $\alpha = 0.05$, corrispondente ad una confidenza del 95%)

Andamento della statistica F

Esempio:

Il valore limite di T^2 al 95% per un campione di $n = 18$ misure di $p = 8$ variabili ciascuna è

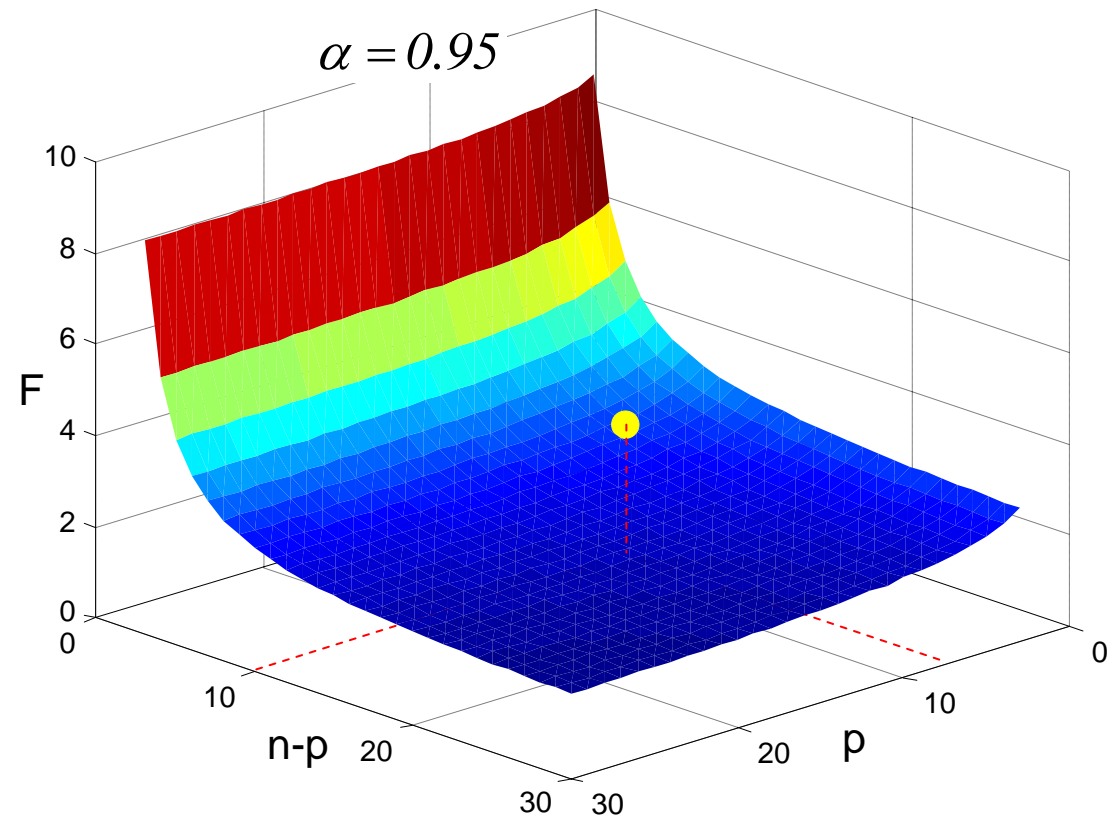
```
x = finv(0.95,8,18-8)
```

```
x = 3.0717
```

Ciò significa che si può osservare per puro caso un valore di F superiore a 4.3468 solamente nel 5% dei casi

