Facoltà: : BioScienze e Tecnologie Agro-Alimentari e Ambientali

Denominazione Corso di Laurea: Biotecnologie Avanzate (Laurea Magistrale)

Corso: Statistica e bioinformatica per le biotecnologie

MODULO:

Chemometria applicata (5 CFU, 40 ore)

Docente: Marcello Mascini

(mmascini@unite.it)

Il Docente e' disponibile per chiarimenti al termine della lezione o su richiesta via mail

2UD Analisi multivariata “unsupervised” (2 CFU = 16 ore)

Correlazione; Covarianza; Multiple linear regression; Principal component analysis (PCA); Multiple correspondence analysis (MCA); K-means clustering; Agglomerative hierarchical clustering. Esempi di elaborazione dati “unsupervised”.

Sistemi Lineari e Matrici

- Un sistema lineare è un insieme di n equazioni lineari in n incognite
- È noto che le incognite possono essere determinate con semplicità utilizzando il formalismo matriciale

$$\begin{cases} a \cdot x + b \cdot y = k \\ c \cdot x + d \cdot y = g \end{cases} \Rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} k \\ g \end{bmatrix} \Rightarrow \boldsymbol{A} \cdot x = k \Rightarrow x = \boldsymbol{A}^{-1} \cdot k$$

- Il sistema quindi ammette soluzione, cioè è invertibile, se la matrice A ammetta la sua inversa quindi se det(A)≠0
- Quindi se tutte le equazioni sono linearmente indipendenti una rispetto all'altra

Formalismo matriciale

$$y_i = a \cdot x_i + b \Rightarrow \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{bmatrix} \Rightarrow Y_{1xn} = K_{1x2} \cdot X_{2xn}$$

esempio: n = 2 senza errori di misura

$$Y_{1x2} = K_{1x2} \cdot X_{2x2} \Rightarrow K_{1x2} = Y_{1x2} \cdot X^{-1}_{2x2}$$

- X^{-1} è detta matrice inversa di X

$$X^{-1}_{2x2} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} \frac{d}{\det X} & \frac{c}{\det X} \\ \frac{b}{\det X} & \frac{a}{\det X} \end{bmatrix}$$

- Inversamente proporzionale al determinante
  - Se due righe o colonne sono proporzionali o combinazione lineare delle altre il determinante della matrice è 0 e la matrice inversa diverge!
    - Problema della co-linearità nella soluzione dei problemi lineari.

In presenza di errori di misura

$$y_i = a \cdot x_i + b + e_i \Rightarrow \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix} = \begin{bmatrix} a & b \end{bmatrix} \cdot \begin{bmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{bmatrix} + \begin{bmatrix} e_1 & \dots & e_n \end{bmatrix}$$

$$\Rightarrow Y_{1xn} = K_{1x2} \cdot X_{2xn} + E_{1xn}$$

- In presenza di errori di misura ci sono n righe indipendenti con n>2
- La soluzione minimi quadrati, in cui cioè la norma del vettore E_{1xn} si ottiene, formalmente come nel caso precedente scambiando l'operazione di inversione con quella di pseudo-inversione o inversa generalizzata (teorema di Gauss-Markov)

esempio: n = 2 con errori di misura

$$Y_{1xn} = K_{1x2} \cdot X_{2xn} + E_{1xn}; \quad \min \left\| E_{1xn} \right\| \Rightarrow K_{1x2} = Y_{1xn} \cdot X^{+}_{nx2}$$

Matrice Pseudo-inversa

- a.k.a. inversa generalizzata o inversa di Moore
- Definita attraverso le relazioni di Moore:

$$X \cdot X^{+} \cdot X = X \qquad \left(X \cdot X^{+}\right)^{*} = X \cdot X^{+}$$

$$X^{+} \cdot X \cdot X^{+} = X^{+} \qquad \left(X^{+} \cdot X\right)^{*} = X^{+} \cdot X$$

- Calcolo pratico:

$$X^{+} = \left(X^{T} X\right)^{-1} X^{T}$$

  - Esiste se $det\ X^{T}X \neq 0$.
- Pinv(x) in Matlab
- PseudoInverse[x] in Mathematica

Regressione Polinomiale

- Una relazione funzionale polinomiale può essere scritta in forma matriciale. Il problema ai minimi quadrati si può risolvere applicando il teorema di Gauss-Markov
- Polinomio di grado n; n+1 parametri; m>n+1 misure sperimentali

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

$$\begin{bmatrix} y_1 & \ldots & y_m \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & \ldots & a_n \end{bmatrix} \cdot \begin{bmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_m \\ \ldots & \ldots & \ldots \\ x_1^n & \ldots & x_m^n \end{bmatrix} + \begin{bmatrix} e_1 & \ldots & e_m \end{bmatrix}$$

$$Y_{1xm} = K_{1xn+1} \cdot X_{N+1xm} + E_{1xm}$$

$$\min \left\| E_{1xm} \right\| \Rightarrow K_{1xn+1} = Y_{1xm} X^{+}_{mxn+1}$$

Sistemi Lineari e matrici di dati

- In un sistema lineare, ogni equazione definisce una funzione di più variabili.
- Nella scienza analitica i sistemi lineari consentono di estrarre informazioni da metodi di misura non specifici, cioè nei quali il risultato della misura non è funzione solo di una grandezza ma di più grandezze caratterizzanti un campione.
- Quando però si considerano dati sperimentali bisogna tenere conto di tre fattori:
  - Errori di misura
  - Dipendenza da più grandezze rispetto a quelle considerate
  - Limiti della relazione lineare
- Tutto ciò fa si che ad esempio la risposta di una misura (y) rispetto a due variabili (x_1, x_2) si possa scrivere come:

$$y = k_1 \cdot x_1 + k_2 \cdot x_2 + e$$

  - Il termine e contiene i tre fattori sopra elencati

Sistemi lineari e matrici di dati

- Dato un metodo di misura come il precedente per la misura delle grandezze x_1 e x_2 è necessario almeno disporre di un altro metodo di misura ma con coefficienti differenti

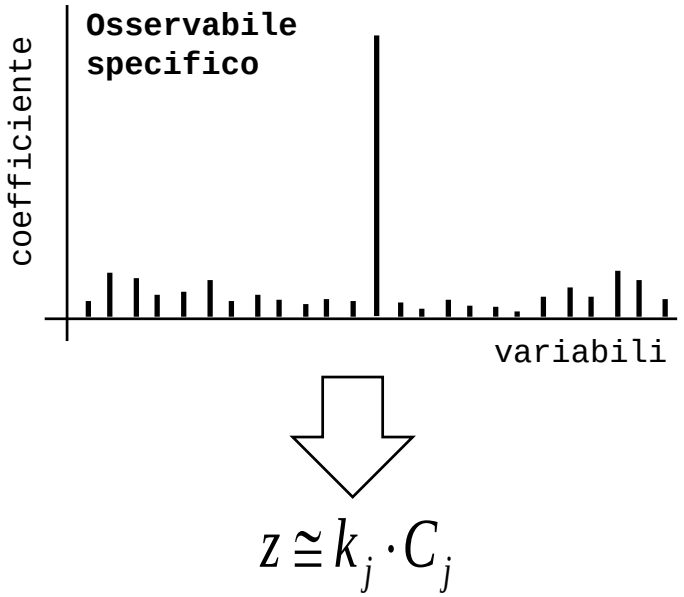
$$y_1 = k_1 \cdot x_1 + k_2 \cdot x_2 + e_1$$

$$y_2 = w_1 \cdot x_1 + w_2 \cdot x_2 + e_2$$

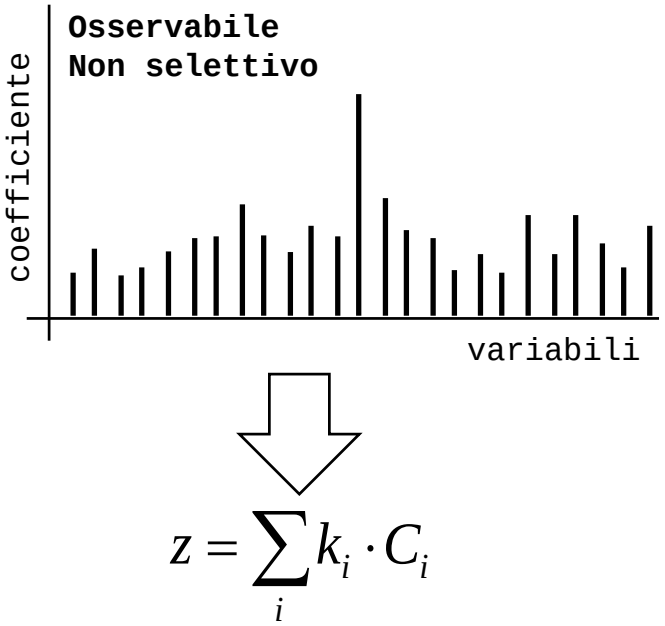
  - Si noti che in assenza dei termini **e** il sistema sarebbe deterministico, e due equazioni sarebbero il massimo necessario per ricavare due variabili
    - Nel senso che ogni altra equazione sarebbe necessariamente combinazione lineare delle prime due
- La presenza dei termini **e** fa si che il problema della determinazione di x_1 e x_2 sia un problema statistico simile al problema della regressione e quindi risolvibile con il metodo dei minimi quadrati
  - In particolare, si possono avere più equazioni che termini incognite anzi maggiore è il numero di equazioni minore è l'errore di stima

Osservabili selettivi e non selettivi

- Le quantità osservabili possono essere, rispetto alle variabili da misurare, selettivi o non selettivi
  - Selettivi: L'osservabile dipende **in maniera dominante** da una variabile
  - Non selettivi: l'osservabile dipende da più variabili
- Gli osservabili non selettivi sono gli oggetti della analisi multivariata



$$z \cong k_j \cdot C_j$$



$$z = \sum_i k_i \cdot C_i$$

Osservabili non selettivi

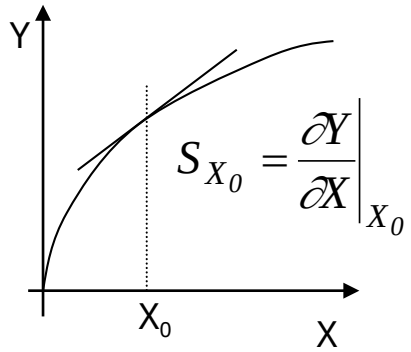
- Esempio:
  - Spettri ottici
    - L'assorbimento ad una data frequenza dipende dalla concentrazione di più specie
  - Gas cromatogrammi
    - L'intensità di una riga può risultare dalla concentrazione di più composti con tempi di eluizione simili
  - Sensori chimici
    - La risposta di una sensore è data dalla combinazione di più sostanze a seconda della loro concentrazione e della loro affinità con il sensore stesso.

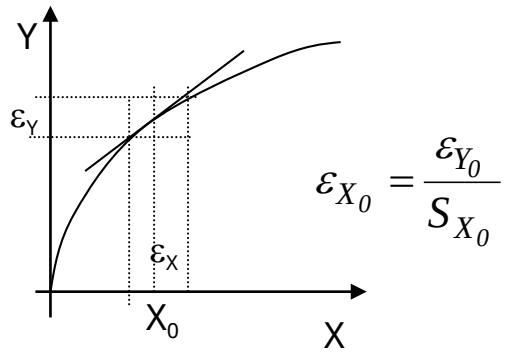
Parametri importanti dei metodi analitici: la sensibilità e la risoluzione

- Dato un metodo analitico, si definisce sensibilità il rapporto tra la variazione del risultato del metodo e la corrispondente variazione della variabile misurata.
  - Tale quantità corrisponde alla seguente derivata: $S = \frac{\partial Y}{\partial X}$

- La risoluzione indica la quantità minima misurabile della variabile considerata, essa è generata dall'errore di misura ed è definita da:

$$R = \varepsilon_X = \frac{\varepsilon_Y}{S}$$



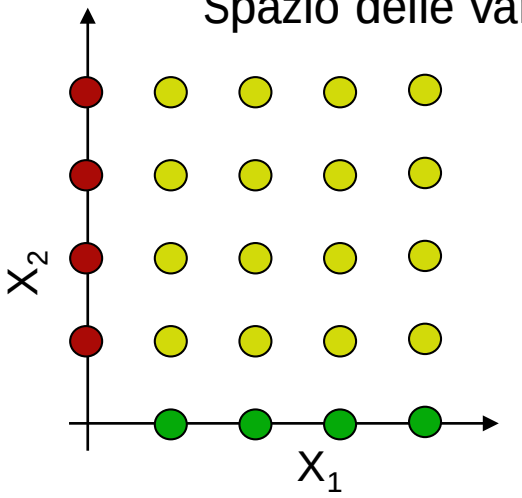


Spazio delle variabili e spazio degli osservabili: caso di osservabili selettivi

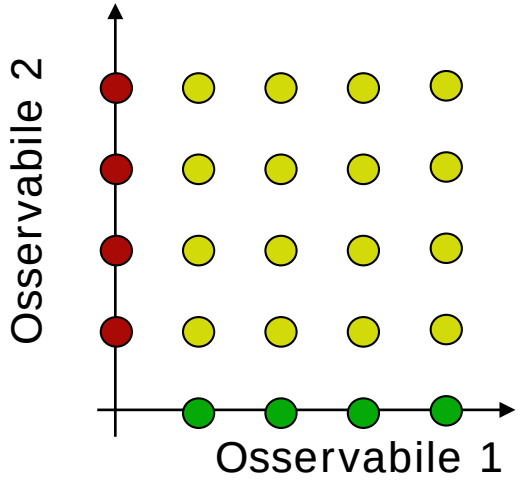
$$Y_1 = aX_1 + bX_2$$

$$Y_2 = cX_1 + dX_2$$

Spazio delle variabili



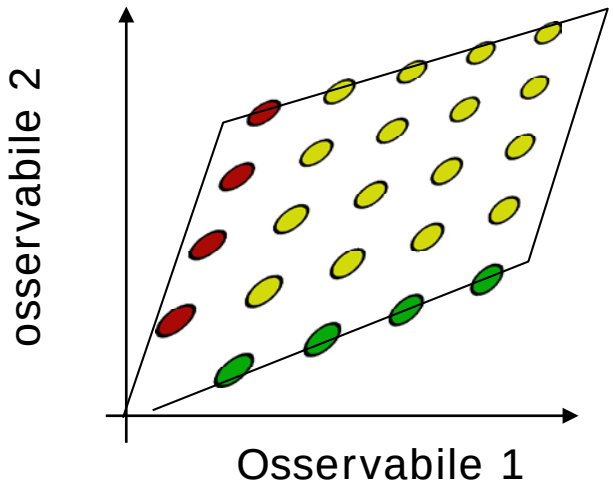
Se I due osservabili sono i



$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

Correlazione=1-det(K)=0

Spazio delle variabili e spazio degli osservabili: caso di osservabili non selettivi

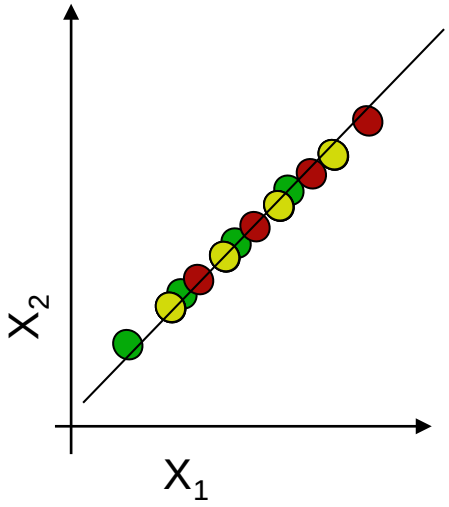


Correlazione parziale

0 < C < 1

$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

$$a, b, c, d \neq 0$$



Correlazione totale

C = 1

$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

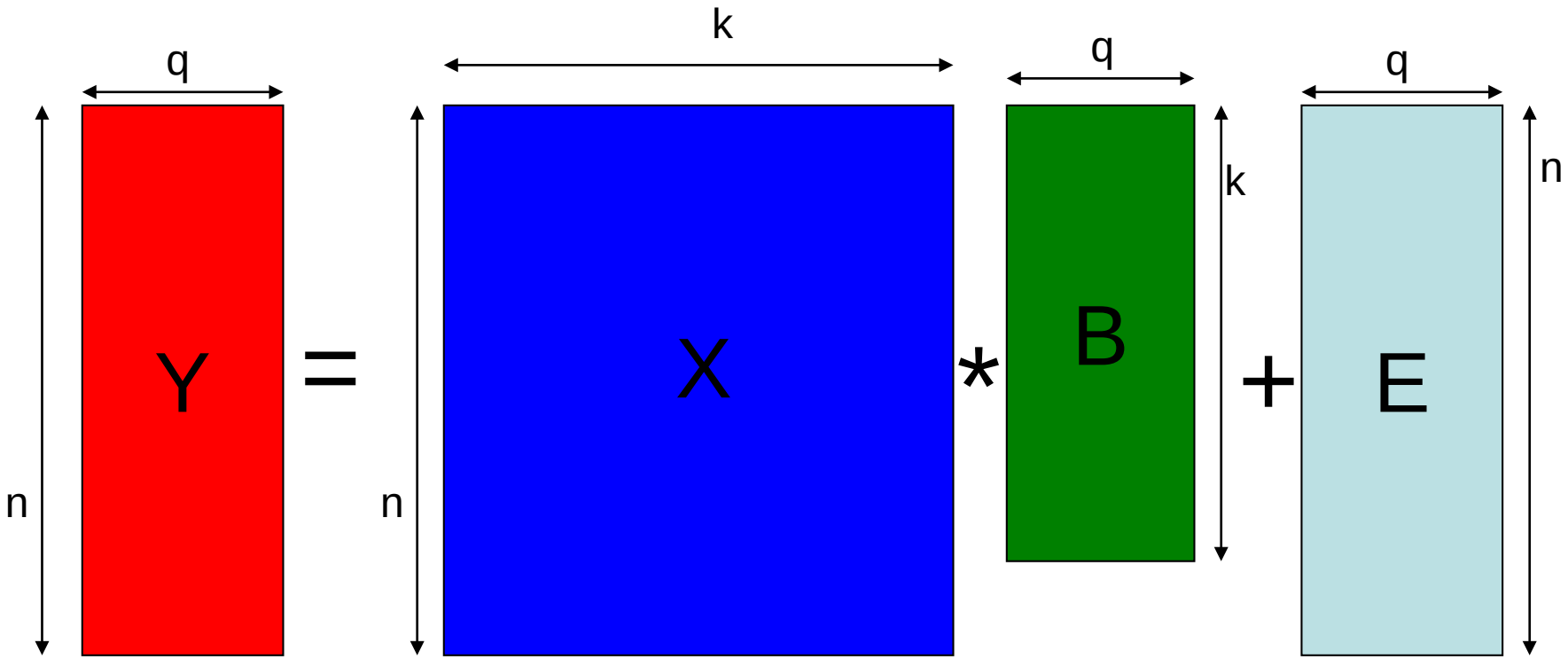
$$a \cdot d - b \cdot c = 0$$

Multiple Linear Regression

- Dati n osservabili ognuno dipendente da m variabili e caratterizzato da errore di misura nel senso esteso del termine, le m variabili possono essere statisticamente stimate utilizzando il metodo dei minimi quadrati.
- Ovviamente dovranno essere rispettate le 4 condizioni del metodo dei minimi quadrati:

1	L'errore su y è molto maggiore dell'errore su x
2	Y è distribuita normalmente
3	Gli eventi osservati sono indipendenti
4	Le misure hanno varianza uguale

Multiple Linear Regression



k = n° osservabili
n =n° misure
q= n° variabili misurabili

Y=XB+E

Multiple Linear Regression

- Come nel caso dei minimi quadrati monovariati, si identificano due fasi:
  - Calibrazione: misurando gli osservabili Y relativi a variabili note X si determina la matrice B
  - Utilizzo: conoscendo la matrice B si ricavano le migliori stime per le quantità X dalle misure degli osservabili Y

- Calibrazione:
  - Noti X e Y la migliore stima per B è data dal teorema di Gauss-Markov:

$$B_{MLR} = X^{+} \cdot Y$$

  - Se X è di rango massimo si può calcolare la pseudoinversa come:

$$B_{MLR} = \left(X^{T} \cdot X\right)^{-1} \cdot X^{T} \cdot Y$$

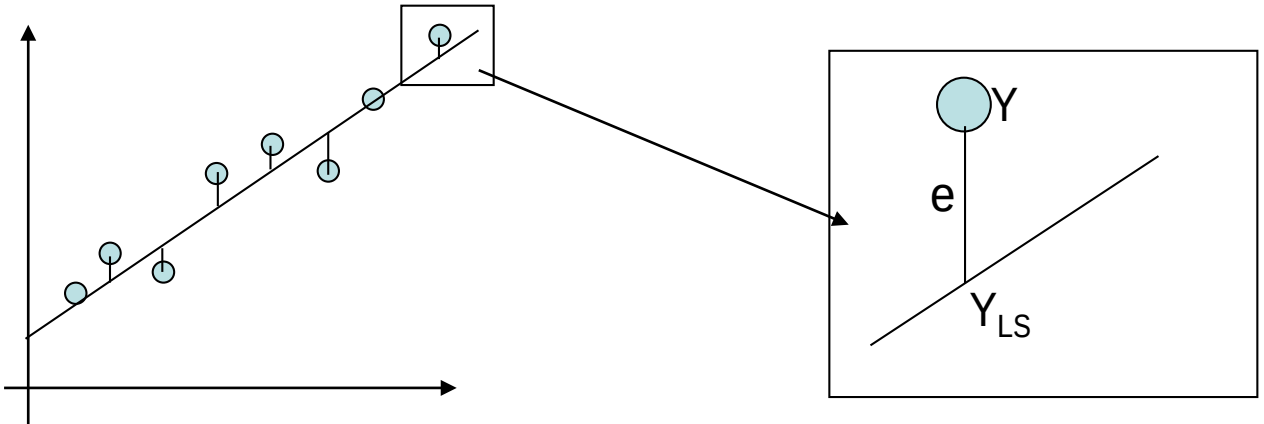
    - **Significa imporre che ogni osservabile sia indipendente dagli altri**

Significato della MLR

- La soluzione del teorema di Gauss-Markov è detta anche estimatore BLUE (best linear unbiased estimator) cioè è lo stimatore di varianza minima
- In pratica B_{MLR} massimizza la correlazione tra X e Y
- Geometricamente la soluzione trovata corrisponde ad una proiezione ortogonale di Y in un sottospazio di X.

$$Y_{MLR} = X \cdot B_{MLR} = X \cdot \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y = \Omega \cdot Y$$

- Ω è una matrice di proiezione ortogonale in un sottospazio di X



Utilizzo pratico

- In pratica conviene imporre la dipendenza lineare tra le grandezze da misurare e gli osservabili ipotizzando un errore distribuito normalmente

$$X = Y \cdot B + e$$

- La soluzione minimi quadrati è quindi data da:

$$X_{MLR} = Y \cdot B_{MLR}$$

- E la matrice B_{MLR} viene stimata dalla seguente:

$$B_{MLR} = Y^{+} \cdot X$$

Limitazioni della MLR

- La Pseudoinversa può essere risolta agevolmente nel caso in cui il rango di X sia massimo e quindi gli osservabili siano indipendenti
  - Questa condizione non è sempre vera:
    - In una riga spettrale, tutte le frequenze della riga sono verosimilmente formate dalle stesse variabili con coefficienti pressochè simili
- Nel caso in cui il rango non sia massimo siamo nella condizione di avere osservabili fortemente correlati, questo grossi errori nel calcolo della pseudoinversa che portano ad errori di stima delle variabili non accettabili se la correlazione è troppo elevata.
- In questi casi bisogna trovare un metodo che riduca la correlazione tra gli osservabili.
- Bisogna in pratica trovare delle nuove variabili dipendenti (funzione degli osservabili) che non siano soggette alla eccessiva correlazione.