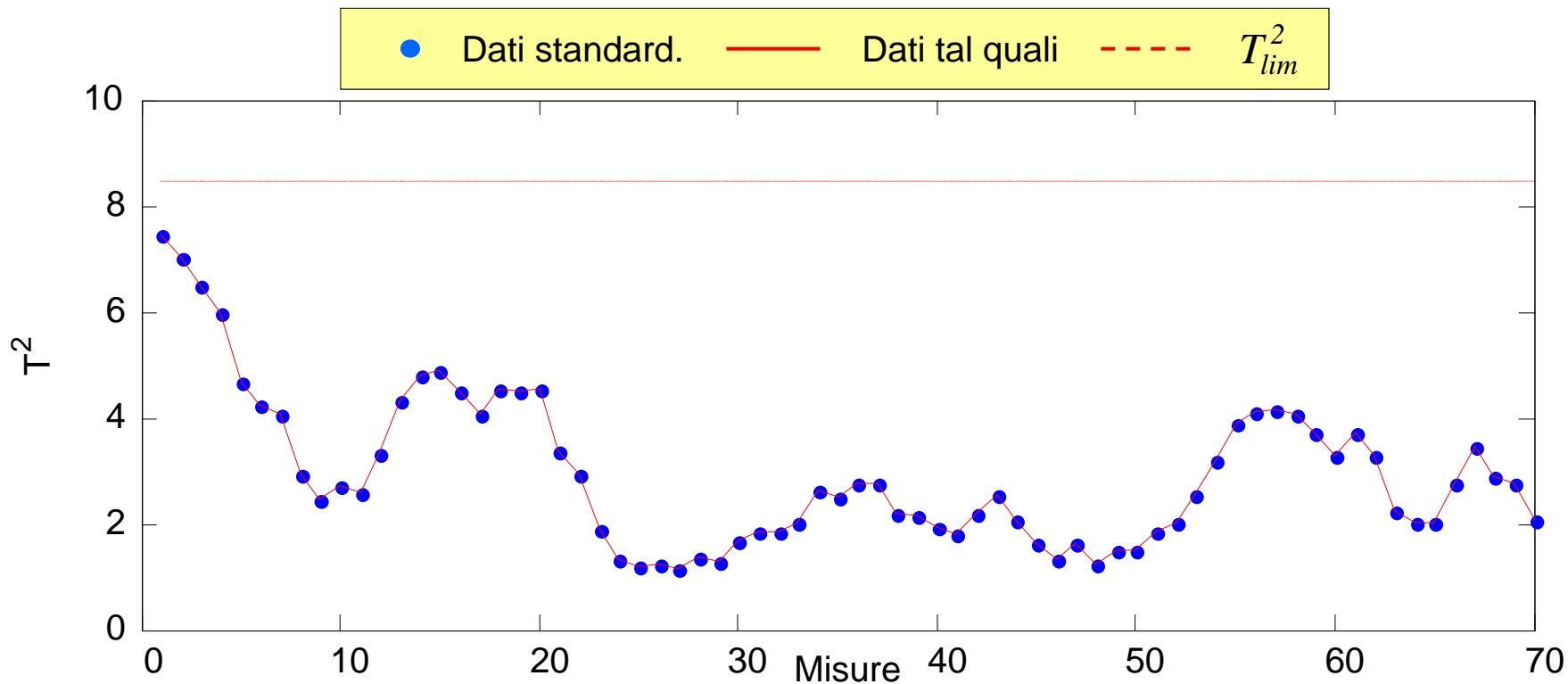
In questo caso il T^2_{lim} sarebbe

$$T^2_{lim} = \frac{8(18-1)}{18-8} 3.0717 = 0.41775$$



Applicazione del T^2 ai dati di qualità

- 👉 L'indice Hotelling's T^2 è lo stesso sia per i dati originali che per quelli standardizzati
- 👉 Ovviamente l'indice per ogni dato è inferiore a T^2_{lim} perché sono gli stessi dati usati per determinare le PC



Riduzione della dimensionalità

- ☞ Se alcune PC non sono molto importanti (vedi scree plot) si può ridurre la dimensionalità dell'analisi trattenendo solamente le prime $a < p$ PC
- ☞ La matrice di trasformazione è la sotto matrice di \mathbf{W} che ritiene i primi a autovettori

$$\mathbf{W} = \begin{array}{c} \xleftarrow{a} \quad \quad \quad \xrightarrow{p-a} \\ \left[\begin{array}{cccc|cccc} w_{11} & w_{12} & \dots & w_{1a} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2a} & \dots & w_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{p1} & w_{p2} & \dots & w_{pa} & \dots & w_{pp} \end{array} \right] \end{array} \quad \mathbf{W}_a = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1a} \\ w_{21} & w_{22} & \dots & w_{2a} \\ \dots & \dots & \dots & \dots \\ w_{p1} & w_{p2} & \dots & w_{pa} \end{bmatrix} \in \mathbb{R}^{n \times a}$$

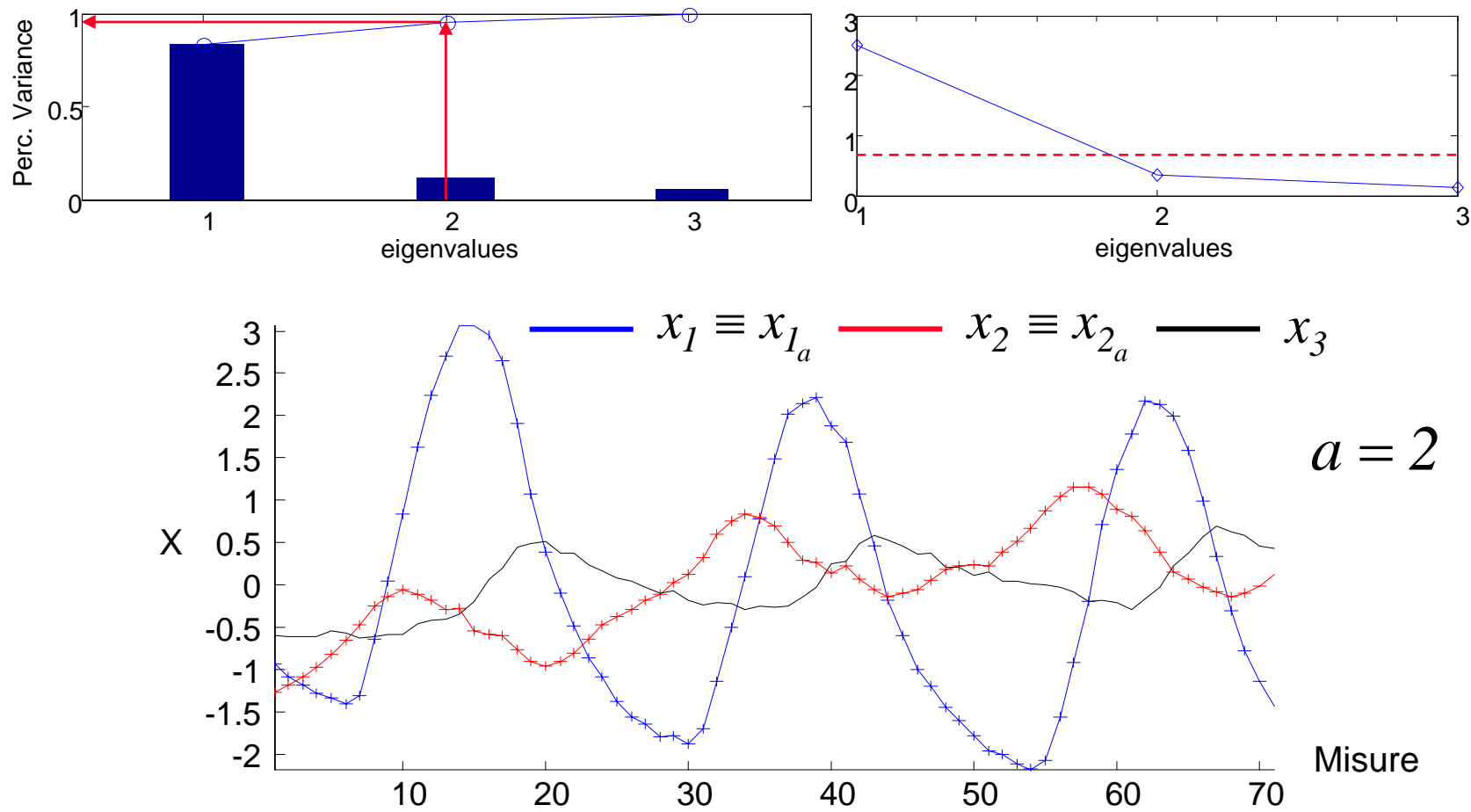
- ☞ La trasformazione PCA ridotta diviene