Definizione di MLR

- Dati suddivisi in due insiemI: calibrazione e validazione
- Ipotesi: le grandezze **X** sono calcolabili come combinazione lineare dei valori spettrali **Y** più un errore che distribuito normalmente.

$$X = Y \cdot B + e$$

- La stima LS di **X** è quindi:

$$X_{MLR} = Y \cdot B_{MLR}$$

- E la matrice B_{MLR} è data da:

$$B_{MLR} = Y^{+} \cdot X$$

L’Analisi delle Componenti Principali (PCA)

Analisi della Varianza

PCA e diagonalizzazione della matrice di covarianza

Scores e Loadings

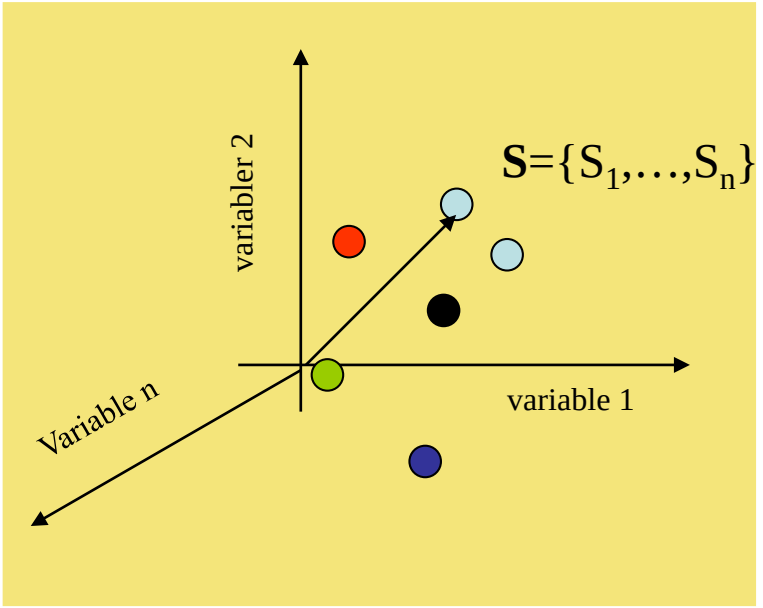
Sviluppo di matrici e residui

Applicazioni all’analisi delle immagini

Applicazione alla regressione multivariata: Principal Components Regression (PCR)

Lo spazio degli osservabili

- Ogni misurazioni multivariata è rappresentata da un vettore in uno spazio a N dimensioni
  - N è pari alla dimensione del vettore che esprime la osservazione
- La distribuzione statistica dei punti (vettori) definisce le proprietà dell'intero set di dati.
- Per ogni grandezza multivariata rappresentabile in uno spazio vettoriale a dimensione N possiamo definre una PDF multivariata.
  - Corollario di grande importanza: osservazioni che descrivono campioni simili sono rappresentate da punti vicini
  - Relazione quindi tra distanza reciproca e similitudine tra campioni (Ipotesi base della *pattern recognition*)

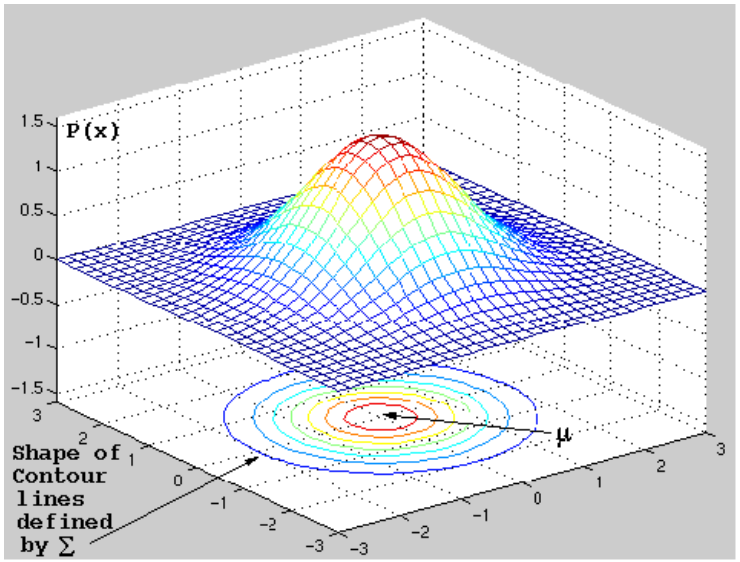


Statistica descrittiva multivariata

- Come per una distribuzione univariata possiamo definire i descrittori fondamentali:
  - Media scalare $\Rightarrow$ vettore
  - Varianza scalare $\Rightarrow$ matrice (matrice di covarianza)
  - ....
- La distribuzione normale definita per una variabile univariata conserva la sua importanza nella statisica multivariata

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$

$$p(x) = \frac{1}{\sqrt{2\Pi}\sqrt{\Sigma}} \exp\left[\frac{-1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu) \right]$$



La matrice di covarianza

- La varianza di una distribuzione univariata definisce la ampiezza della distribuzione stessa, in pratica il range di valori della variabile che hanno una probabilità “reale” di essere osservati
  - In pratica il 99% della probabilità si ottiene in un range ampio 3σ attorno al valore medio.
- Poiché la normale è simmetrica attorno al valore medio i punti di isoprobabilità sono 2 a distanza uguale dalla media
- In una distribuzione multivariata la matrice di covarianza definisce l’ampiezza della PDF e definisce il grado di correlazione tra le variabili stesse
- Il luogo dei punti di isoprobabilità è una ellisse ottenuta come forma quadratica avente come matrice la matrice di covarianza
  - Esponente della PDF multivariata
- La matrice di covarianza può essere stimata dai dati come: $\mathrm{cov}(xy)=x^{T}y$.

Colinearità

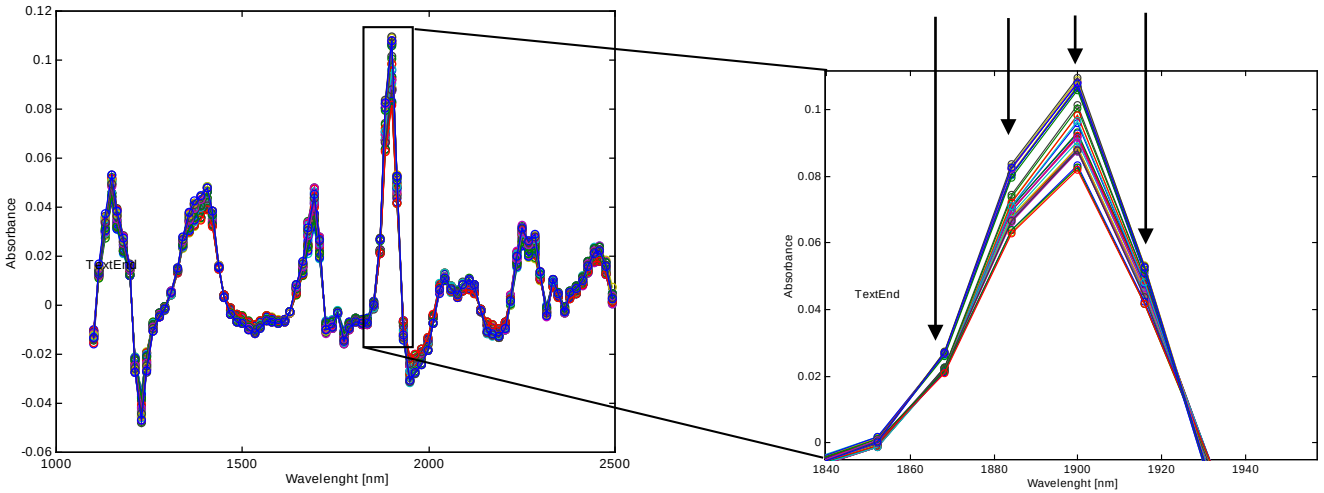
- In un problema MLR la soluzione, stima della variabili **x**, si ottiene invertendo (meglio pseudoinvertendo) la matrice degli osservabili **y**
- Tale operazione è possibile se il rango della matrice **y** è massimo cioè se il numero di colonne linearmente indipendenti coincide con il numero di colonne della matrice.
  - Cioè se tutte gli osservabili sono linearmente indipendenti tra di loro
- Se esiste una parziale dipendenza, cioè se i coefficienti della combinazione lineare sono rigorosamente non nulli, l'inversione numerica della matrice comporta grossi errori di calcolo
- Questo effetto si chiama "colinearità"

Esempio di colinearità

- In uno spettro ottico le righe spettrali coprono un intervallo di lunghezze d'onda, tale intervallo è generalmente coperto da più canali spettrali, di modo che più variabili concorrono a formare una riga spettrale.
- Se la riga è proporzionale ad una caratteristica del campione (es. concentrazione di glucosio) tutti i canali spettrali relativi alla riga saranno in egual modo proporzionali alla caratteristica del campione, quindi le relative variabili (colonne nella matrice dei dati) risulteranno colineari
  - Sono colineari le variabili che dipendono quantitativamente da caratteristiche del campione

Spettro NIR di frutti

Variabili colineari

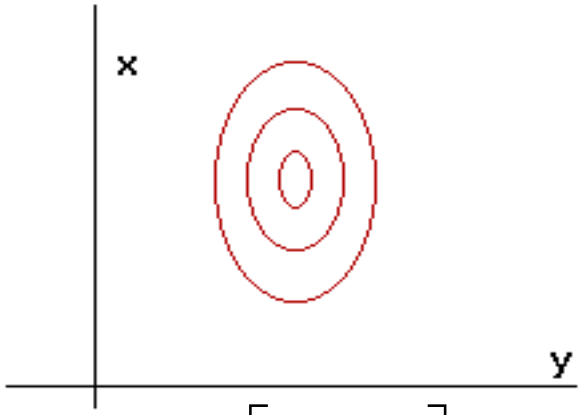


Colinearità e matrice di covarianza

- La colinearità si esprime attraverso la matrice di covarianza.
- In caso di colinearità i termini non diagonali della matrice di covarianza sono diversi da zero.
- Rimuovere la colinearità quindi significa ridurre la matrice di covarianza in forma diagonale introducendo delle nuove variabili latenti.
- La tecnica della analisi delle componenti principali consente, tra le altre cose, di ottenere questo risultato.

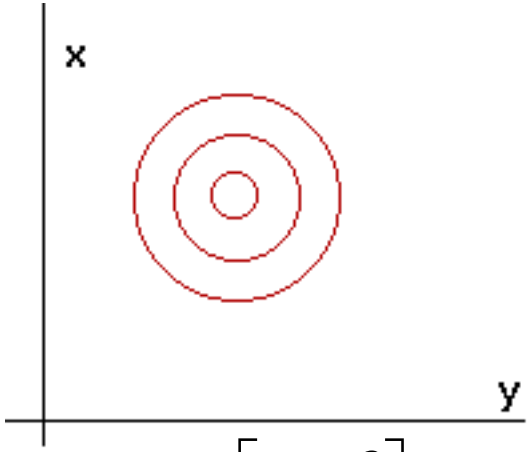
Esempio di matrici di covarianza e luoghi di punti isoprobabili

- Ci sono tre esempi notevoli di matrici di covarianza in termini di correlazione tra le variabili.
  - Come esempio usiamo uina distribuzione bivariata



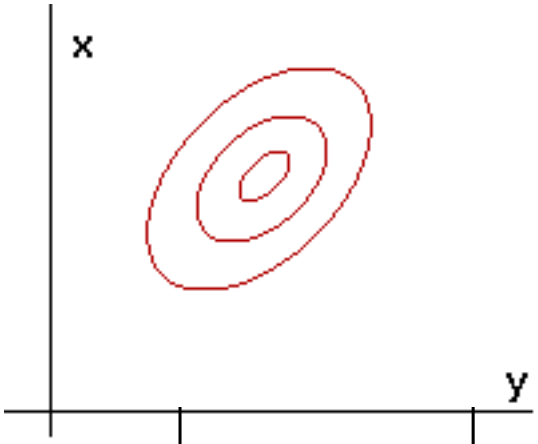
$$\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & 2\sigma \end{bmatrix}$$

Variabili scorrelate
Varianze diverse



$$\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}$$

Variabili scorrelate
Varianze uguale



$$\Sigma = \begin{bmatrix} \sigma_x & \sigma_{xy} \\ \sigma_{xy} & \sigma_y \end{bmatrix}$$

Variabili correlate
Varianze diverse

PDF multivariata e matrice di covarianza

- La normale multivariata ha senso solo se la matrice di covarianza descrive grandezze correlate tra loro, cioè se la matrice è non diagonale.
- Infatti per due grandezze (x e y) tra loro non correlate ed indipendenti la probabilità di osservare contemporaneamente il valore di x e di y è semplicemente il prodotto delle due distribuzioni univariate:

$$P(x,y) = P(x) \cdot P(y)$$

La matrice di covarianza in forma canonica

- La matrice di covarianza può essere scritta in forma diagonale con un adeguato cambiamento del sistema di riferimento.
- Tale sistema di riferimento corrisponde agli autovettori della matrice di covarianza, cioè agli principali dell'ellisse costruita come forma quadratica dalla matrice di covarianza stessa.
- Tale operazione rende le variabili scorrelate e la PDF prodotto di PDF univariate.
- D'altro canto le nuove variabili non sono più degli osservabili fisici (oggetto di misurazioni) ma sono combinazioni lineari di queste.
- Le nuove variabili prendono il nome di Componenti Principali e l'insieme di procedure di calcolo e interpretazione delle componenti principali si chiama analisi delle componenti principali (PCA)

$$a \cdot x^2 + 2b \cdot xy + c \cdot y^2 = \begin{bmatrix} x & y \end{bmatrix} \cdot \begin{bmatrix} a & b \\ b & c \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\Rightarrow \quad \lambda_1 \cdot u^2 + \lambda_2 \cdot w^2 = \begin{bmatrix} u & w \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 \\ o & \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} u \\ w \end{bmatrix}$$

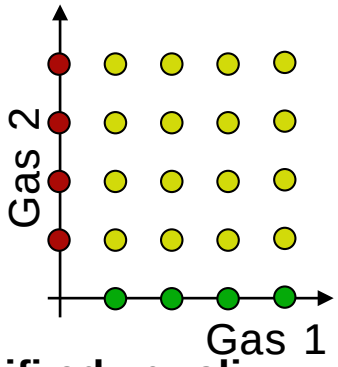
Dimensioni del data set

- Se le variabili di un fenomento multivariato hanno un certo grado di correlazione allora i vettori rappresentativi del fenomeno tenderanno ad occupare solo una porzione dello spazio degli osservabili.
- Quindi una veriabile di dimensione N riempie uno spazio di dimensione minore

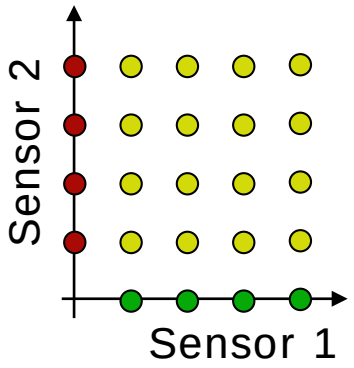
Example: linear sensors

$$\begin{cases} s_1 = k_{11} \cdot g_1 + k_{12} \cdot g_2 \\ s_2 = k_{21} \cdot g_1 + k_{22} \cdot g_2 \end{cases}$$

Spazio delle variabili indipendenti

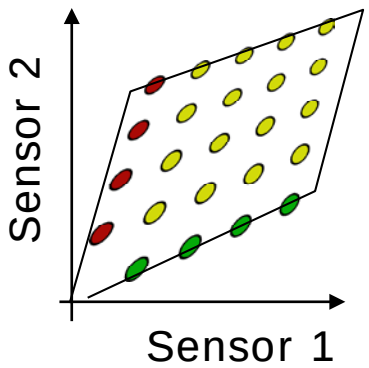


Sensori specifici
$k_{12}=k_{21}=0$



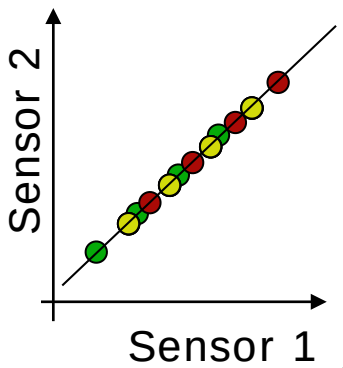
C=0 Dim=2

Sensori non specifici ma diversi
K_{11}; k_{12}; k_{22}; k_{22} diversi



C>0 and <1 Dim intermedia

Non specifi ed uguali

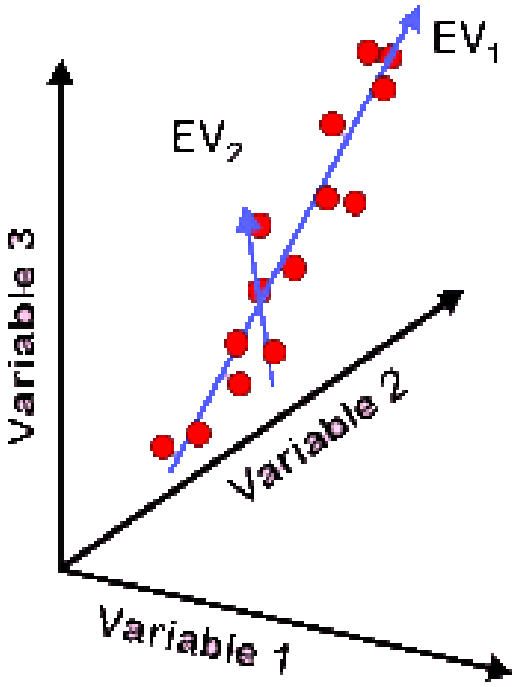


C=1 Dim=1

Principal Component Analysis

- Lo scopo della PCA è la rappresentazione di un insieme di dati con matrice di covarianza non diagonale e di dimensione N in uno spazio di dimensione minore di N in cui gli stessi dati siano rappresentati da una matrice di covarianza diagonale.
- La diagonalizzazione si ottiene con una rotazione delle coordinate nella base degli autovettori (componenti principali)
- Ad ogni autovettore è associato un autovalore a cui corrisponde la varianza della componente principale associata. Se le variabili originarie erano parzialmente correlate tra loro alcuni autovalori avranno un valore trascurabile.
- In pratica gli autovettori corrispondenti possono essere trascurati e limitare la rappresentazione solo agli autovettori con glu autovalori più grandi.
- Poiché la matrice di covarianza nella base delle componenti principali è diagonale la varianza totale è la somma delle varianze delle singole componenti principali.

PCA e proiezione



- La PCA è uno dei possibili modelli che danno luogo alla riduzione delle dimensioni, in pratica si tratta di una proiezione ortogonale dallo spazio originale allo spazio delle componenti principali i cui autovalori associati siano quelli di valore maggiore.

$$\boldsymbol{s} = \boldsymbol{W} \cdot \boldsymbol{x}$$

PCA

- PCA è un metodo detto del secondo ordine poiché sia le nuove coordinate che il criterio per la riduzione delle dimensioni si basano unicamente sulle proprietà della matrice di covarianza
  - La varianza è detta momento secondo dei una distribuzione ed è proporzionale al quadrato della variabile
    - Momento primo: media; secondo: varianza; terzo: skewness;.....
- Quindi la PCA si basa sulla ipotesi che la variabile **x** sia distribuita normalmente
  - La media è in genere resa nulla e quindi tutta l'informazione statistica è contenuta nella matrice di covarianza
- Solo in questo caso le singole componenti principali saranno indipendenti e la probabilità multivariata diventa il prodotto delle probabilità univariate
- Nel caso contrario si ottiene unicamente lo scorrelazione delle componenti principali

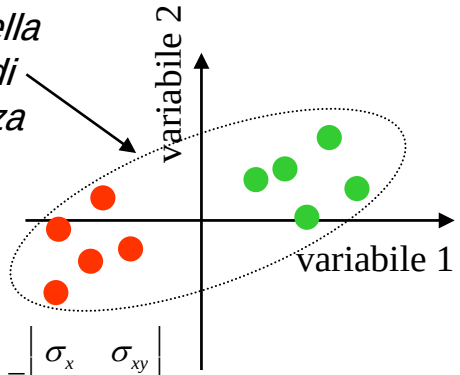
PCA, autovalori ed autovettori

- Le componenti principali della matrice **X** sono gli autovettori (**P**) della matrice $\mathbf{X}^T\mathbf{X}$.
- La matrice **X** risulta decomposta nel prodotto dei **P** (loadings) e delle coordinate dei patterns originali nella base degli **P** (scores): $\mathbf{X}=\mathbf{T}\mathbf{P}^T$
- Le coordinate delle righe della matrice **X** nella base delle componenti principali si calcolano come: $\mathbf{T}=\mathbf{X}\mathbf{P}$
- Gli autovalori sono proporzionali alla varianza dei dati lungo la direzione principale identificata dall'autovettore corrispondente

Interpretazione geometrica della PCA

Spazio osservabili

Ellisse della matrice di covarianza



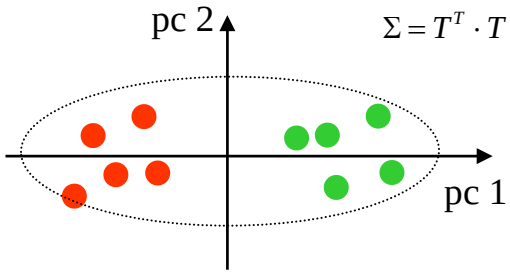
$$\Sigma = X^T \cdot X = \begin{bmatrix} \sigma_x & \sigma_{xy} \\ \sigma_{xy} & \sigma_y \end{bmatrix}$$

PCA

$$\Sigma = X^T \cdot X \Rightarrow \Lambda \cdot P^T$$

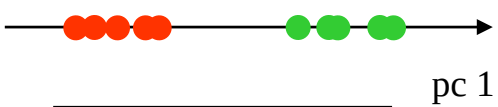
$$T = X \cdot P \;;\; X = T \cdot P^T$$

Spazio comp. Princ.



$$\Sigma = T^T \cdot T = \begin{bmatrix} \lambda_1 & 0 \\ o & \lambda_2 \end{bmatrix}$$

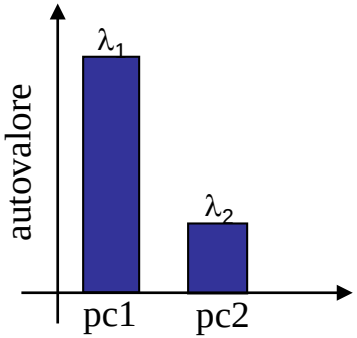
Spazio ridotto



$$T_1 = X \cdot P_1 + E$$

2 ➡ 1

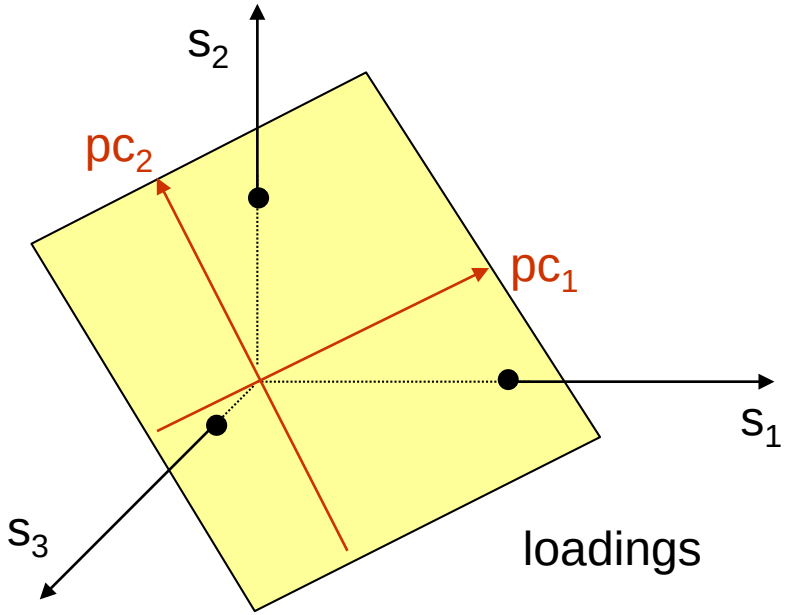
Riduzione delle dimensioni

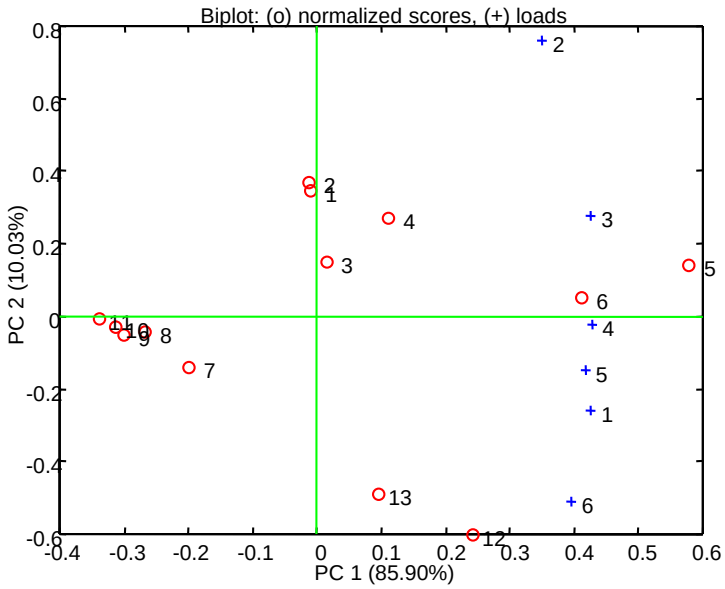


Confronto autovalori: una PC ha un contenuto di informazione maggiore rispetto all'altra

PCA: scores e loadings

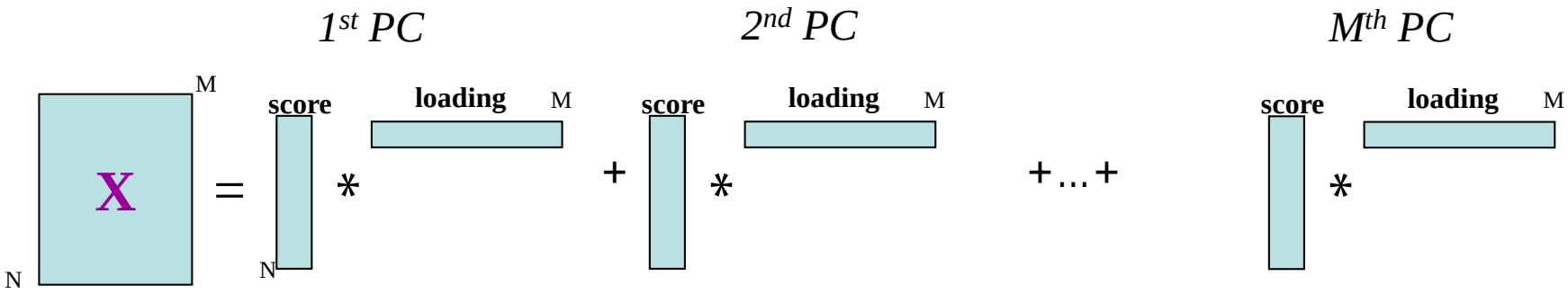
- Le nuove coordinate dei vettori corrispondenti alle osservazioni (le righe della matrice **x**) nella base delle componenti principali prendono il nome di **scores**
- I coefficienti delle combinazioni lineari che definiscono le componenti principali sono detti **loadings**
  - Il loading quindi fornisce una misura del contributo di ogni osservabile alle componenti principali
- I loadings sono anche rappresentabili come scores in quanto sono la proiezione degli assi originali nel sottospazio identificato dalla componenti principali, quindi scores e loadings possono essere graficati insieme





PCA matrix Decomposition

- PCA Può essere considerata come la scomposizione della matrice X nella base delle componenti principali.