$$\mathbf{Z}_a = \mathbf{X} \cdot \mathbf{W}_a \in \mathbb{R}^{n \times a}$$

ovvero i dati \mathbf{X} vengono proiettati nelle prime a componenti di \mathbf{Z}

Decomposizione parziale

La PCA ridotta contiene solo le prime due componenti (vedi scree plot) che comunque spiegano 83% della variabilità totale.



Ricostruzione da PCA ridotta

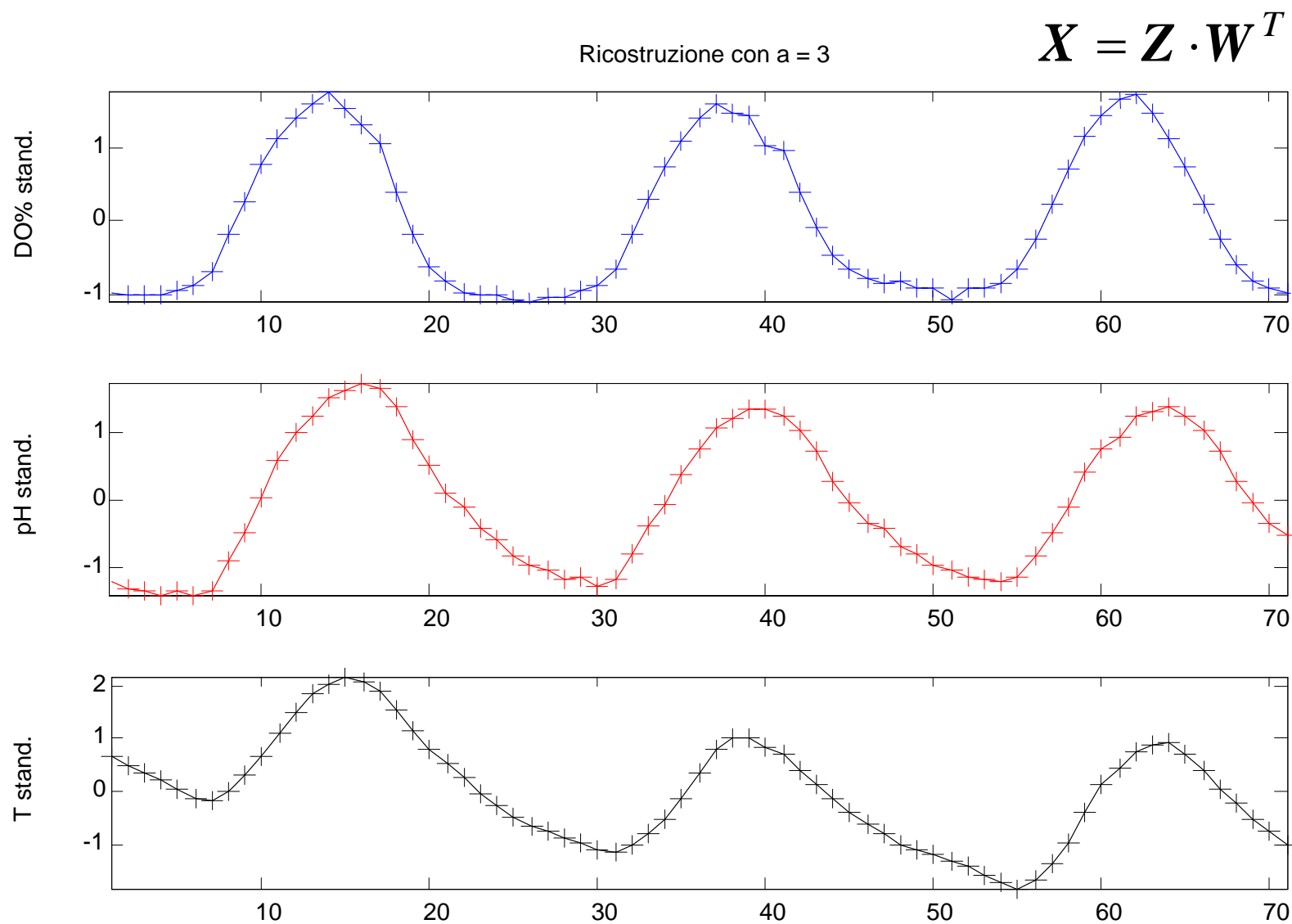
- 👉 La ricostruzione delle variabili originali dalla PCA ridotta contiene ovviamente degli errori
- 👉 Se la mappa inversa completa era $X = Z \cdot W^T$
- 👉 nel caso ridotto sarà $X_a = Z_a \cdot W_a^T \Rightarrow X = Z_a \cdot W_a^T + E$
- 👉 La matrice E contiene gli errori di ricostruzione