- Limitare la scomposizione alla componente p (p<m) significa reiettare una parte dei dati

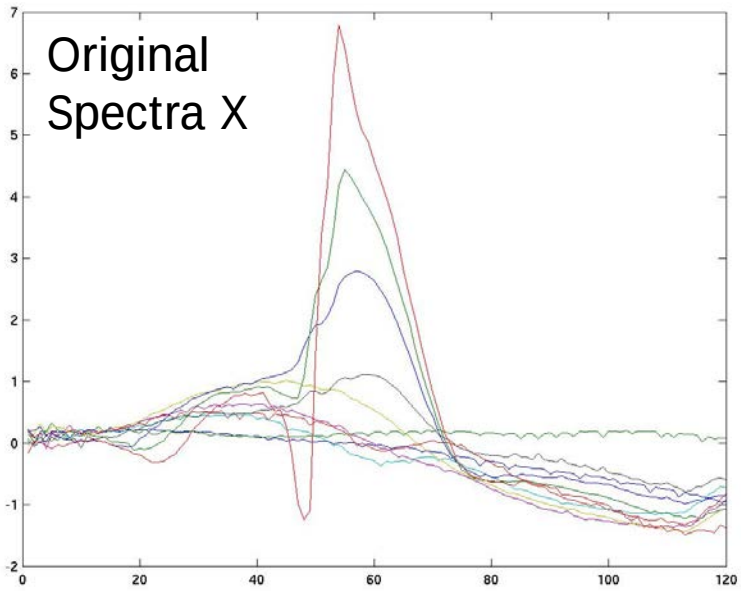
$$X_{nm} = S_{np} \cdot L^{T}_{pm} + Residual$$

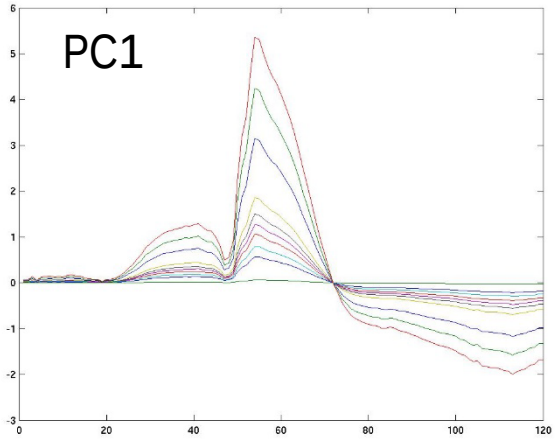
PCA, correlazione e rumore

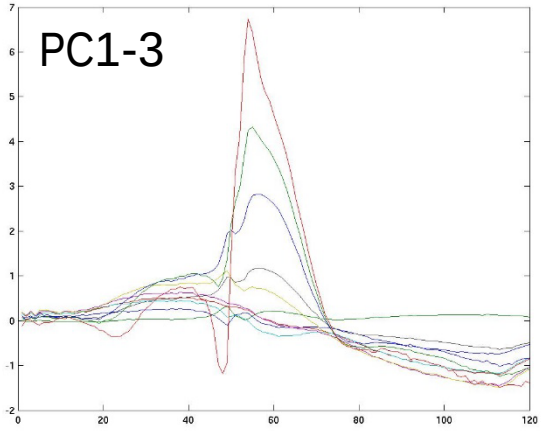
- Il rumore è un termine aggiuntivo stocastico proprio di ogni grandezza osservabile.
- Il rumore è il termine che rende statistica l'operazione di misura.
- Date N variabili il rumore che affetta ognuna di esse è scorrelato rispetto al rumore che affetta le altre.
- Le componenti principali descrivono le direzioni di massima correlazione tra I dati, per cui le PC di ordine più elevato sono orientate verso le direzioni di massima correlazione e quelle di ordine inferiore verso le direzioni di scarsa correlazione
- Limitare la decomposizione alle componenti principali di ordine più elevato significa quindi trattenere le direzioni di massima correlazione e rimuovere quelle non correlate, nella parte non correlata c'è sicuramente il rumore
- La PCA quindi è un metodo per ridurre la quantità di rumore in un set di dati multivariati.
  - esempio: spettroscopia, GC,...

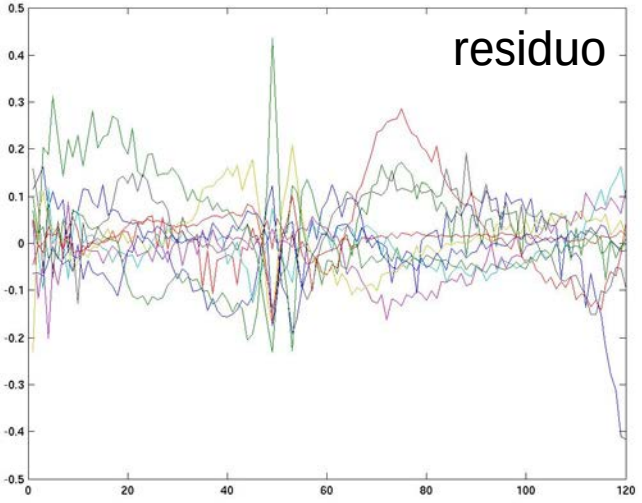
Esempio di rimozione del rumore:

Reflectance Anisotropy Spectroscopy di superfici organiche ordinate



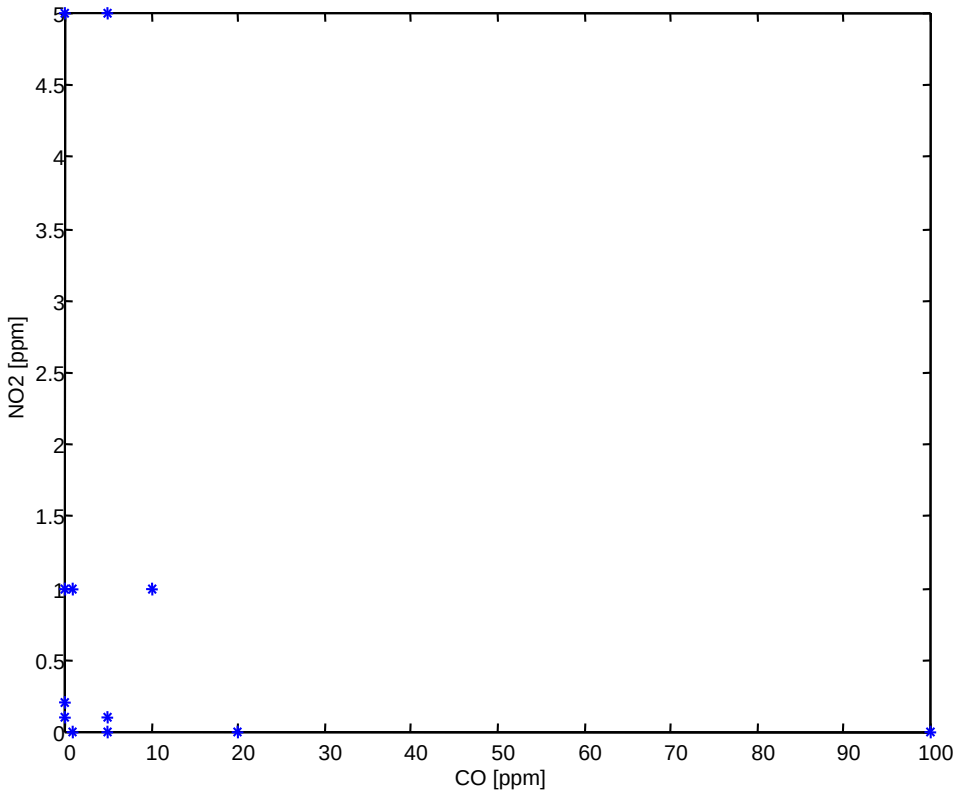




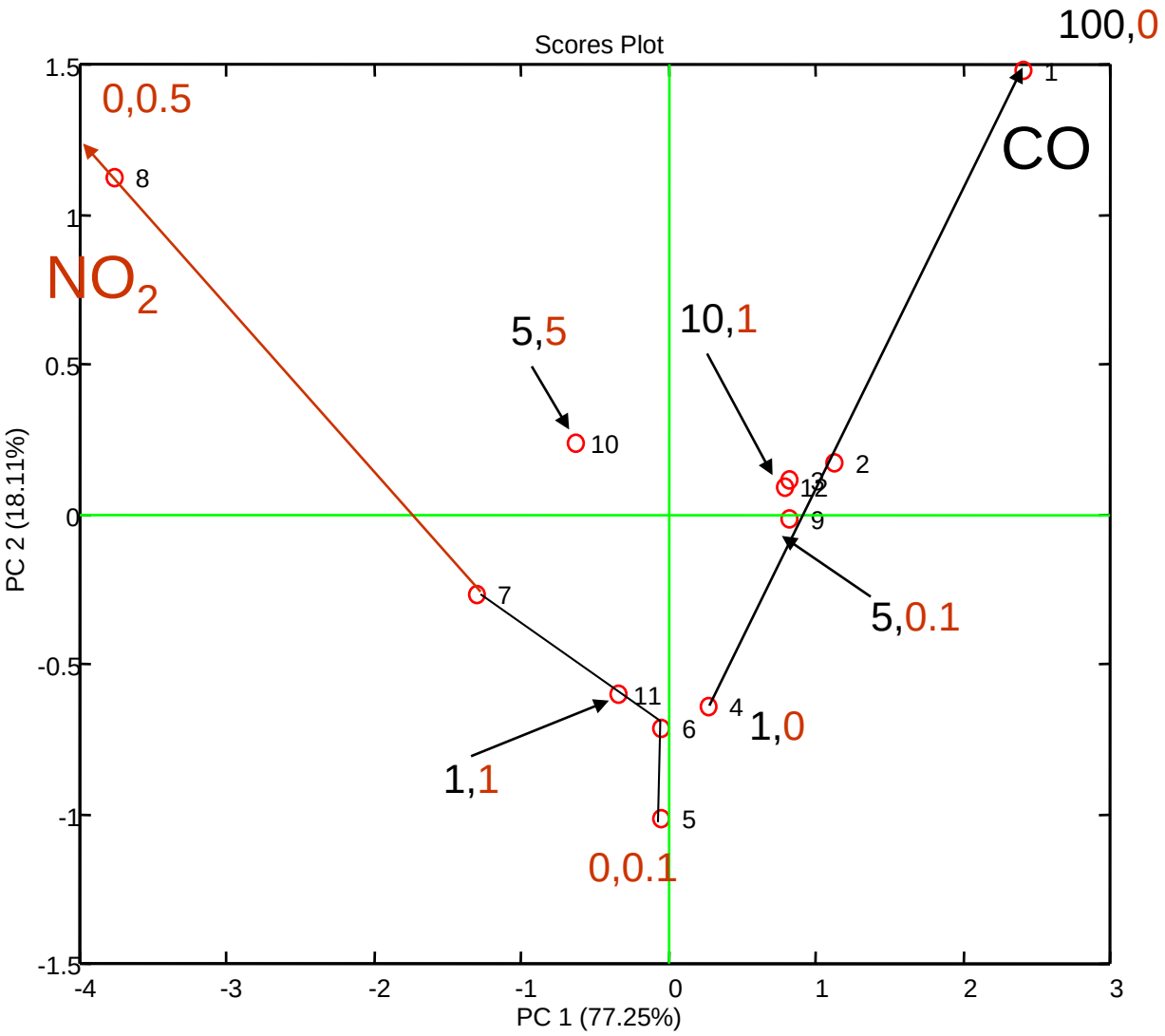


Esempio: 3 SnO2 sensori for 2 gas

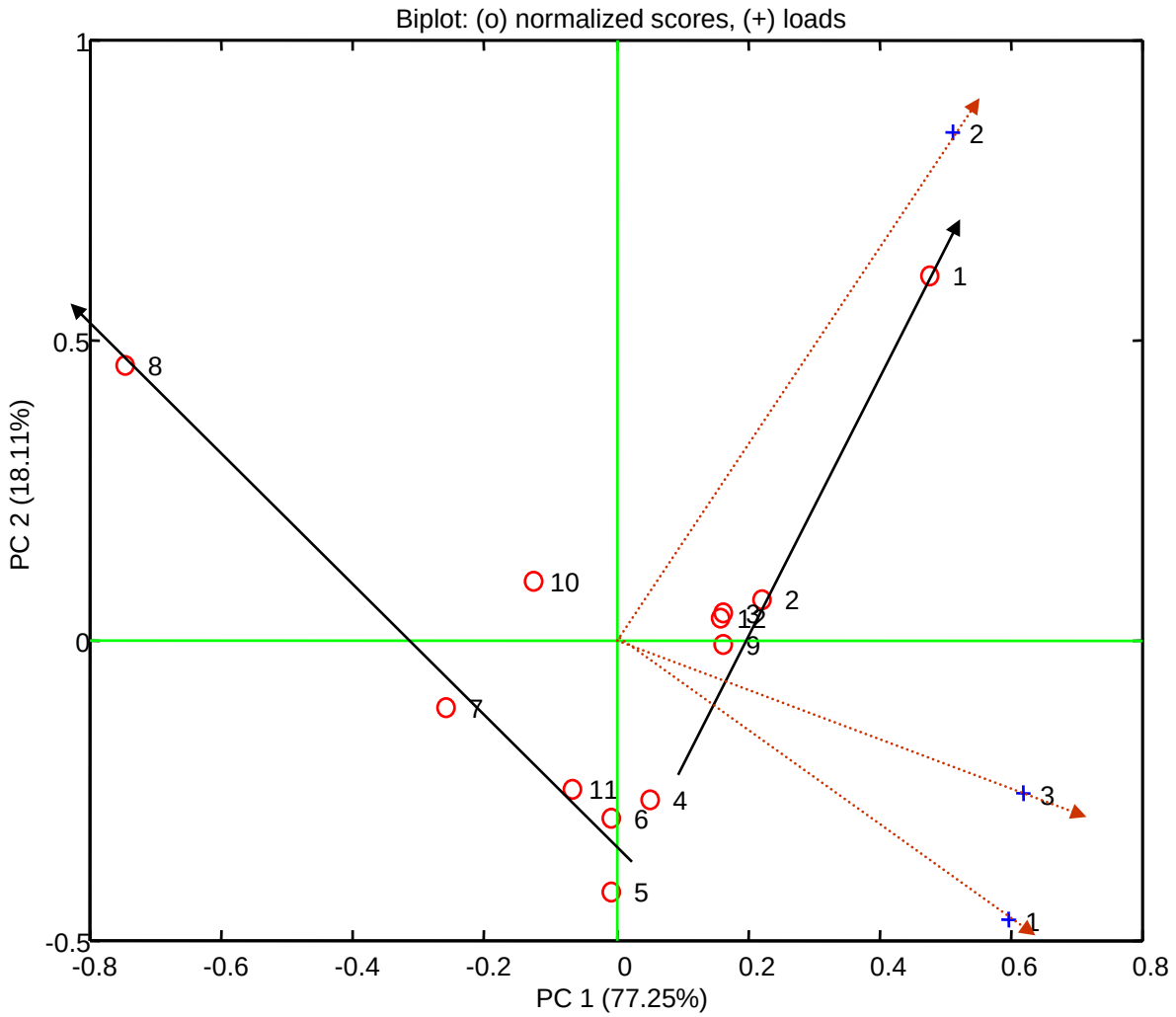
G_f/G_i			CO	NO_2
0.25482	0.63354	0.77832	100.00	0.0000
0.093899	0.27108	0.39692	20.000	0.0000
0.043410	0.23361	0.079543	5.0000	0.0000
0.0097185	0.043353	-0.0021311	1.0000	0.0000
-0.018016	-0.053860	-0.073648	0.0000	0.10000
-0.028579	0.0023183	-0.36593	0.0000	0.20000
-0.25167	-0.028831	-2.4367	0.0000	1.0000
-1.6960	-0.075037	-3.8650	0.0000	5.0000
0.057521	0.21072	0.16777	5.0000	0.10000
-0.13089	0.13002	-2.1376	5.0000	5.0000
-0.068079	-0.0027190	-0.90852	1.0000	1.0000
0.050023	0.22771	0.020198	10.000	1.0000



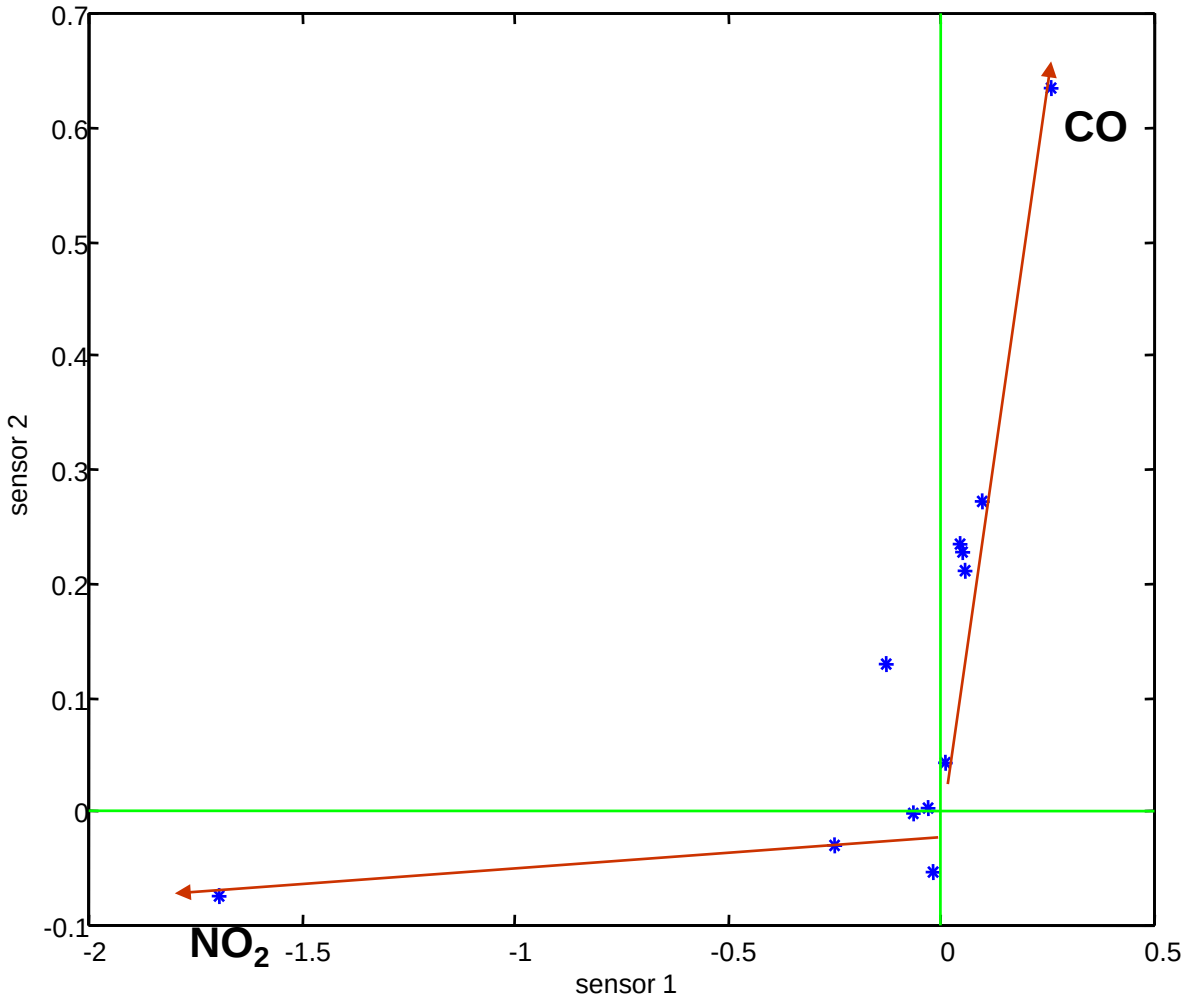
PCA score plot



PCA bi-plot


Biplot: (o) normalized scores, (+) loads
PC 2 (18.11%)
PC 1 (77.25%)
1
0.5
0
-0.5
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
2
1
8
10
2
12
9
7
11
4
6
3
5
1

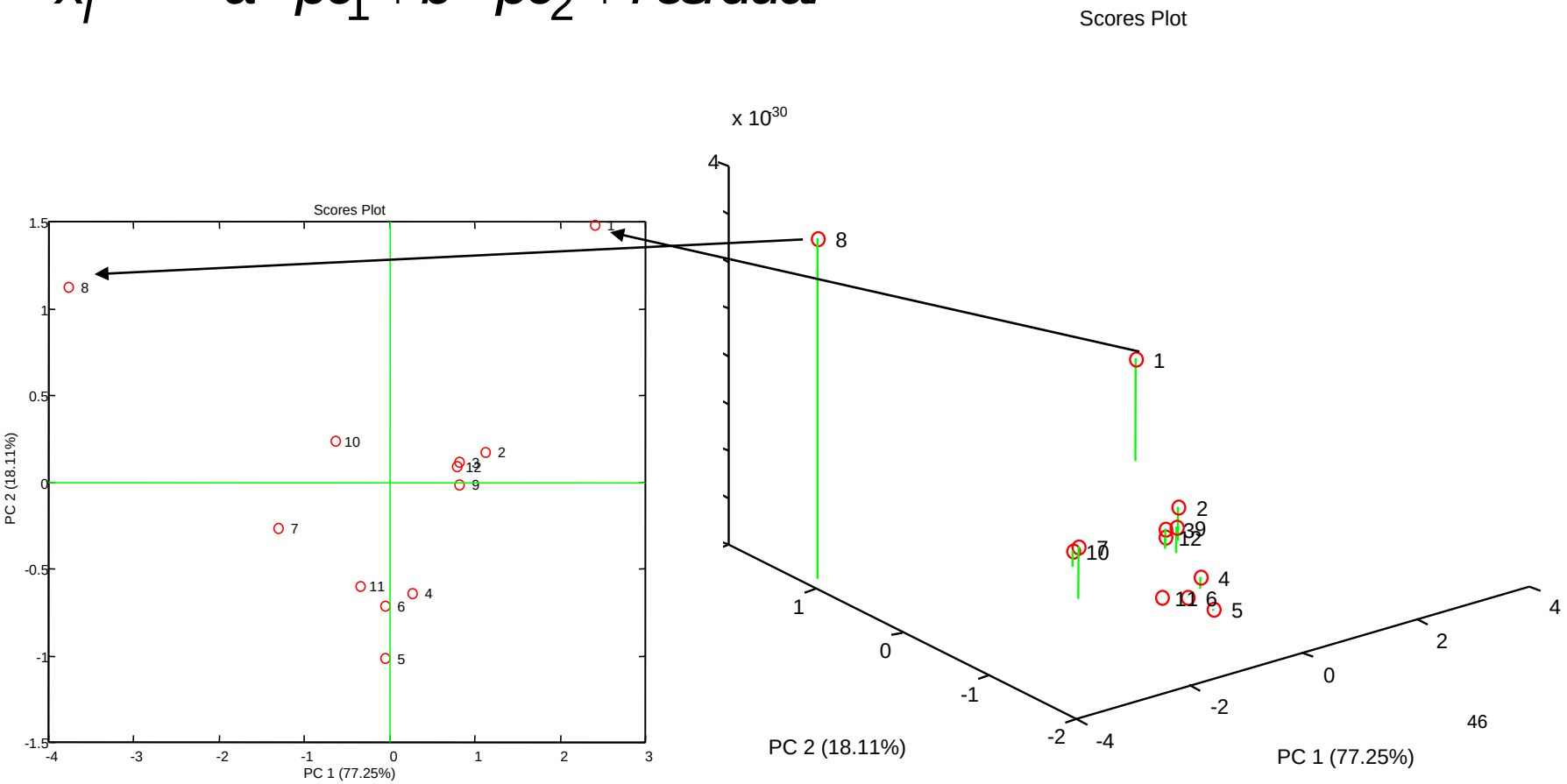
Sensor 1 vs sensor 2



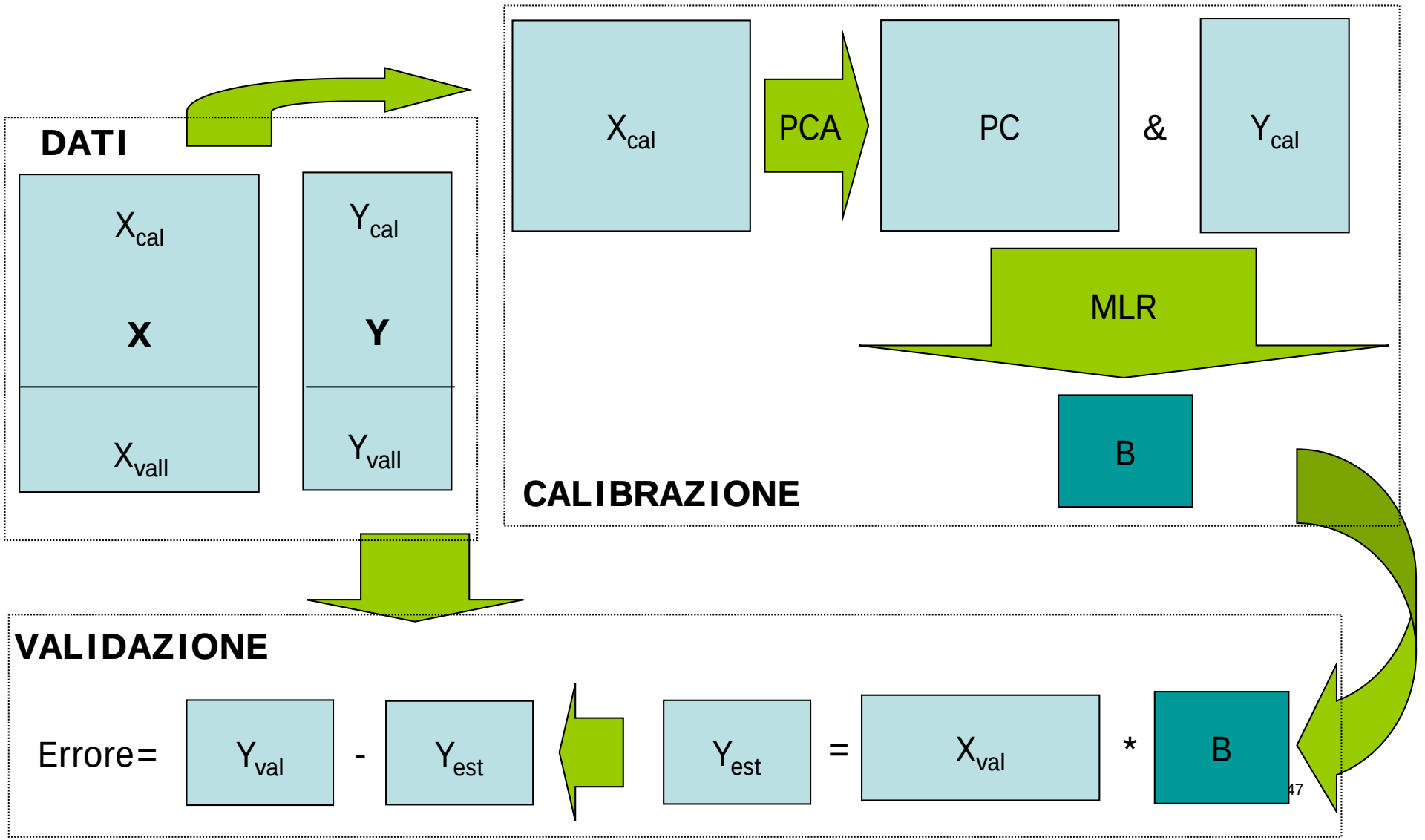
Residual of PCA representation
(leverage)

$$x_i = a \cdot s_1 + b \cdot s_2 + \ldots + n \cdot s_n$$

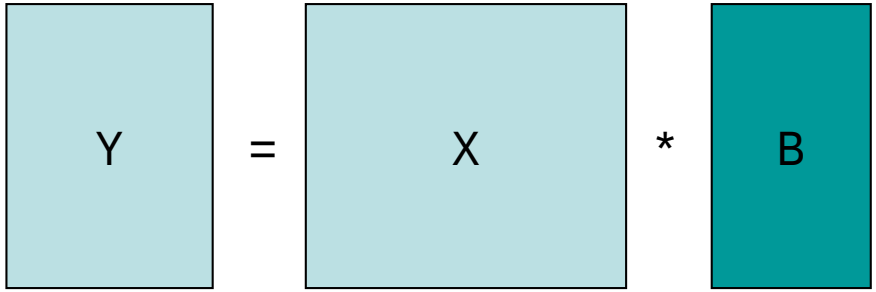
$$x_i^{pca} = a \cdot pc_1 + b \cdot pc_2 + residual$$



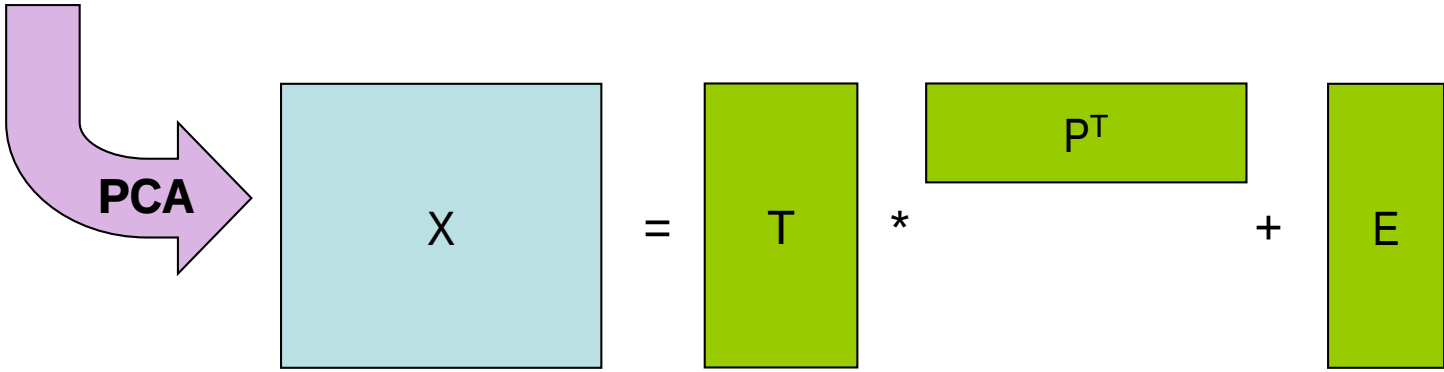
Procedura PCR

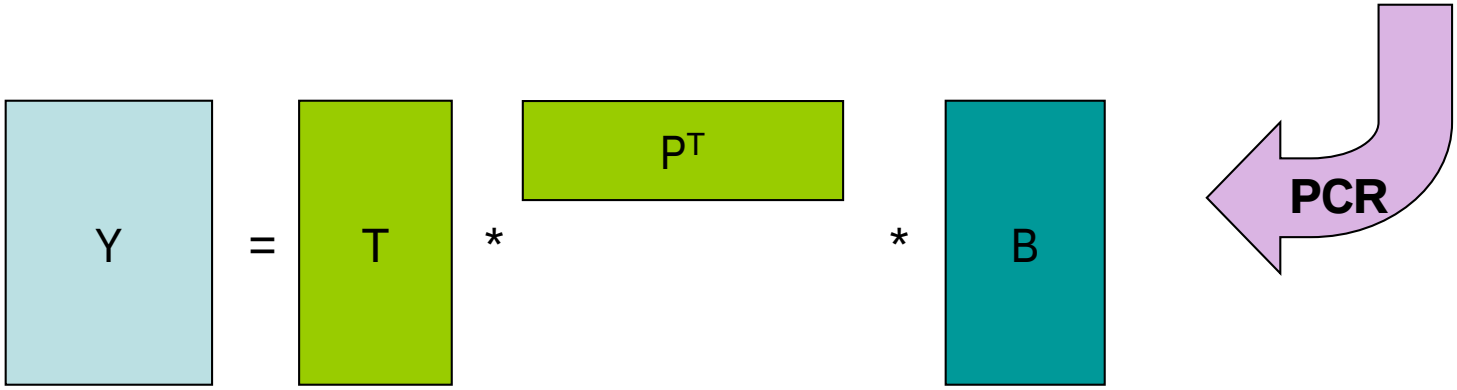


Algoritmo PCR



Original problem





Algoritmo PCR

$$X^T \cdot X \Rightarrow \Lambda \cdot P^T + E$$

$$\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_N \end{pmatrix}$$

$$T = X \cdot P \qquad X = T \cdot P^T$$

$$Y = X \cdot B^T = T \cdot Q^T = T \cdot P^T \cdot B^T \Rightarrow B^T = P \cdot Q^T$$

$$Y = X \cdot \left(P \cdot P^T \right) \cdot B^T$$

$$B^T = \left(X^T X \right)^{-1} \cdot X^T \cdot B^T = \left(P \cdot \Lambda^{-1} \cdot P^T \right) \cdot X^T \cdot Y$$

$$B^T = \left(P \cdot \Lambda^{-1} \cdot P^T \right) \cdot P \cdot T^T \cdot Y = P \cdot \Lambda^{-1} \cdot T^T \cdot Y$$

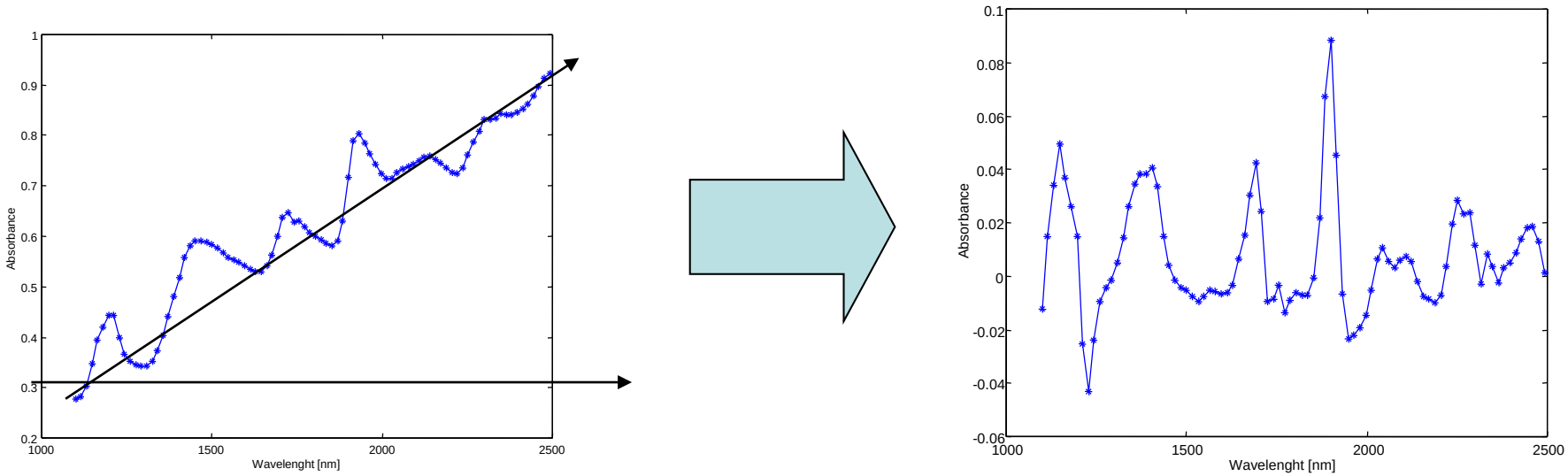
Esempio: Spettri NIR di frutta secca

- Supponiamo di aver raccolto 36 spettri NIR di frutta secca e di voler realizzare un modello per la misura del contenuto di umidità e di acidità totale.
- Ogni spettro è formato da 88 variabili corrispondenti ai canali spettrali nell'intervallo 1.1-2.5 μm.
- Per ogni specie di frutta sono stati misurati umidità ed acidità con metodi di riferimento.
- Vogliamo quindi realizzare il seguente modello che dallo spettro X ci consente di ricavare i due parametri Y. E' necessario quindi stimare il parametro K

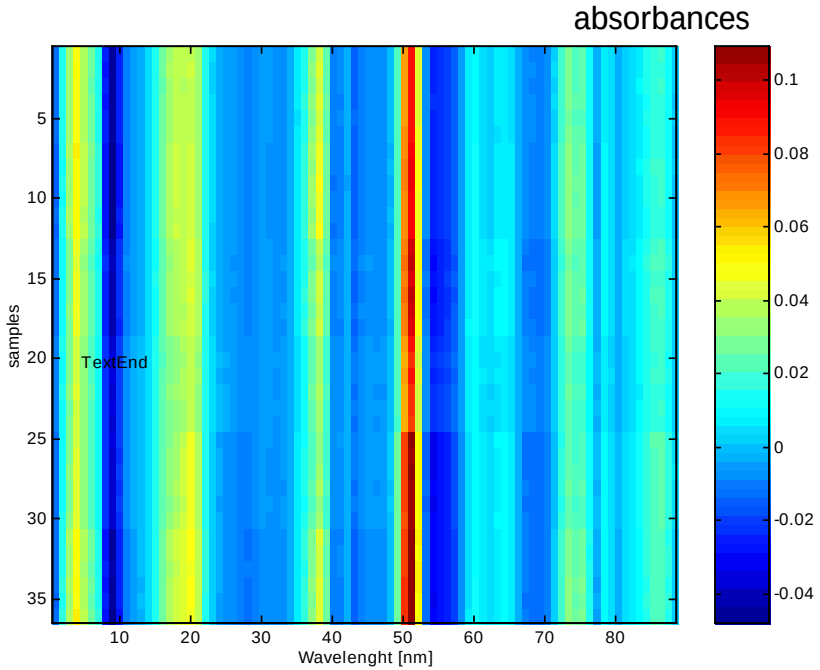
$$Y_{1x2} = X_{1x88} \cdot K_{88x2}$$

Inciso sugli spettri

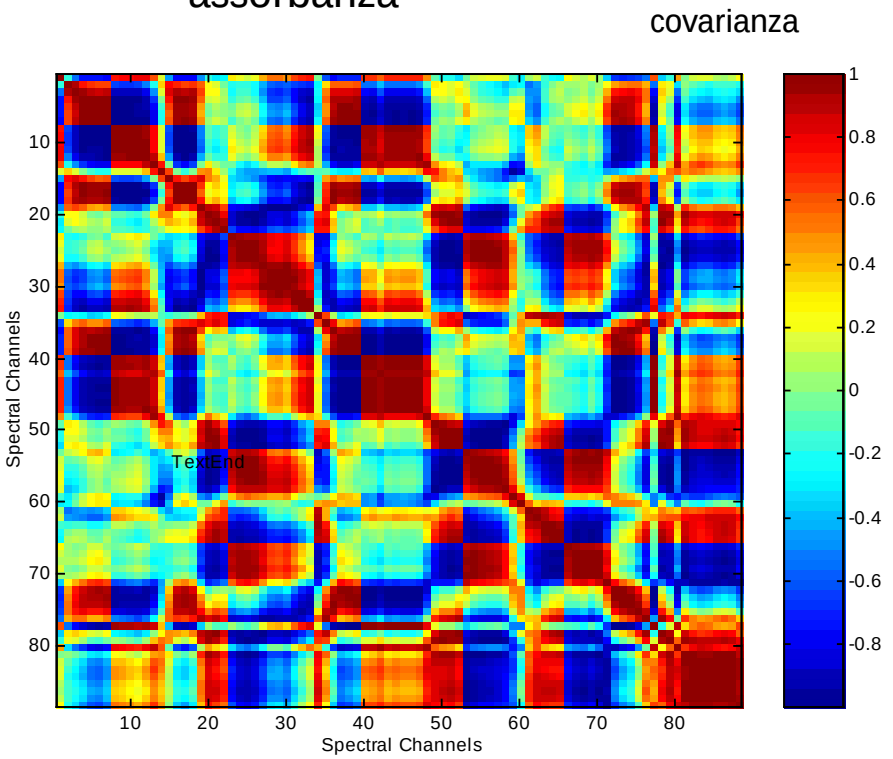
- Spesso gli spettri ottici, in particolare quelli NIR, sono affetti da drift in lunghezza d'onda detto "baseline drift"
- La baseline può essere eliminato derivando numericamente lo spettro
- Un altro metodo, sempre basato sulla derivata. è la normalizzazione di *Savitzky-Golay*



Matrice **X** e matrice di covarianza

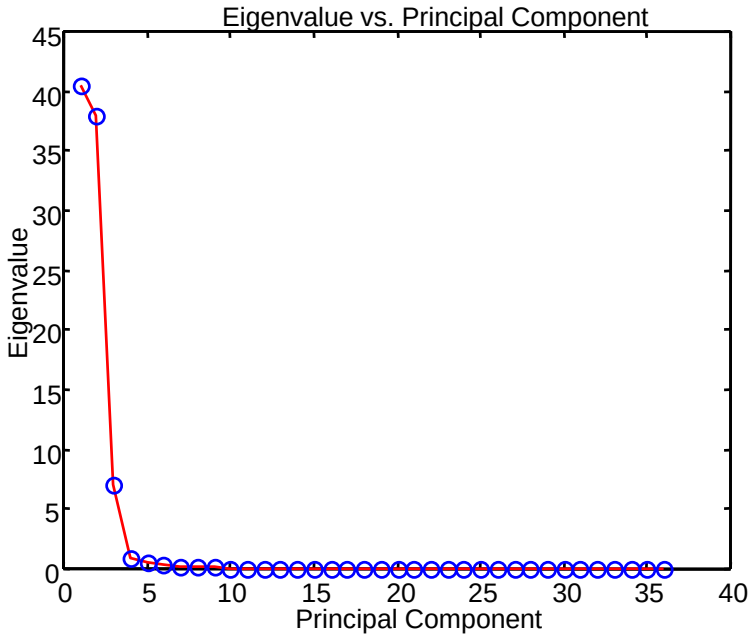


- Colinearità elevata
- I blocchi di elevata correlazione (+ e -) corrispondono alle righe spettrali
  - Colonne colorate nella matrice di assorbanza



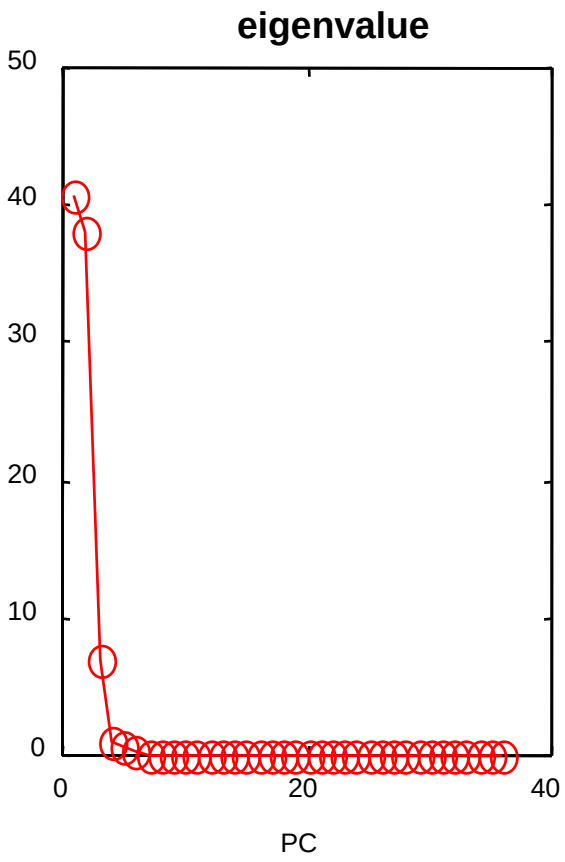
Calcolo della PCA

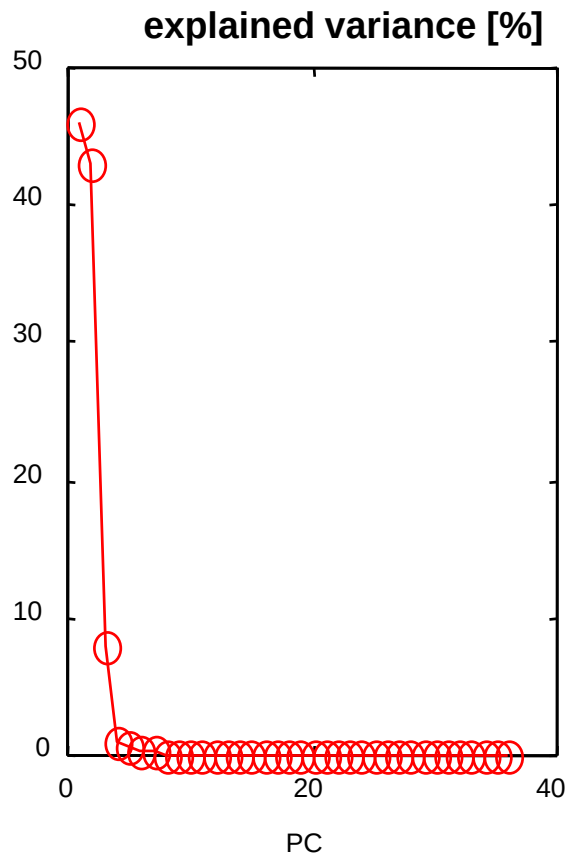
- Riduciamo gli spettri a media nulla in modo che se l'ipotesi di distribuzione normale è soddisfatta, tutta l'informazione è contenuta nella matrice di covarianza.
- Calcolo di autovettori ed autovalori

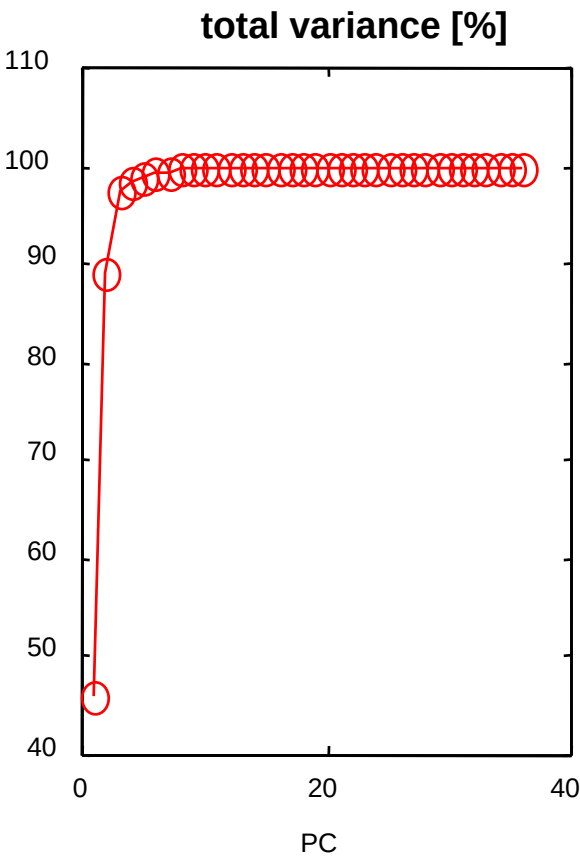


- I primi 3 autovalori hanno un valore considerevolmente diverso da zero.
- Gli 88 spettri, vettori in uno spazio a dimensione 88, sono in buona parte confinati in un sottospazio a dimensione 3.