$$\begin{array}{c} n \\ \boxed{X} \\ p \end{array} = \underbrace{\begin{array}{c} \boxed{Z_a} \\ a \end{array} \times \begin{array}{c} \boxed{W_a} \\ p \end{array}}_{\text{modello ridotto}} + \underbrace{\begin{array}{c} \boxed{E} \\ p \end{array}}_{\text{residui}} \begin{array}{c} n \end{array}$$

The diagram shows the equation $X = Z_a \cdot W_a^T + E$ with dimensions and labels. Matrix X is pink, Z_a is light blue, W_a is yellow, and E is light green. Dimensions n and p are indicated for X , a for Z_a , p for W_a , and p for E . A bracket under Z_a and W_a labels them as 'modello ridotto', and a bracket under E labels it as 'residui'.

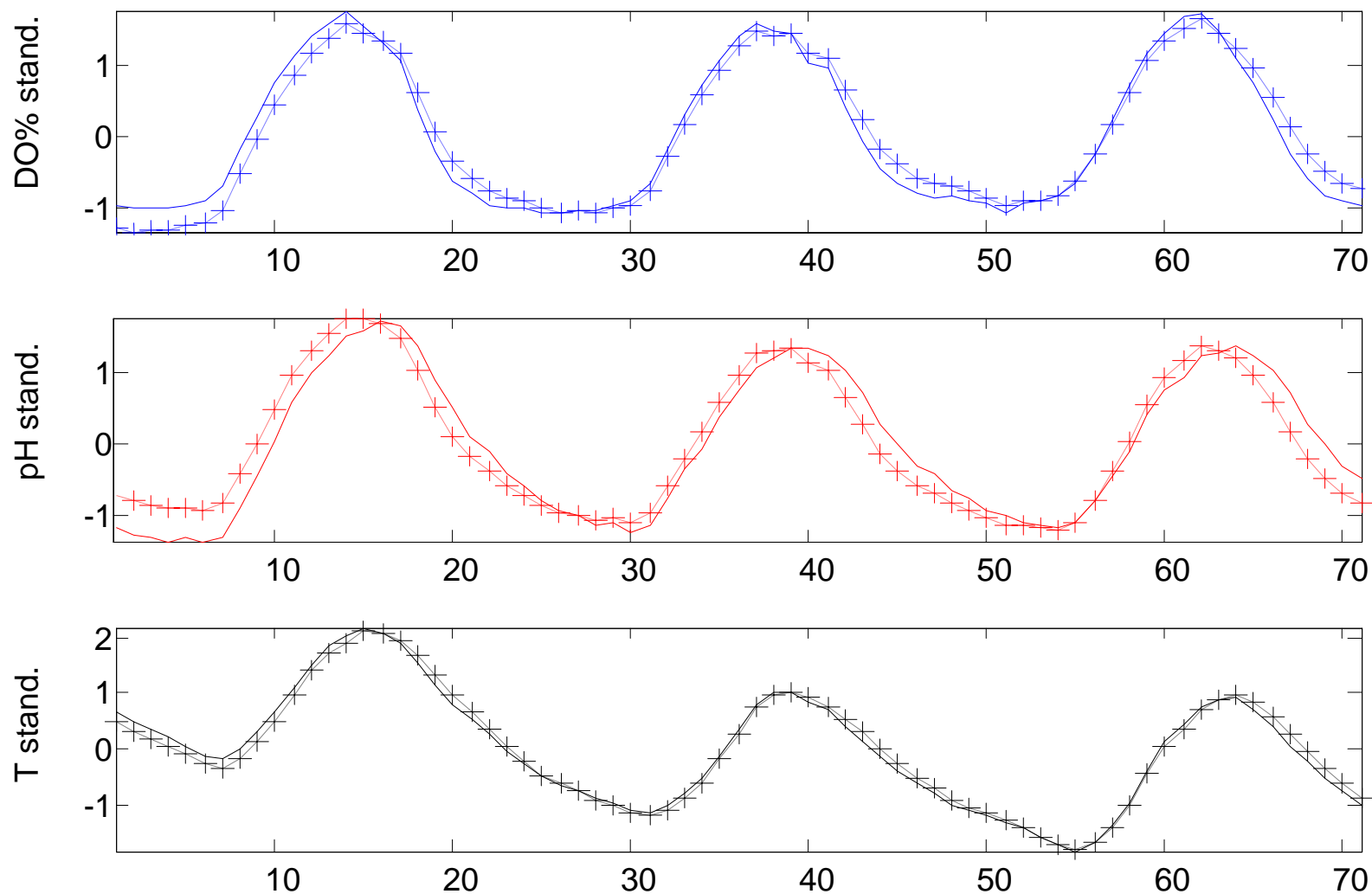
👉 Matlab: `[residuals,reconstructed]=pcarec(X,a);`