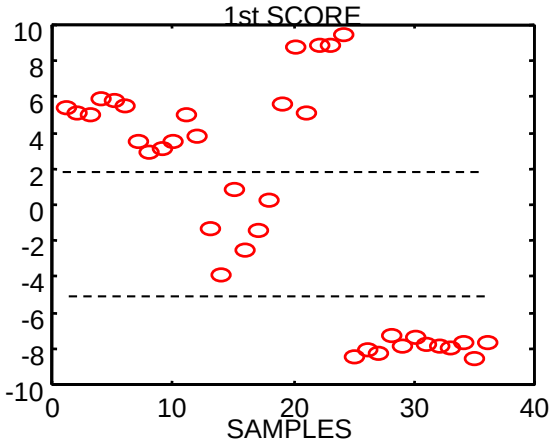
Autovalori e varianza

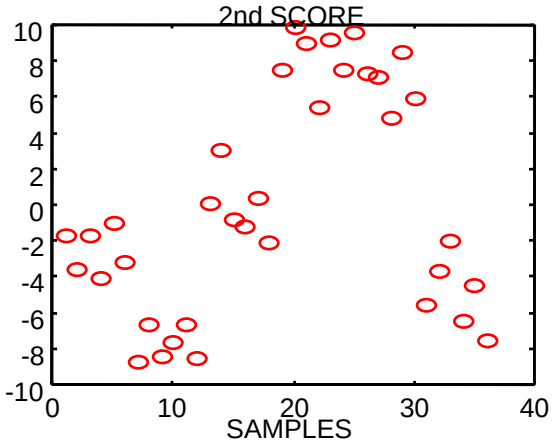


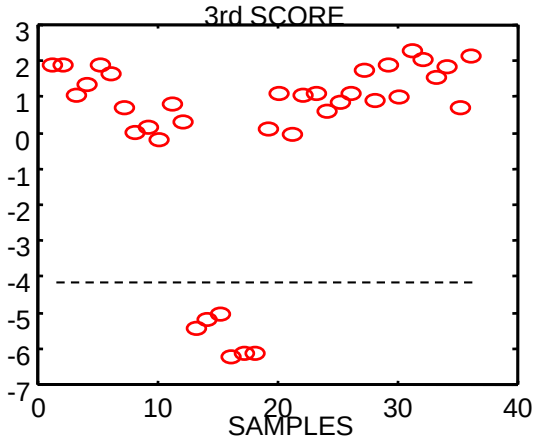


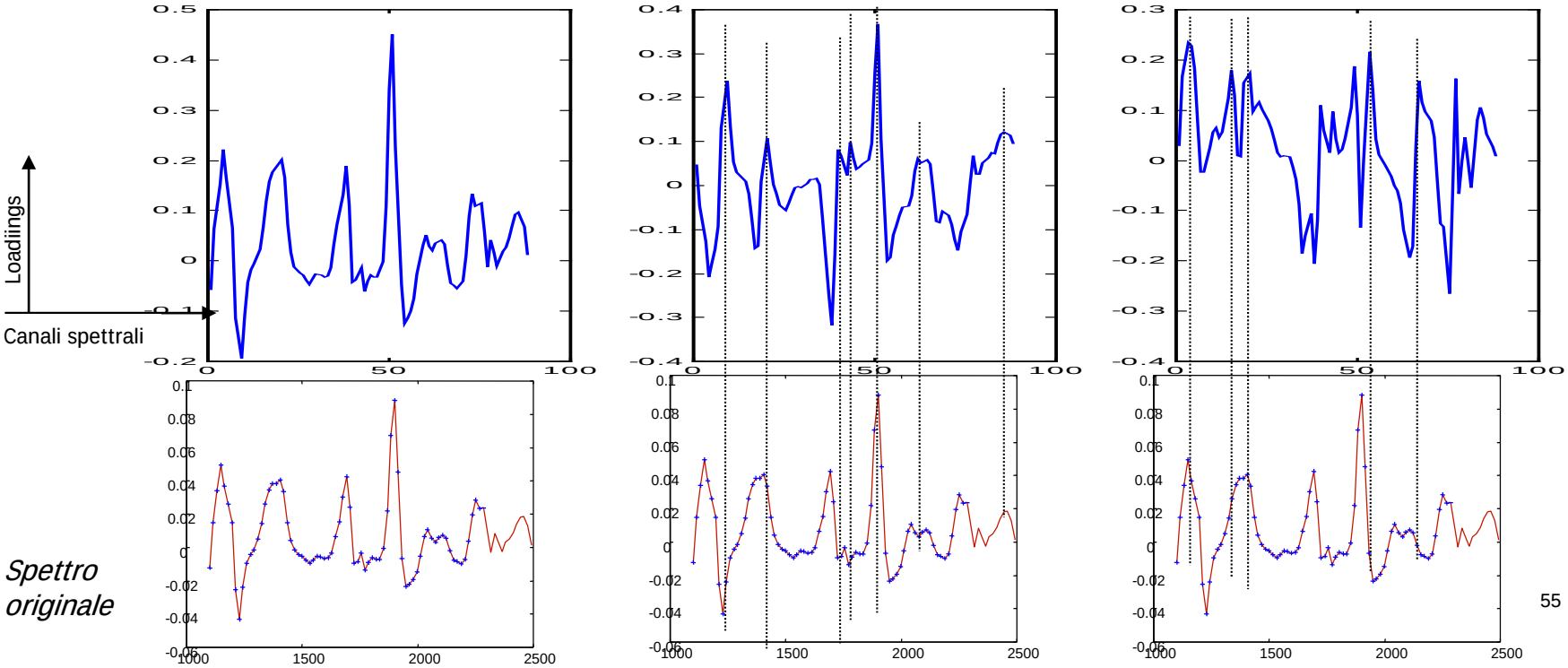


Scores e loadings

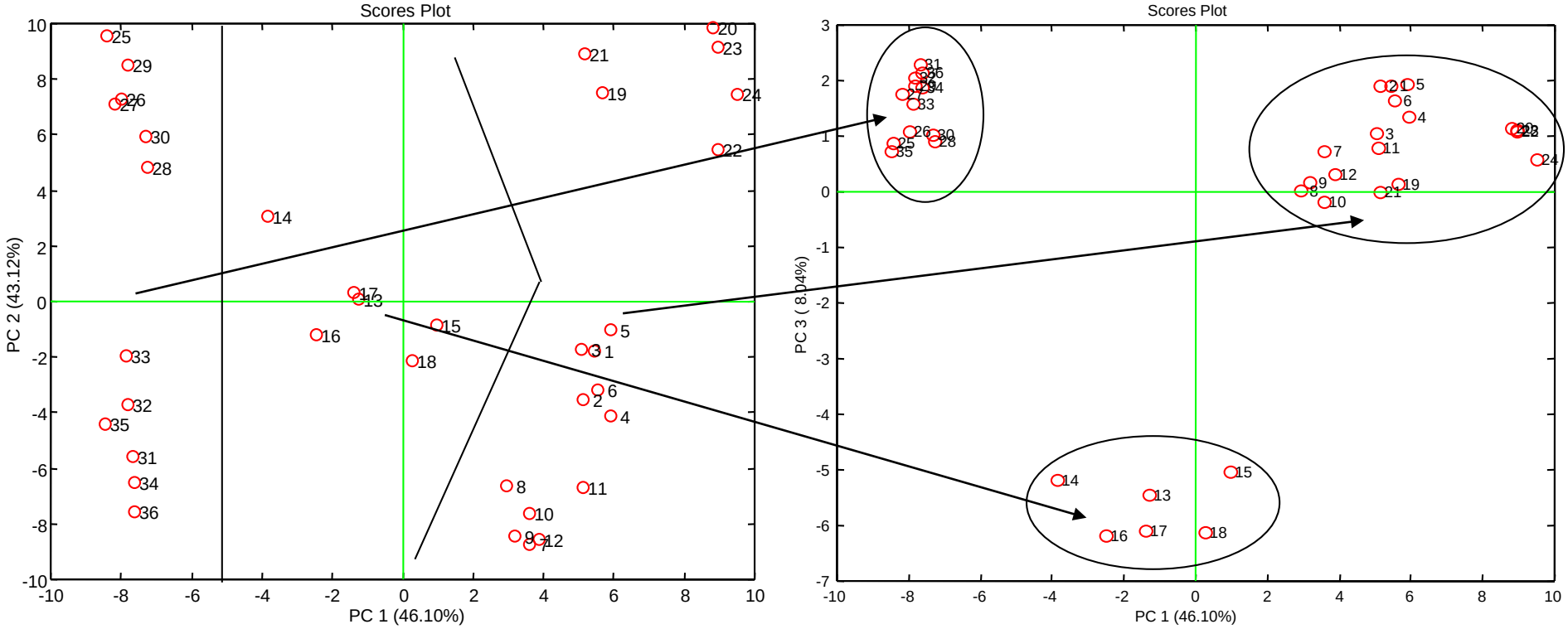






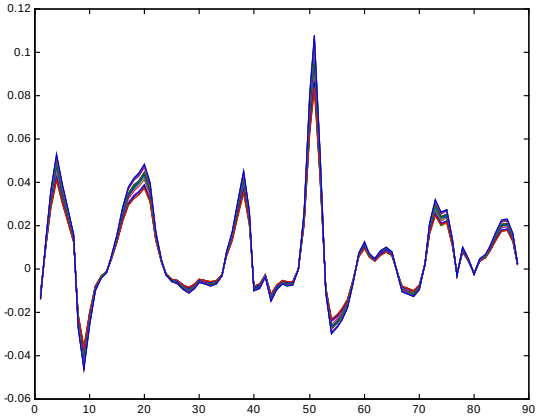


Scores plot

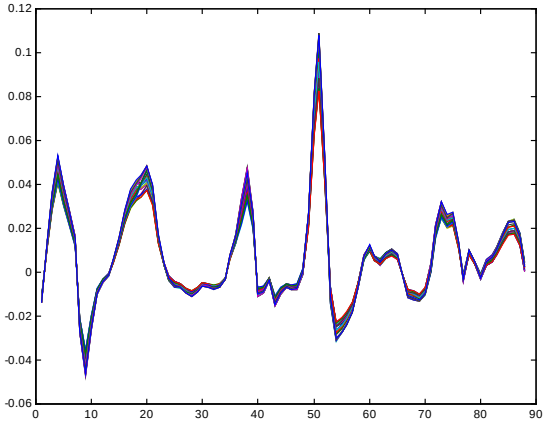


Decomposizione e residui

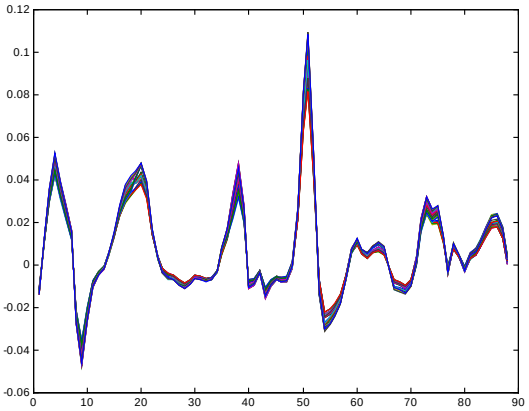
Prima PC



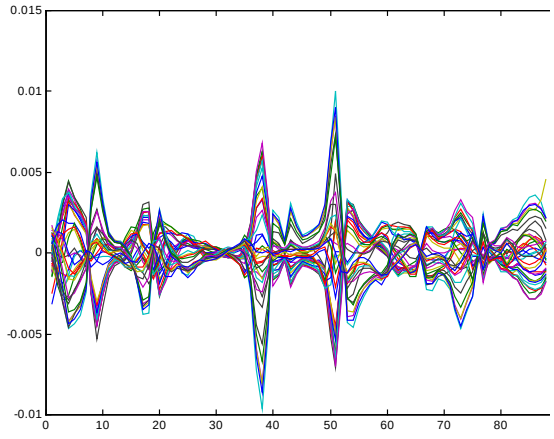
Seconda PC

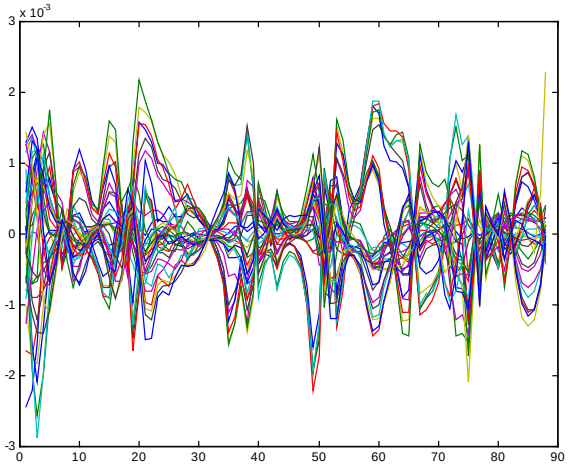


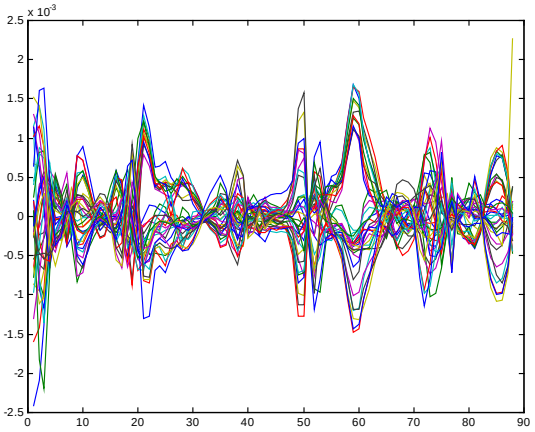
Terza PC



Residui







Calcolo della Principal Components Regression (PCR)

- Separiamo il set di dati in due:
  - 26 per il calcolo del modello PC_{cal}, Y_{cal}
  - 10 per la valutazione dell'errore PC_{val}, Y_{val}
- Dal modello si calcola la matrice di regressione B_{pcr}

$$Y_{cal} = X_{cal} \cdot B^T \Rightarrow B^T = P \cdot \Lambda^{-1} \cdot T^T \cdot Y_{cal}$$

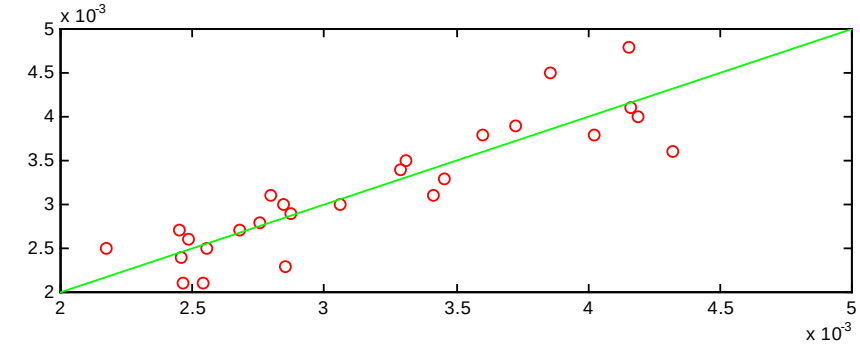
- Si calcola poi la stima sul set di validazione (e per confronto anche su quello di calibrazione)
  - Si valuta RMSEC ed $RMSECV_r$

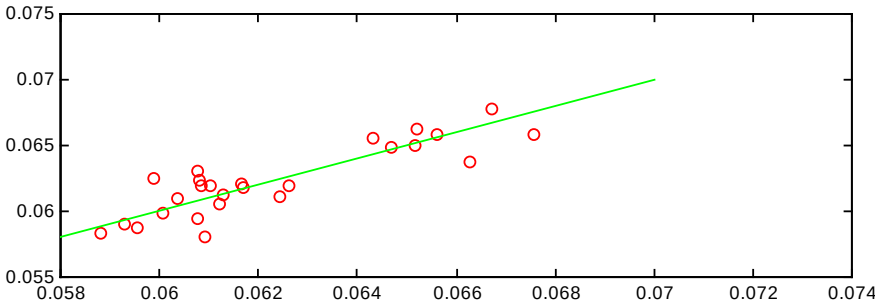
$$stimaY_{cal} = X_{cal} \cdot B^T$$

$$stimaY_{val} = X_{val} \cdot B^T$$

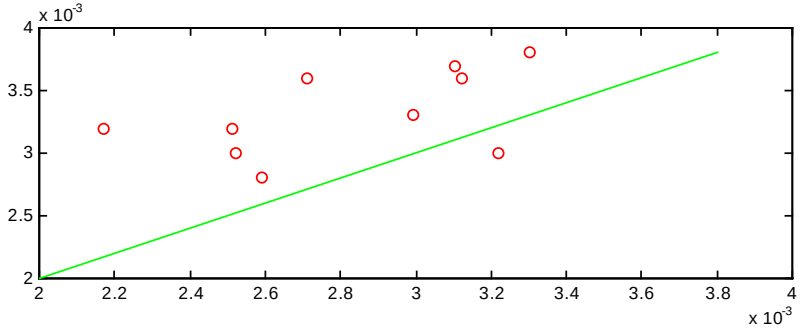
Esempio risultati

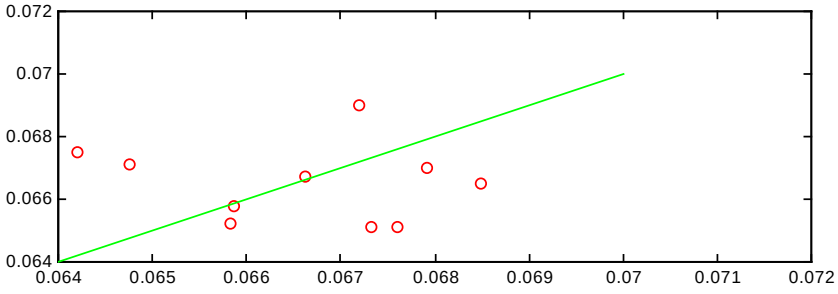
calibrazione





test





$RMSEC_{acidità}=3.1\ 10^{-4}$
$RMSEC_{umidità}=0.0013$

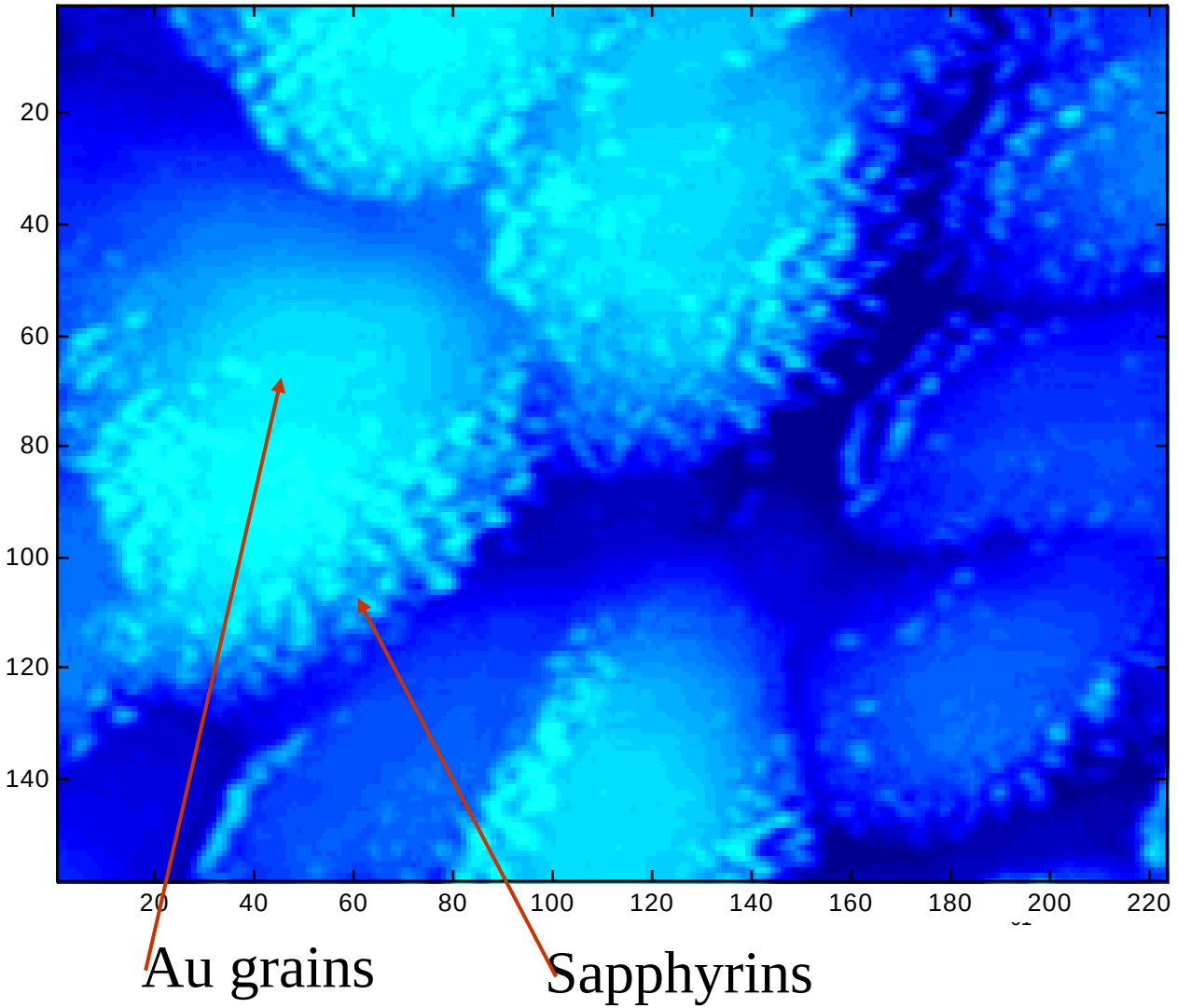
$RMSECV_{acidità}=5.9\ 10^{-4}$
$RMSECV_{umidità}=0.0019$

Applicazione alla analisi delle immagini

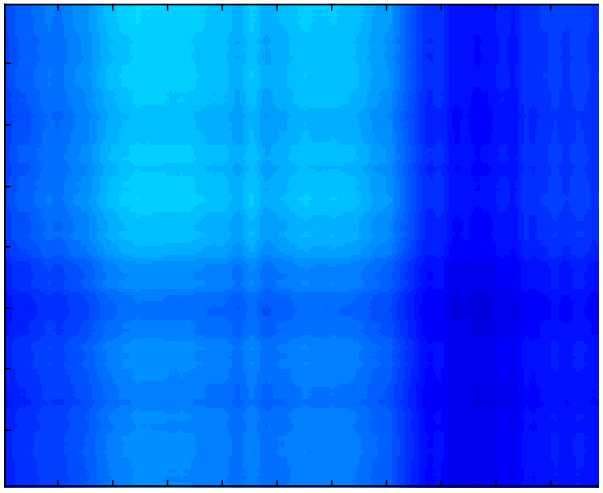
- Un immagine digitalizzata può essere considerata come una matrice NxM nel caso di immagine a scala di grigio (bianco nero) o NxMx3 (nel caso di immagine a colori)
- Considerando una immagine in una scala di tonalità la possiamo considerare come una matrice ed applicare la PCA
  - Più avanti considereremo le strutture 3dimensionali di dati.
- La decomposizione PCA può mettere in evidenza alcune strutture peculiari dell'immagine permettendo quindi di studiare le caratteristiche dell'immagine stessa.

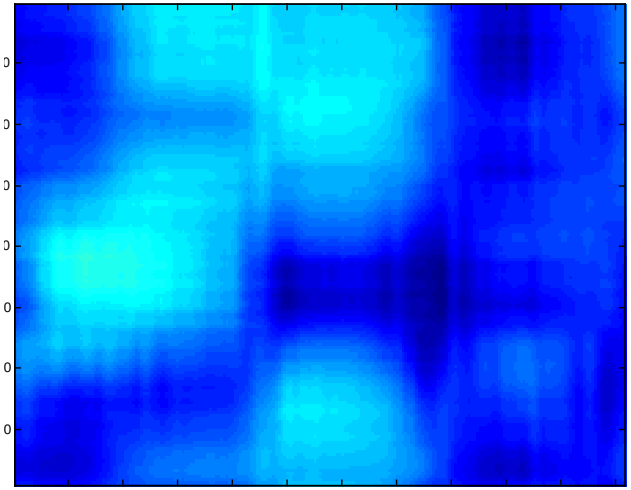
PCA: Application to Image Analysis (example 1: I)

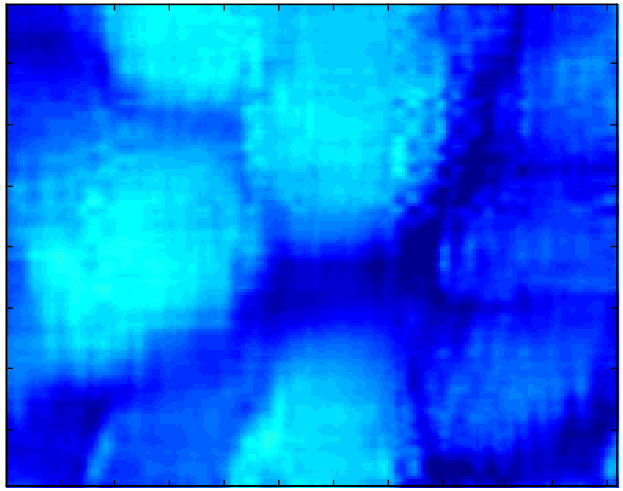
- STM image of Sapphyrin molecules growth as a Langmuir-Blodgett film onto a gold substrate.



PCA: Application to Image Analysis (example 1: II)





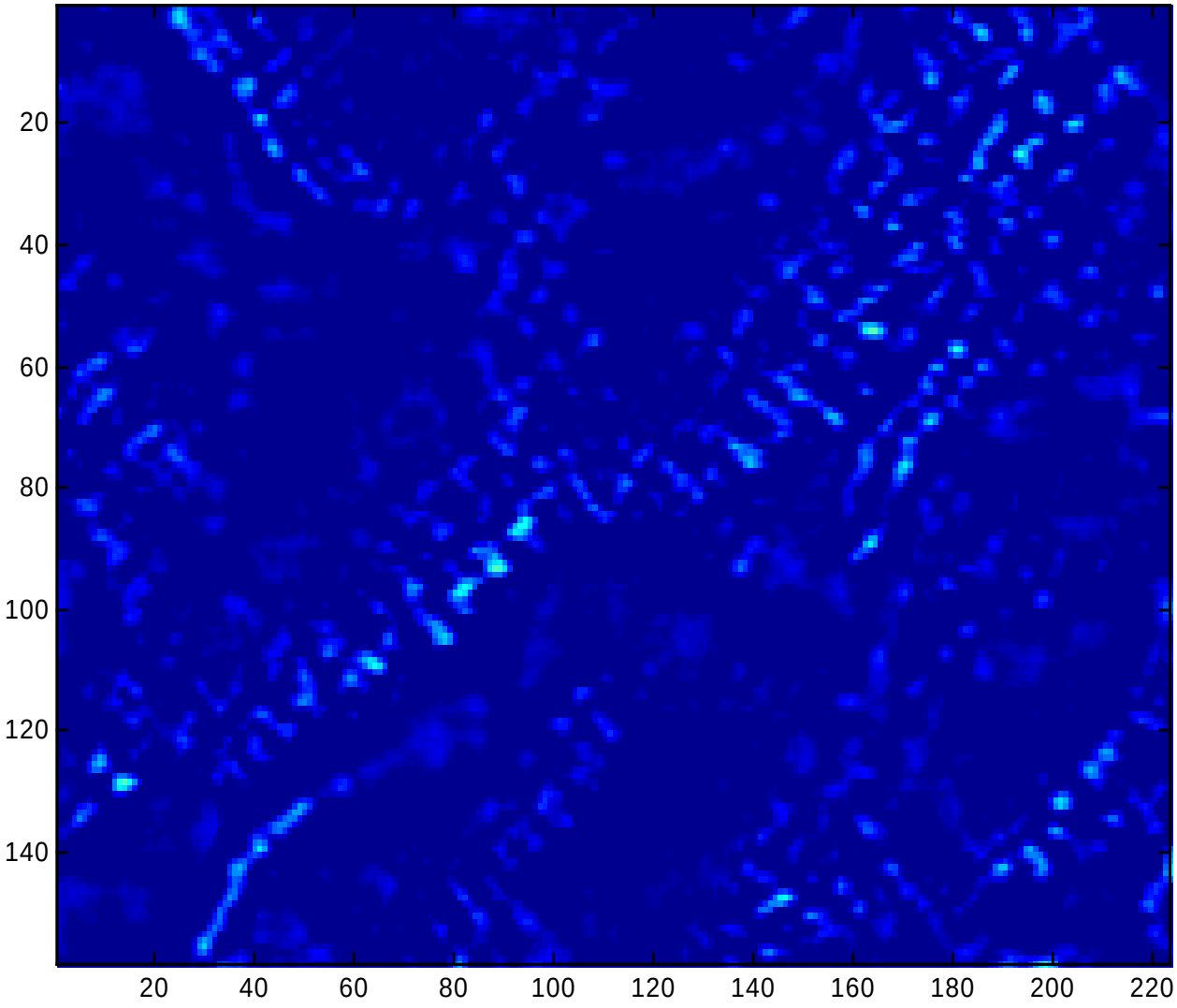


$$X = S_1^T \cdot L_1 \qquad X = S_{1:10}^T \cdot L_{1:10} \qquad X = S_{1:15}^T \cdot L_{1:15}$$

PCA: Application to Image Analysis (example 1: III)

- The residuals of the expansion at the tenth PC put in evidence the sapphyrine film only.

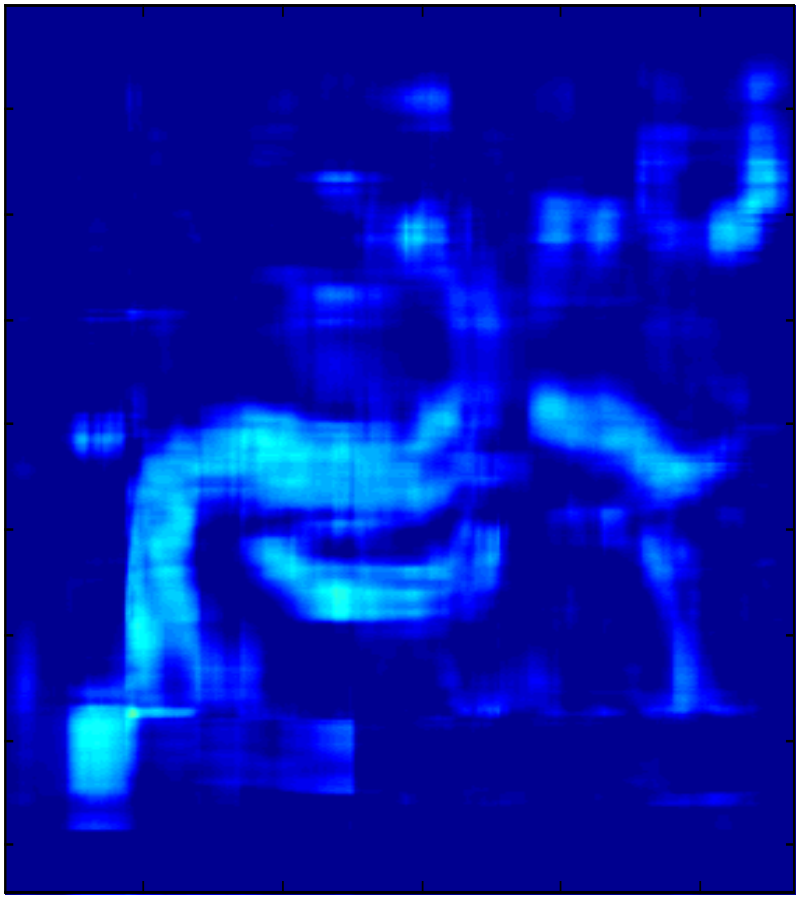
$$X - S_{1:10}^{T} \cdot L_{1:10}$$



PCA: Application to Image Analysis (example 2: I)

- Caravaggio Deposition

PCA: Application to Image Analysis (example 2: II)



$$X = S_{1:10}^{T} \cdot L_{1:10}$$

$$X - S_{1:10}^{T} \cdot L_{1:10}$$

PCA e pattern recognition

- La analisi delle componenti principali è un metodo di analisi che consente:
  - Di definire un nuovo set di features (combinazione lineare delle originali) che risultano scorrelate tra di loro
  - Di decomporre la varianza dei dati nella somma della della varianza dei nuovi assi (componenti principali)
  - Di ridurre la rappresentazione dei pattern ad un sottospazio identificato dalle componenti principali di maggiore varianza
  - Di studiare il contributo delle features originali alle componenti principali più significative identificando le features di maggior contributo.

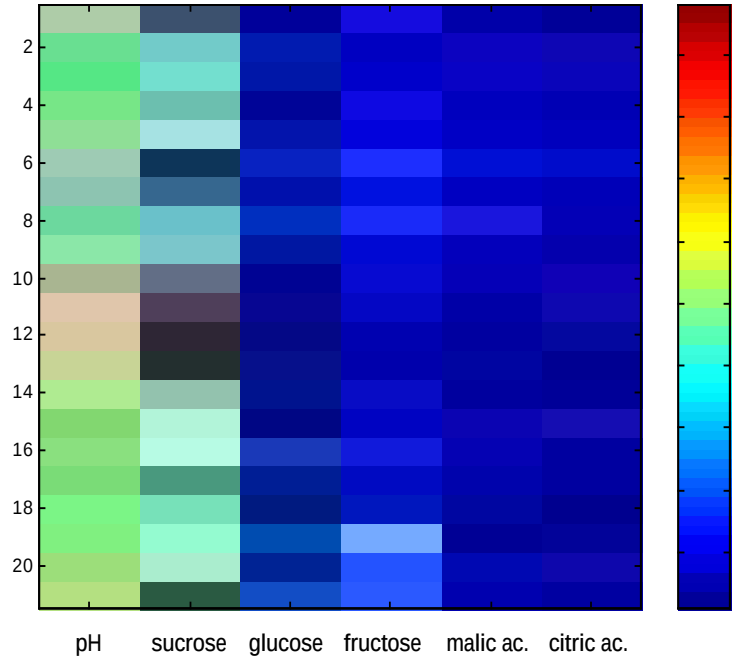
Esempio di applicazione: parametri dei frutti

- Supponiamo di aver misurato le seguenti grandezze in pesche e nettarine di vari cultivars: pH, sucrosio, glucosio, fruttosio, acido malico e acido citrico e di voler studiare la classificazione e la relazione di questa con i singoli parametri.

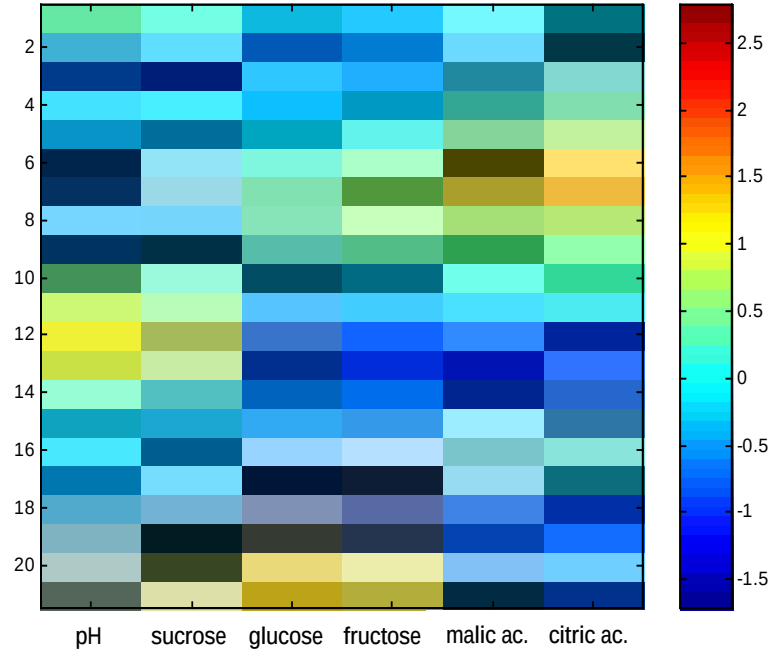
	pH	*sucrose*	*glucose*	*fructose*	*malic acid*	*citric acid*
baby gold	***4.10***	***8.80***	***0.80***	***1.20***	***0.60***	***0.20***
grezzano	***4.0***	***7.0***	***0.60***	***0.80***	***0.50***	***0.10***
iris rosso	***3.50***	***4.30***	***0.90***	***1.0***	***0.40***	***0.60***
maria aurelia	***4.10***	***7.30***	***0.80***	***1.10***	***0.40***	***0.60***
snow queen	***3.90***	***5.70***	***0.80***	***1.30***	***0.50***	***0.50***
spring star	***3.60***	***9.40***	***1.40***	***1.90***	***1.0***	***0.50***
super crimson	***3.70***	***8.20***	***1.0***	***1.10***	***0.90***	***0.60***
venus	***4.10***	***7.40***	***1.60***	***2.20***	***0.70***	***0.40***
argento roma	***3.60***	***4.40***	***0.90***	***1.10***	***0.40***	***0.50***
beauty lady	***3.90***	***8.30***	***0.50***	***0.70***	***0.60***	***0.30***
big top	***4.50***	***8.60***	***0.90***	***1.30***	***0.50***	***0.40***
doucer	***4.40***	***9.80***	***0.70***	***0.80***	***0.40***	***0.10***
felicia	***4.60***	***9.30***	***0.50***	***0.50***	***0.20***	***0.20***
kurakata	***4.40***	***6.90***	***0.60***	***0.80***	***0.20***	***0.20***
lucie	***3.90***	***6.40***	***0.80***	***1.0***	***0.70***	***0.20***
morsinai	***4.10***	***5.80***	***1.60***	***1.90***	***0.50***	***0.60***
oro	***3.80***	***7.70***	***0.40***	***0.40***	***0.60***	***0.20***
royal glory	***4.0***	***6.70***	***0.80***	***0.90***	***0.40***	***0.10***
sensation	***4.70***	***4.60***	***2.0***	***3.40***	***0.30***	***0.20***
sweet lady	***4.20***	***5.50***	***1.30***	***2.10***	***0.50***	***0.40***
youyeong	***4.90***	***8.80***	***1.80***	***2.50***	***0.20***	***0.10***

Parametri delle pesche: mappe di colore dei dati

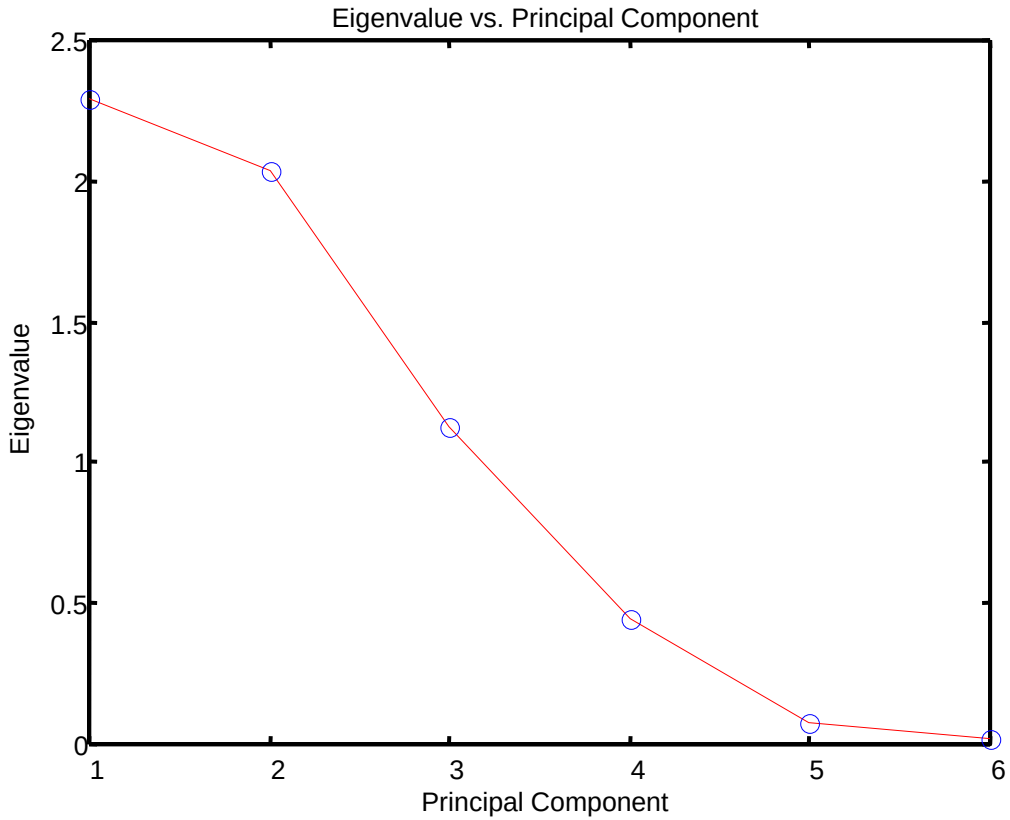
raw data



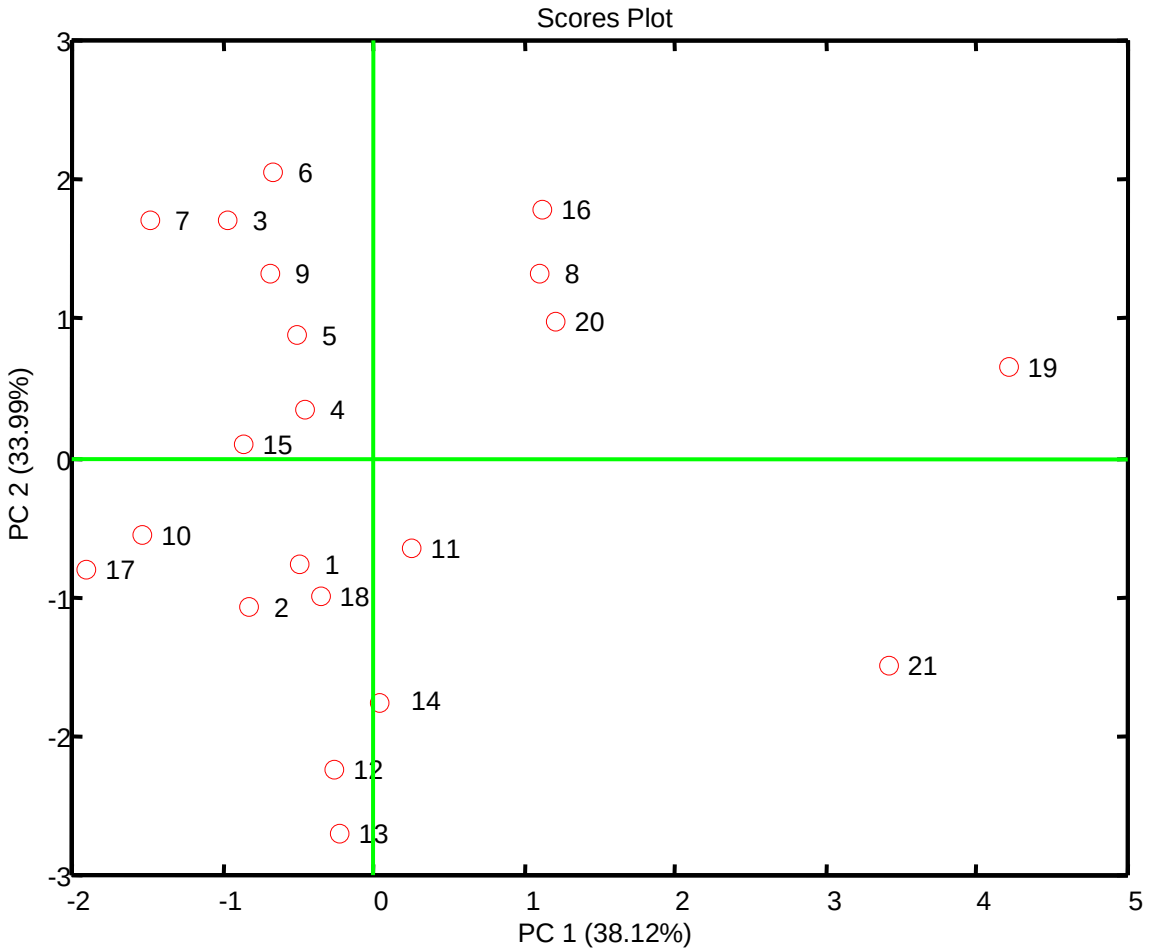
Autoscaled data



PCA: dati pesche
gli autovalori vs. PC (scree plot)