Ricostruzione da PCA completa



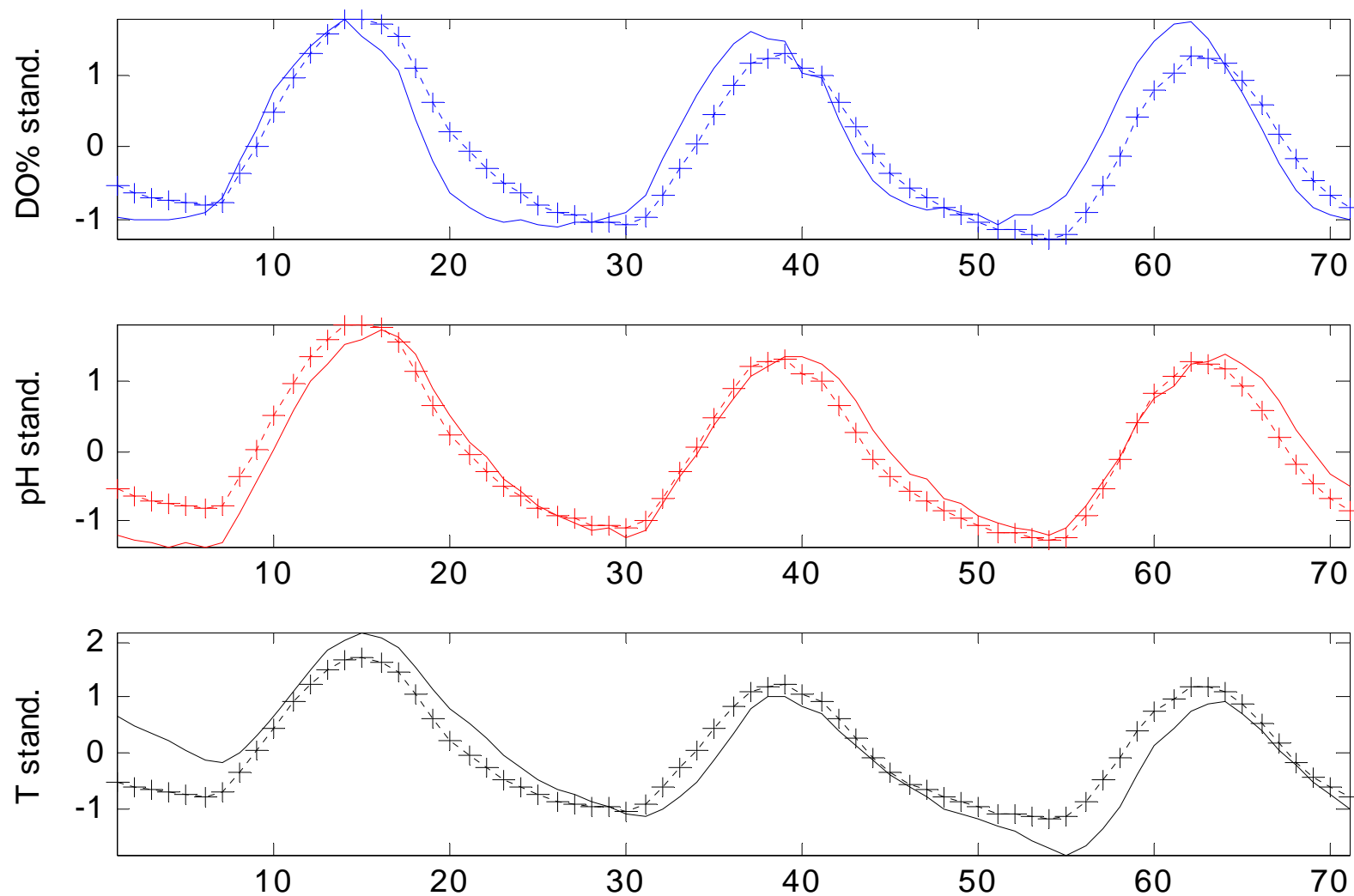
Ricostruzione da PCA ridotta

Ricostruzione con $a = 2$
$$\mathbf{X} = \mathbf{Z}_a \cdot \mathbf{W}_a^T + \mathbf{E}$$



Ricostruzione da PCA ridotta

Ricostruzione con $a = 1$
$$\mathbf{X} = \mathbf{Z}_a \cdot \mathbf{W}_a^T + \mathbf{E}$$



La statistica Q

- ☞ Valuta l'importanza delle PC escluse dall'analisi
- ☞ Ha significato solo quando si esegue una riduzione ($a < p$)
- ☞ Il valore limite della statistica Q per il quantile $c_\alpha=1-\alpha$ è dato da

$$Q_{lim} = \mathcal{G}_1 \left[1 + \frac{h_o c_\alpha \sqrt{2\mathcal{G}_2}}{\mathcal{G}_1} + \frac{\mathcal{G}_2 h_o (h_o - 1)}{\mathcal{G}_1^2} \right]^{\frac{1}{h_o}}$$

$$\mathcal{G}_1 = \sum_{i=a+1}^p \lambda_i$$

$$\mathcal{G}_2 = \sum_{i=a+1}^p \lambda_i^2$$

$$\mathcal{G}_3 = \sum_{i=a+1}^p \lambda_i^3$$

$$h_o = 1 - 2 \frac{\mathcal{G}_1 \mathcal{G}_2}{3\mathcal{G}_3}$$