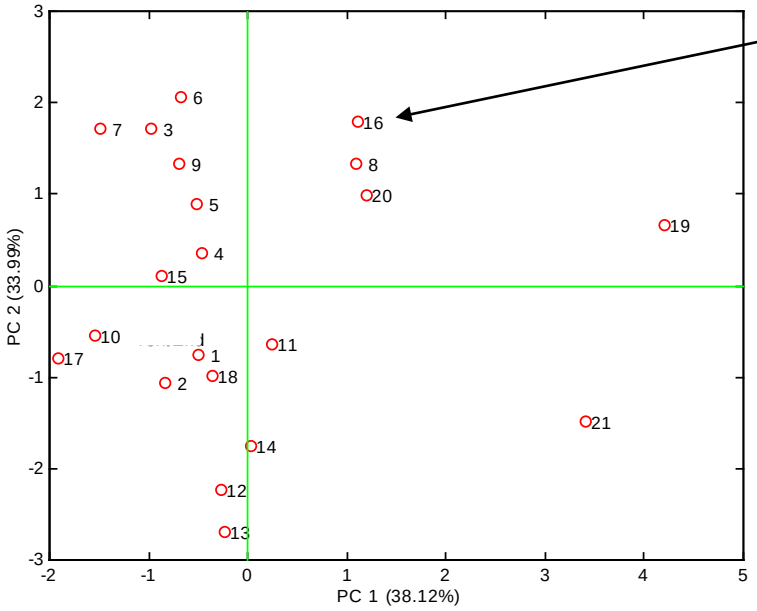
PCA: dati pesche
scores plot

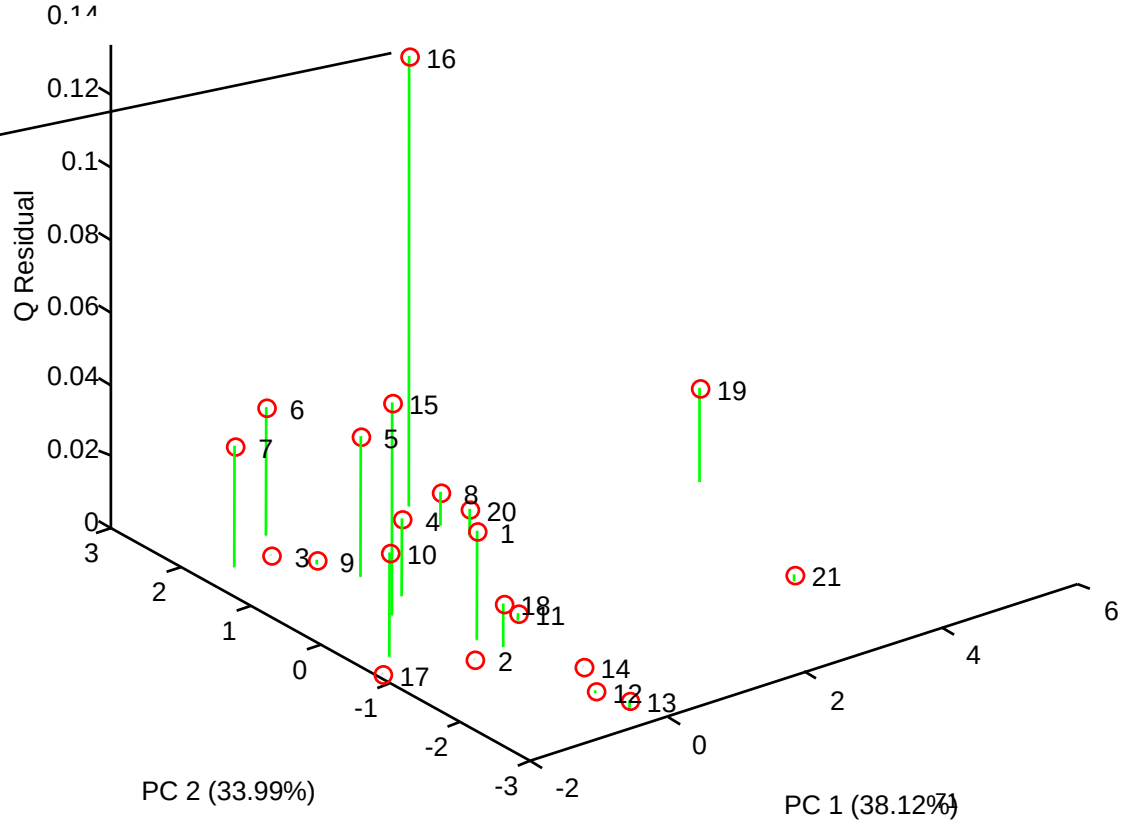


Rappresentazione dei residui

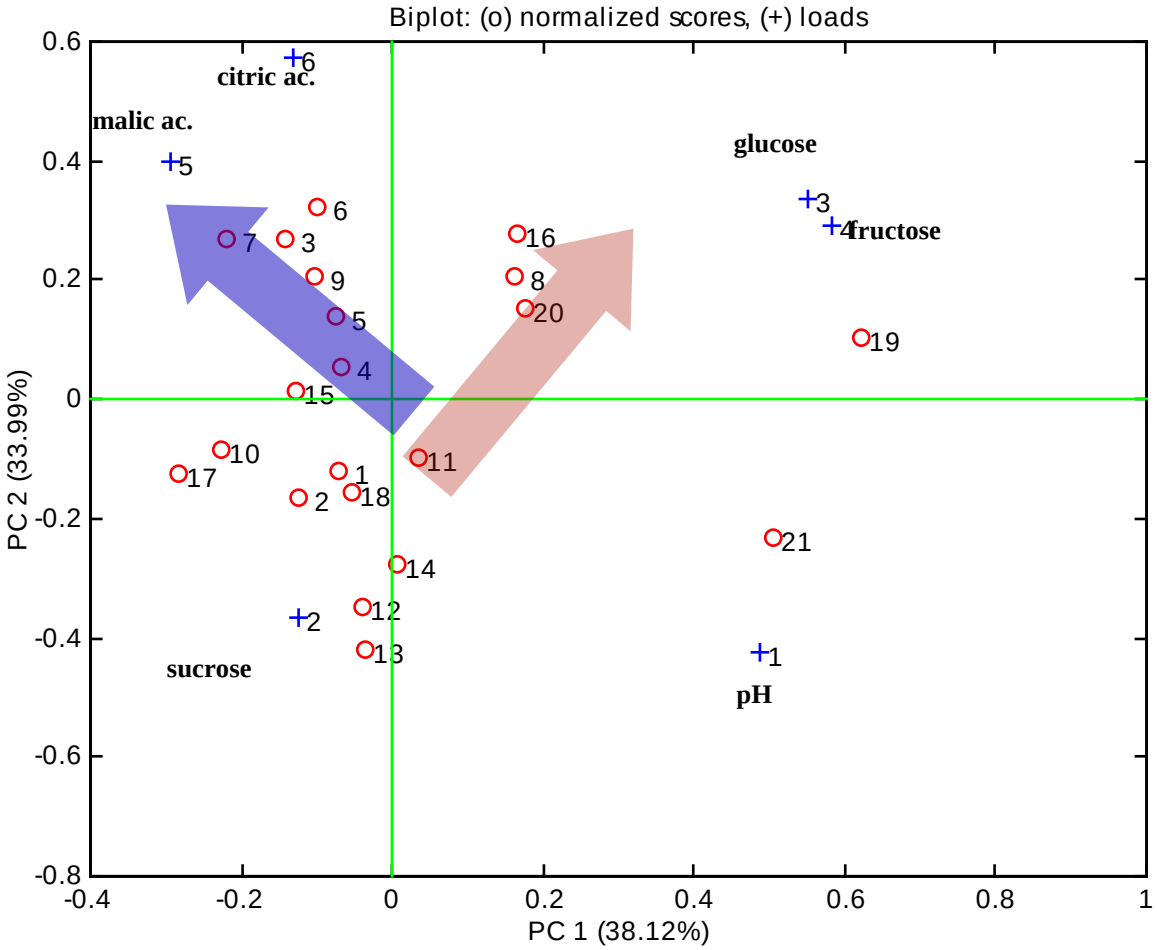
$$x_i = a \cdot s_1 + b \cdot s_2 + \ldots + n \cdot s_n$$

$$x_i^{pca} = a \cdot pc_1 + b \cdot pc_2 + residual$$



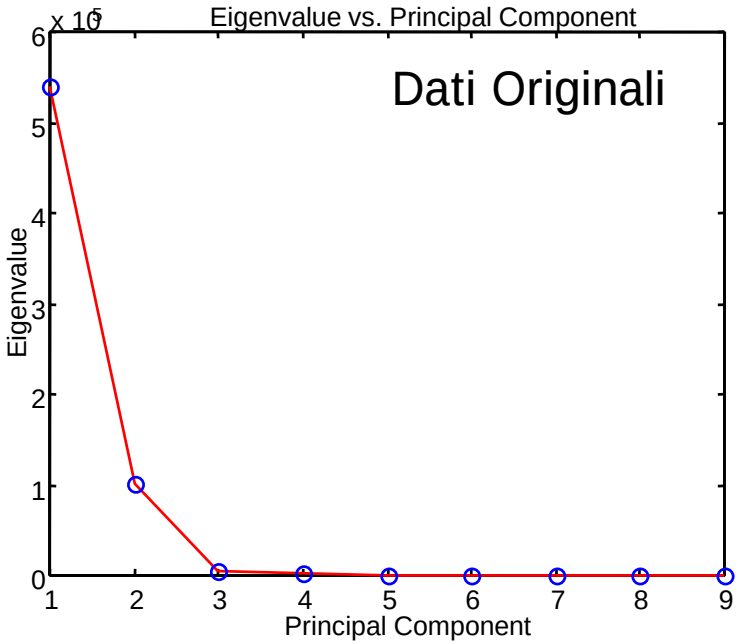


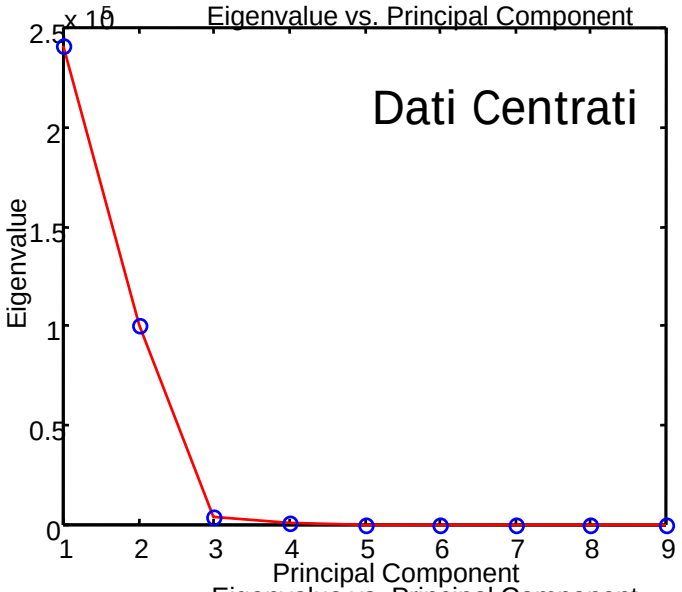
PCA example: peaches data
bi-plot: scores+loadings

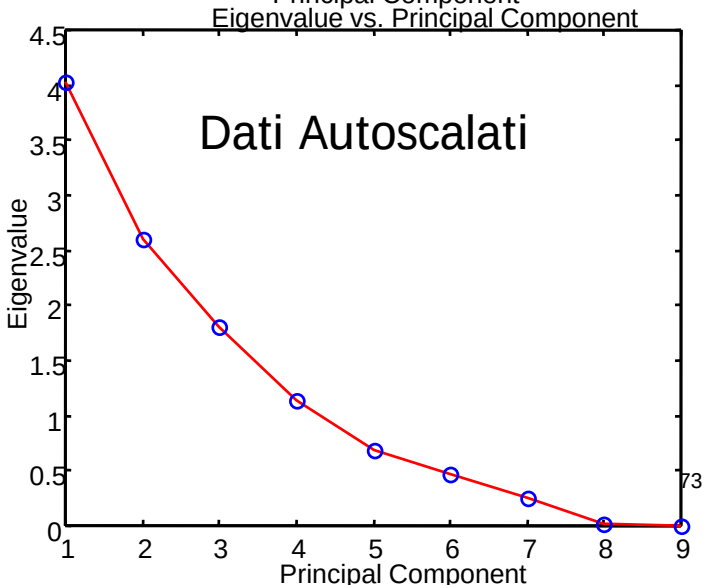


- Gli zuccheri sono ortogonali agli acidi
- Identifichiamo la direzione della acidità e quella della dolcezza
- Il sucrosio è anticorrelato con glucosio e fruttosio
- Il pH ovviamente è anticorrelato con gli acidi

PCA esempio: acque minerali autovalori vs. pc



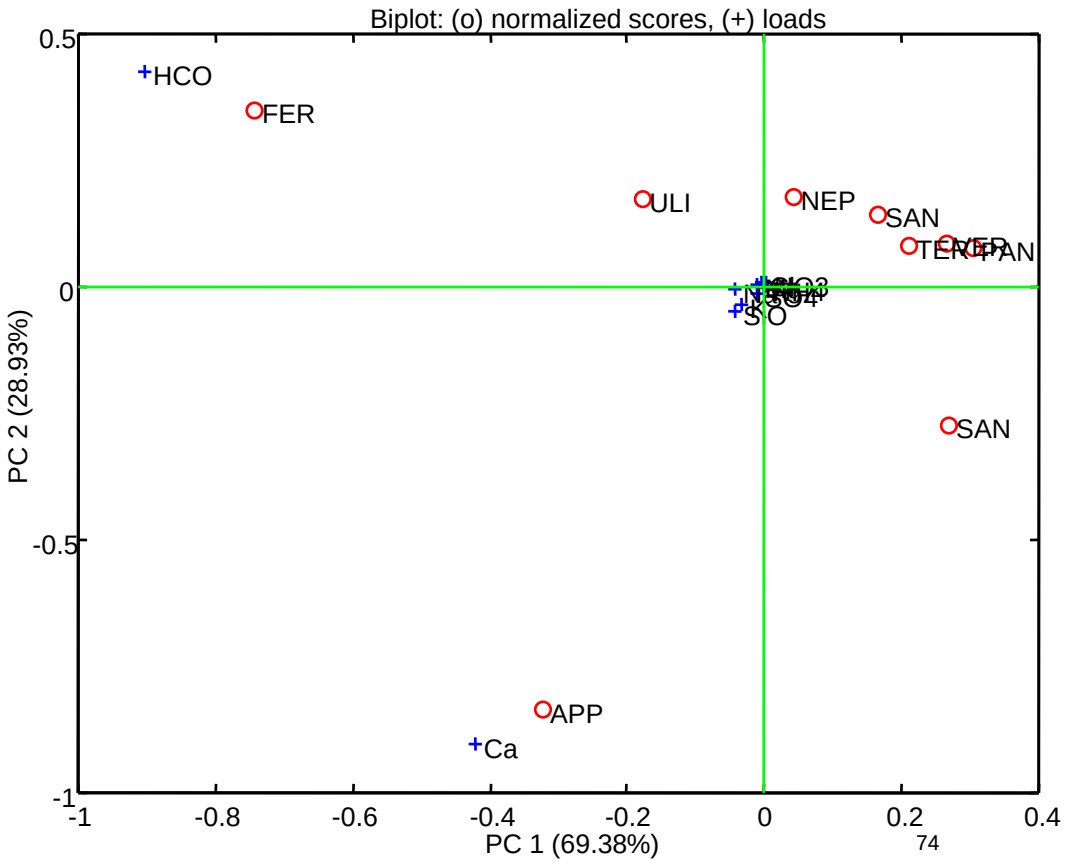




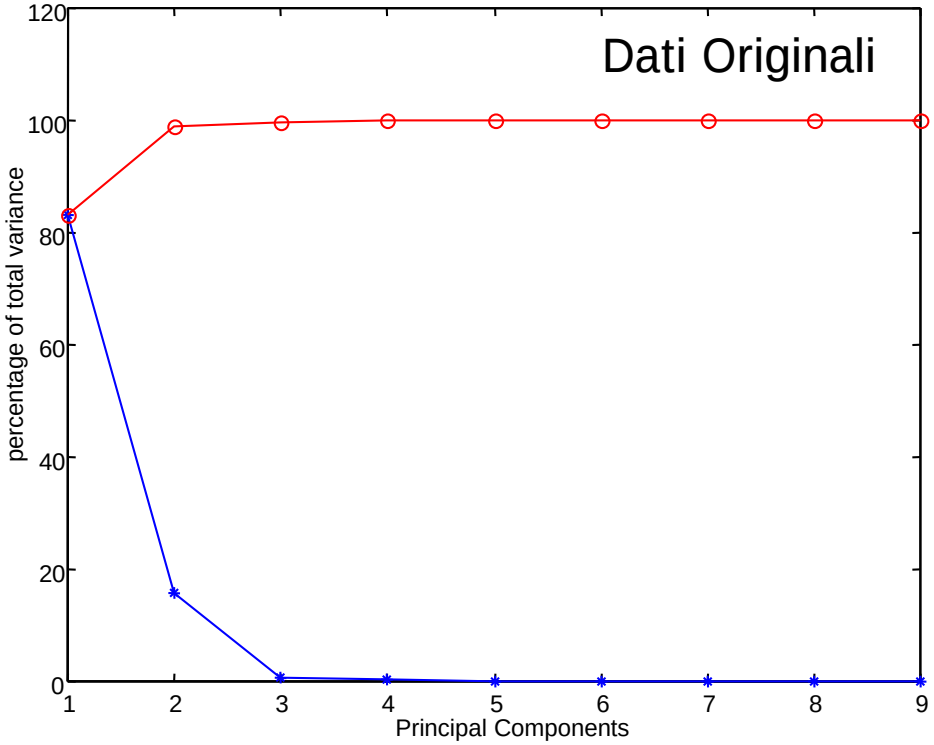
L'autoscaling rende le features omogenee e quindi aumenta il numero di dimensioni importanti

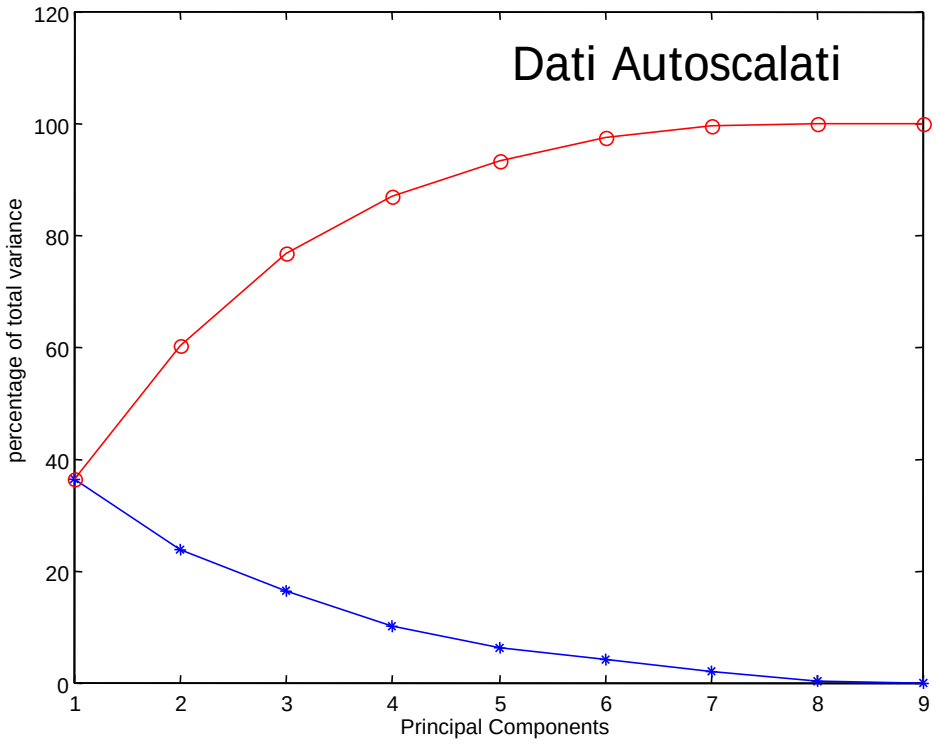
Acque minerali:
PCA biplot dati originali

- Solo le features numericamente importanti contano (HCO e Ca)
- Le altre features sono concentrate attorno all'origine e non contribuscono alla classificazione
- HCO e Ca sono ortogonali
- **Tutto ciò che è ortogonale nello score plot è scorrelato**
- Solo FER e APP si differenziano dalle altre
- 98% di varianza in questo plot



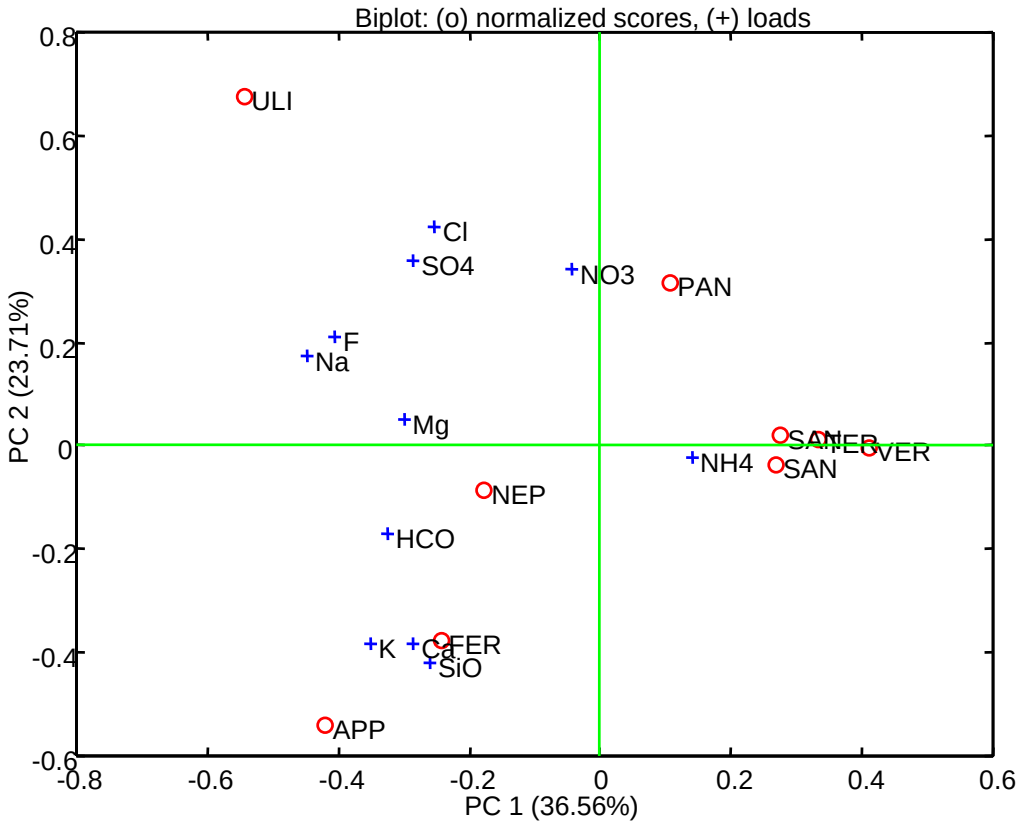
Scree plot in termini di varianza



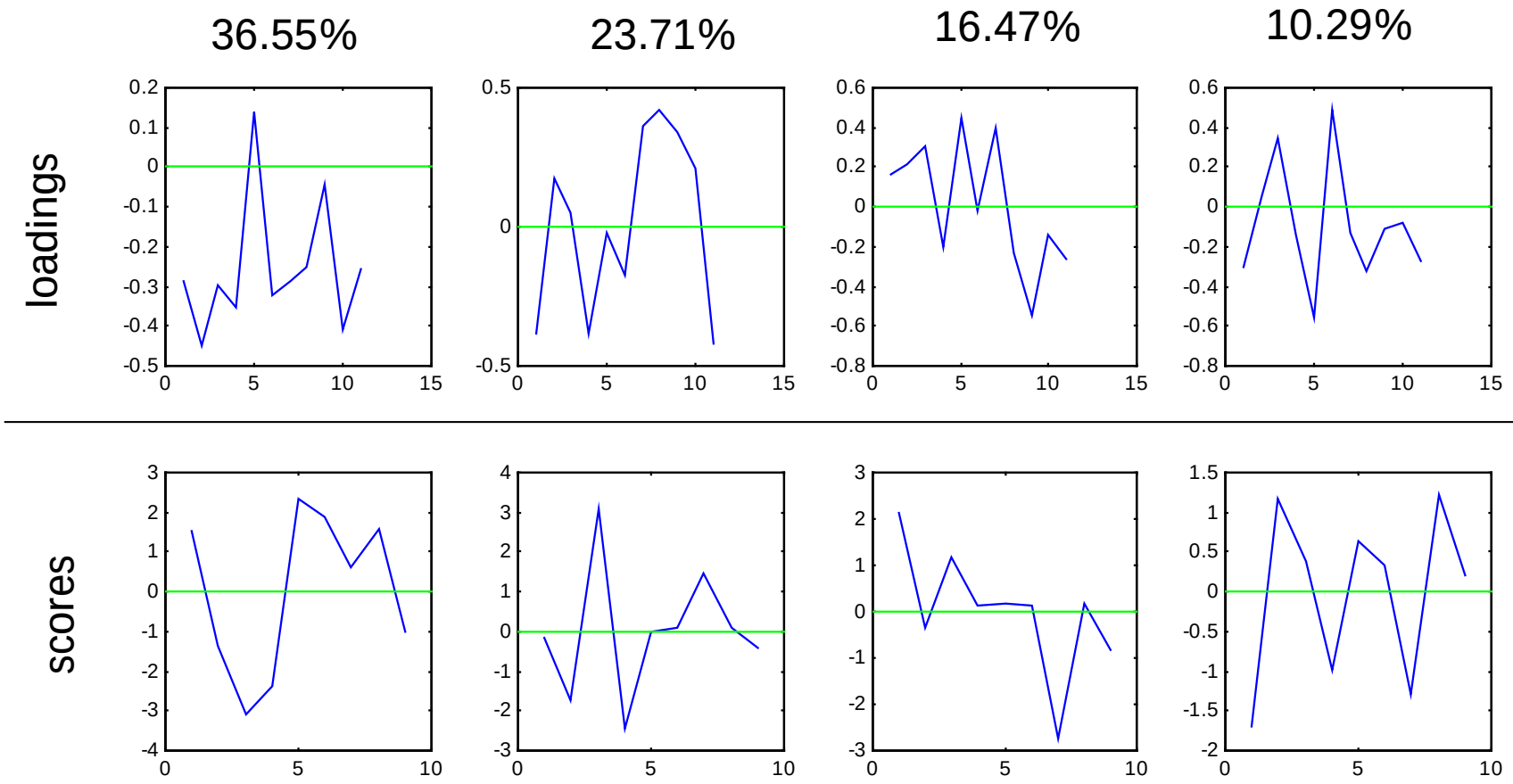


Acque minerali:
PCA biplot dati autoscalati

- Le features contribuiscono omogeneamente
- Sono identificabili I seguenti gruppi di acque:
  - SAN, TER, VER, SAB
    - Oligo minerali
  - PAN
    - Oligo ma con incremento di NO3
  - NEP, FER, APP
    - Incrementi di Mg, HCO, Ca, K
  - ULI
    - Incrementi di Cl, SO4
  - Per ULI,NEP,FER, APP
    - Incremento comune di F, Na
- 60 % di varianza in questo plot
- E l'altro 40%?



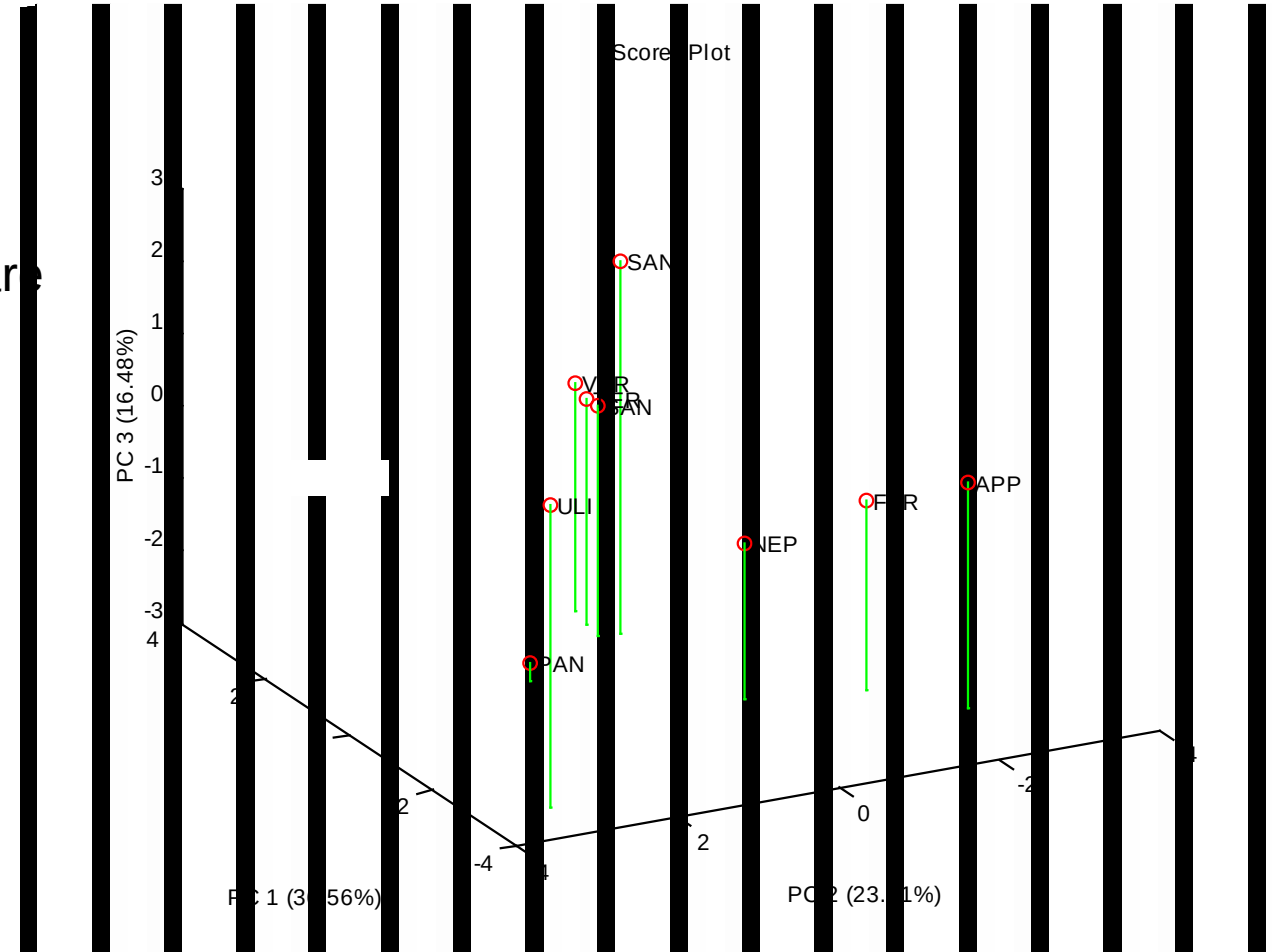
Acque minerali: analisi di loadings e scores



PCA: acque minerali

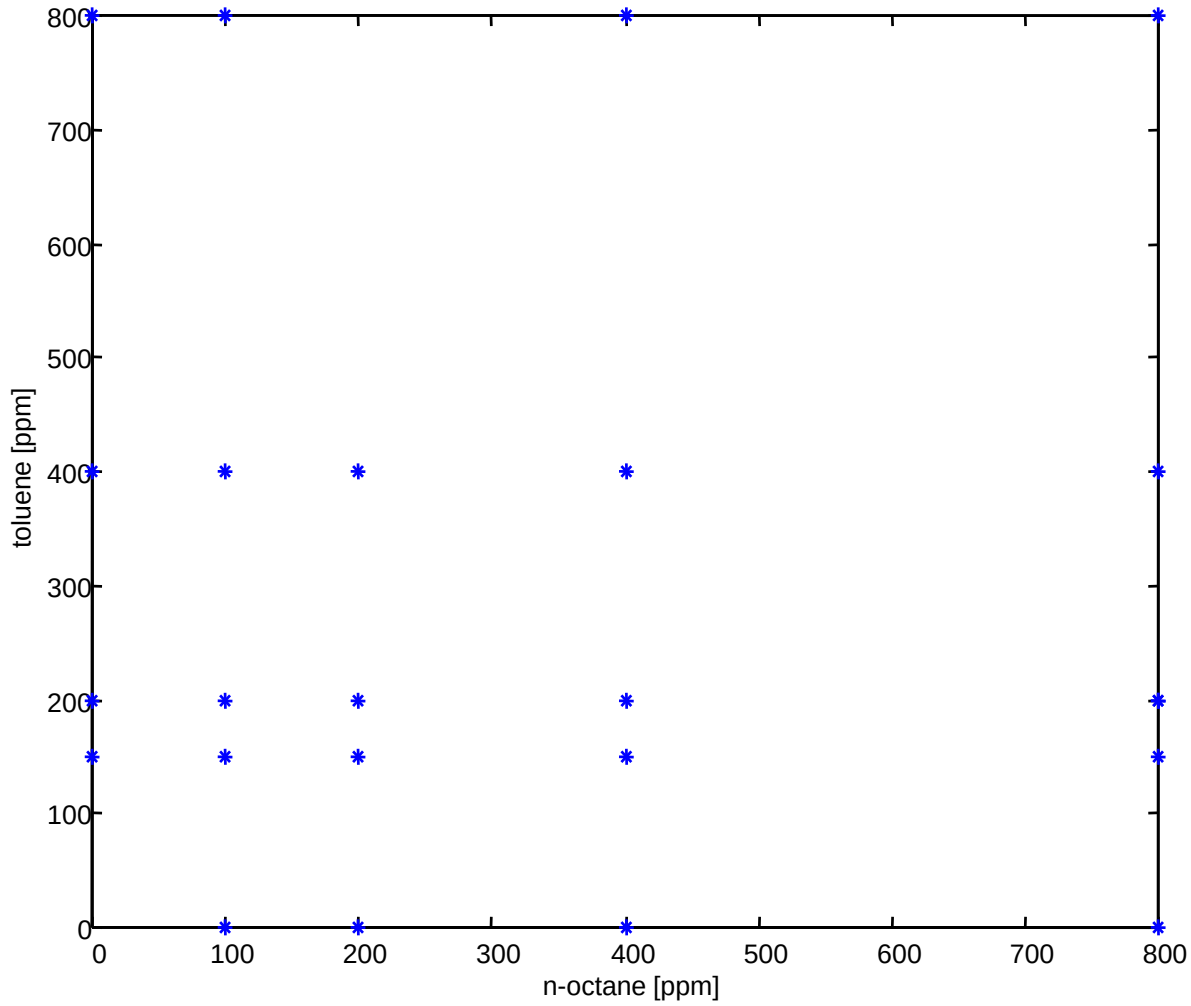
- La rappresentazione 2D non è sufficiente perchè la distribuzione degli autovalori non lo consente
  - Ci sono rappresentazioni 2D parziali che colgono differenti aspetti del problema.

Score plot 3D
76% di varianza
Emerge il carattere peculiare di SAN

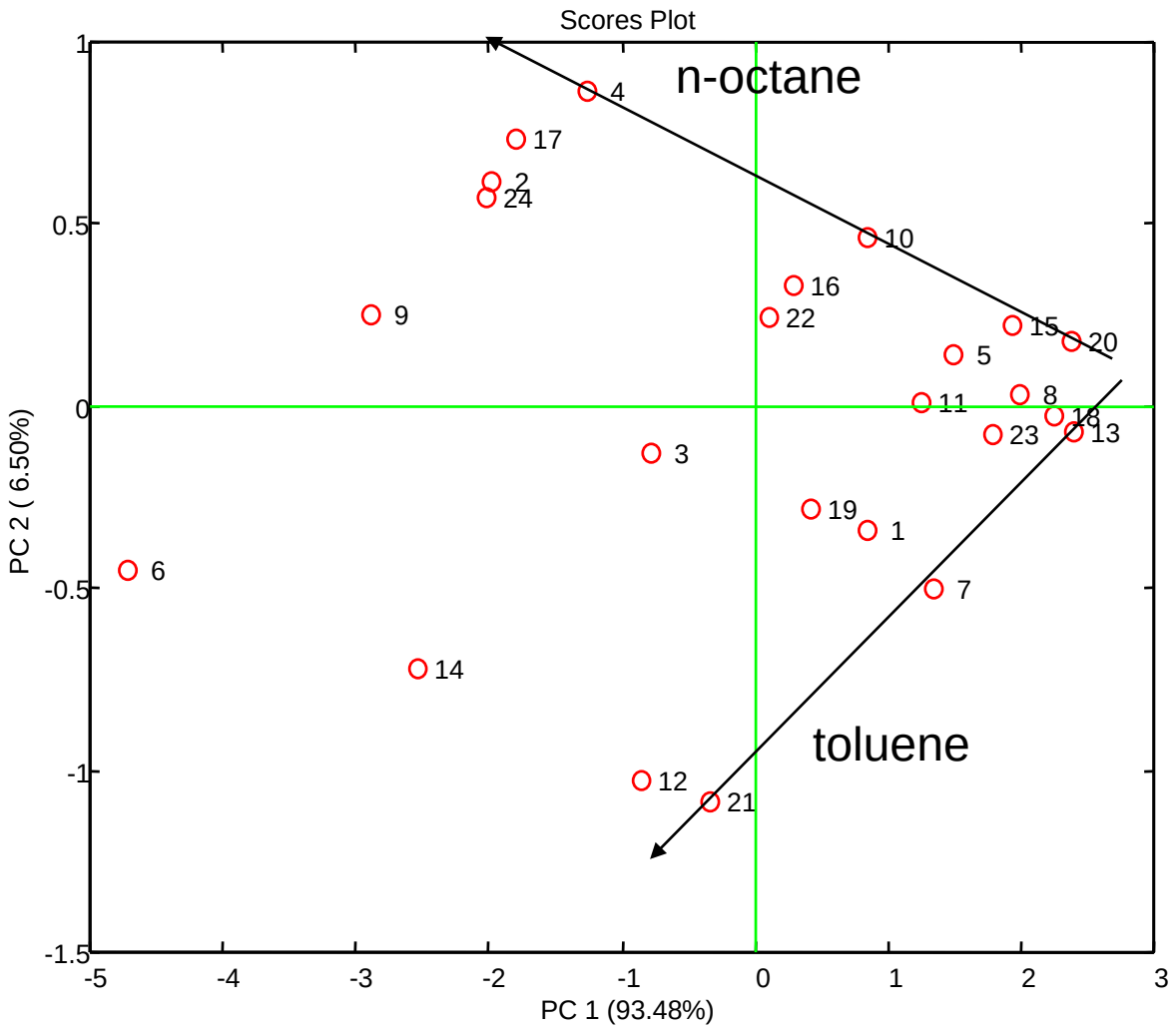


4 sensori for 2 gas

Plot delle concentrazioni misurate



PCA scores


Scores Plot
n-octane
toluene
PC 2 (6.50%)
PC 1 (93.48%)
1
0.5
0
-0.5
-1
-1.5
-5
-4
-3
-2
-1
0
1
2
3
4
17
2
24
10
16
9
22
15
20
5
11
8
18
23
13
3
19
1
6
7
14
12
21

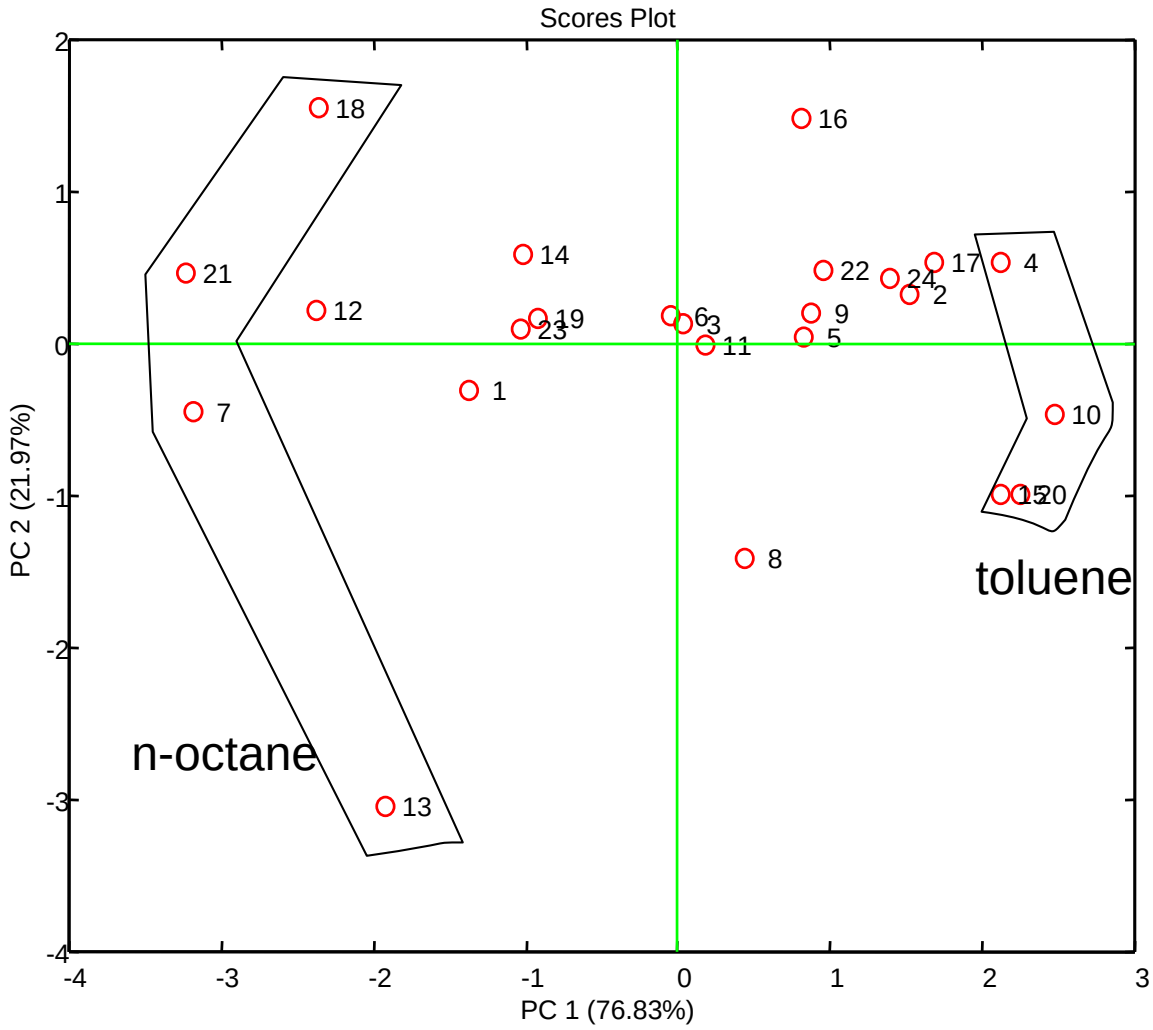
Normalizzazione lineare

Estrazione delle informazioni qualitative

$$s_i = K_{ij} \cdot c_j$$

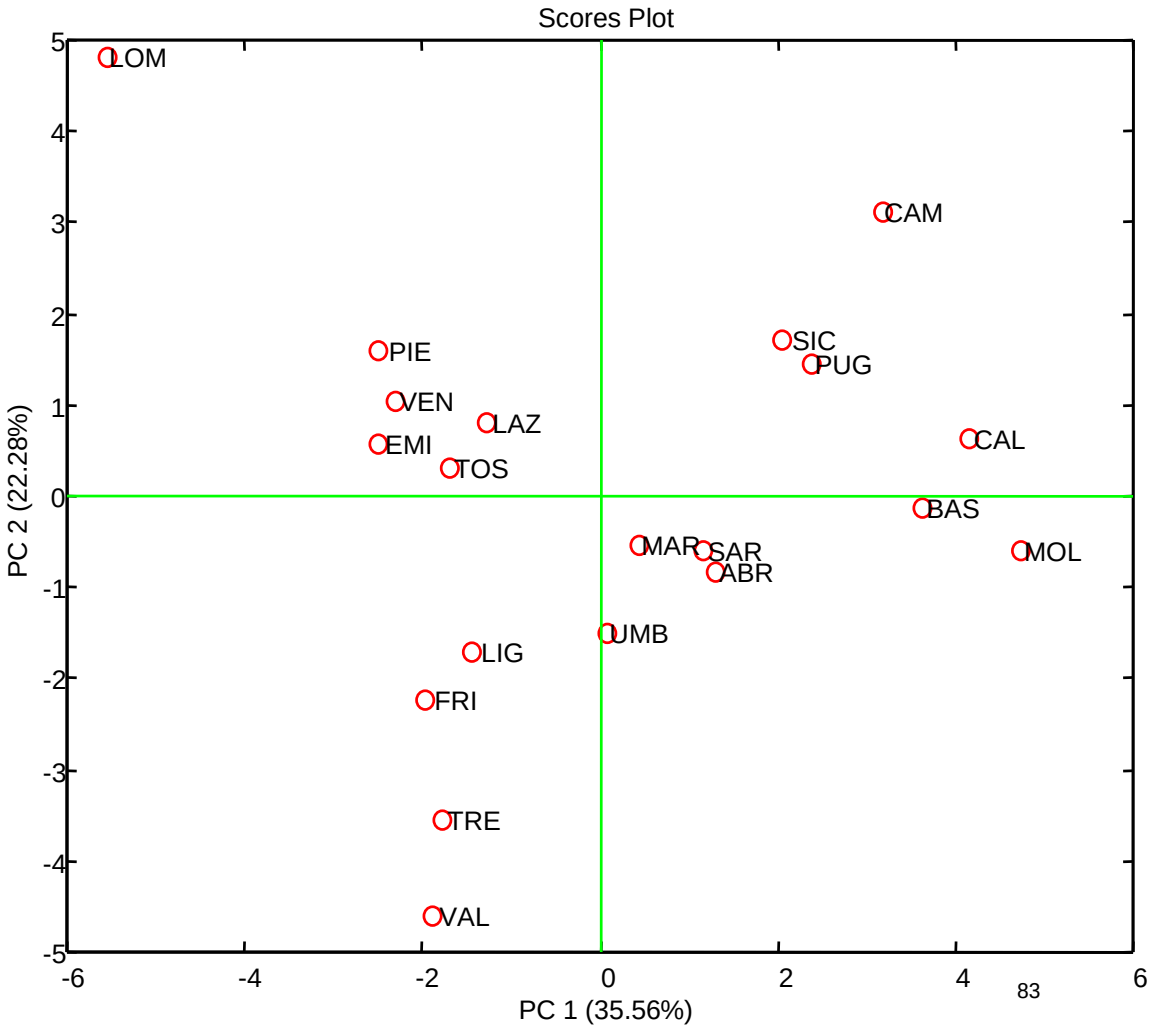
$$\Rightarrow s_i = \frac{s_i}{\sum_m s_m} = \frac{K_{ij} \cdot c_j}{\sum_m K_{mj} \cdot c_j} = \frac{K_{ij}}{\sum_m K_{mj}}$$

Normalizzazione lineare PCA scores

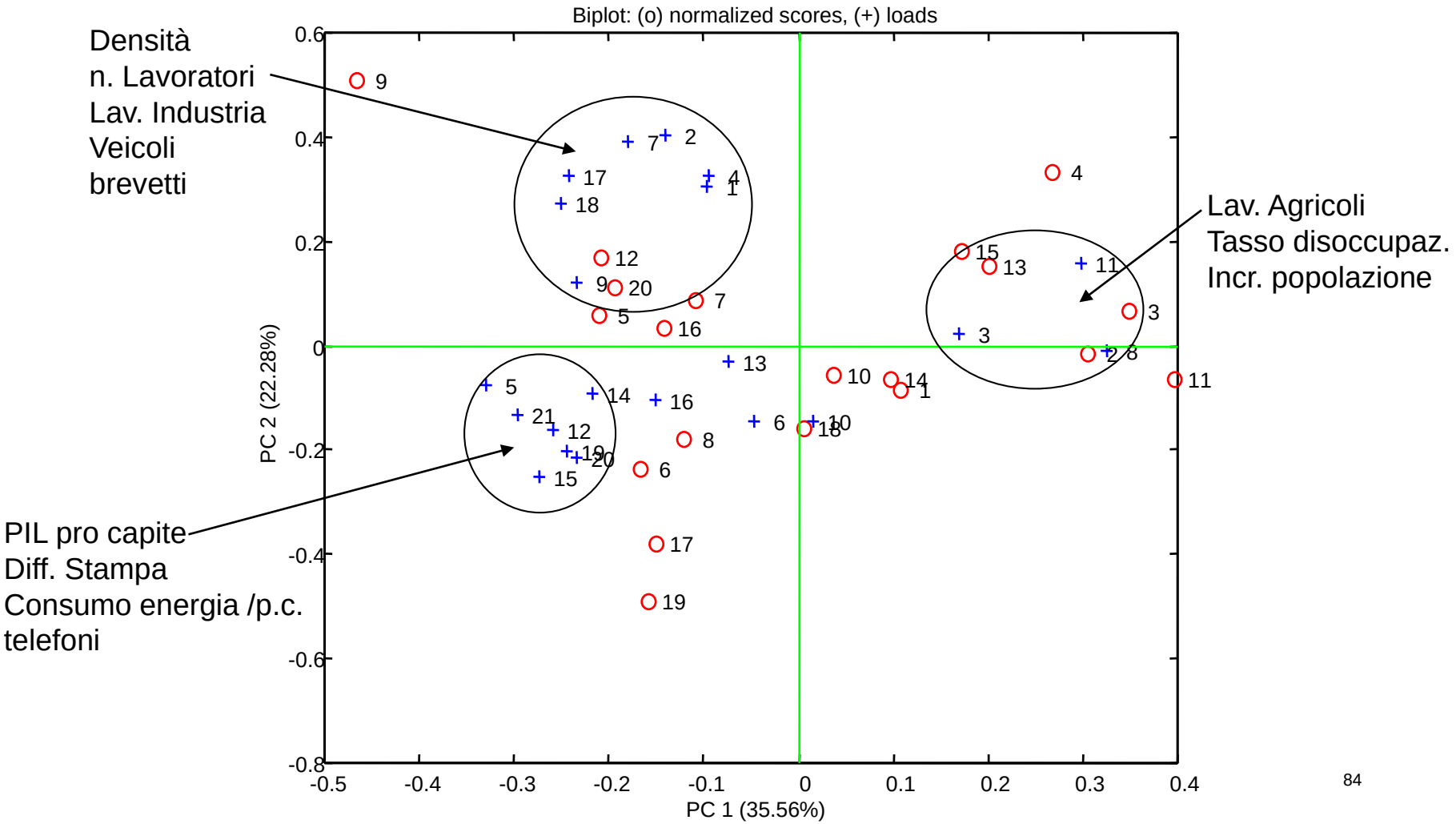


Benessere sociale delle regioni italiane

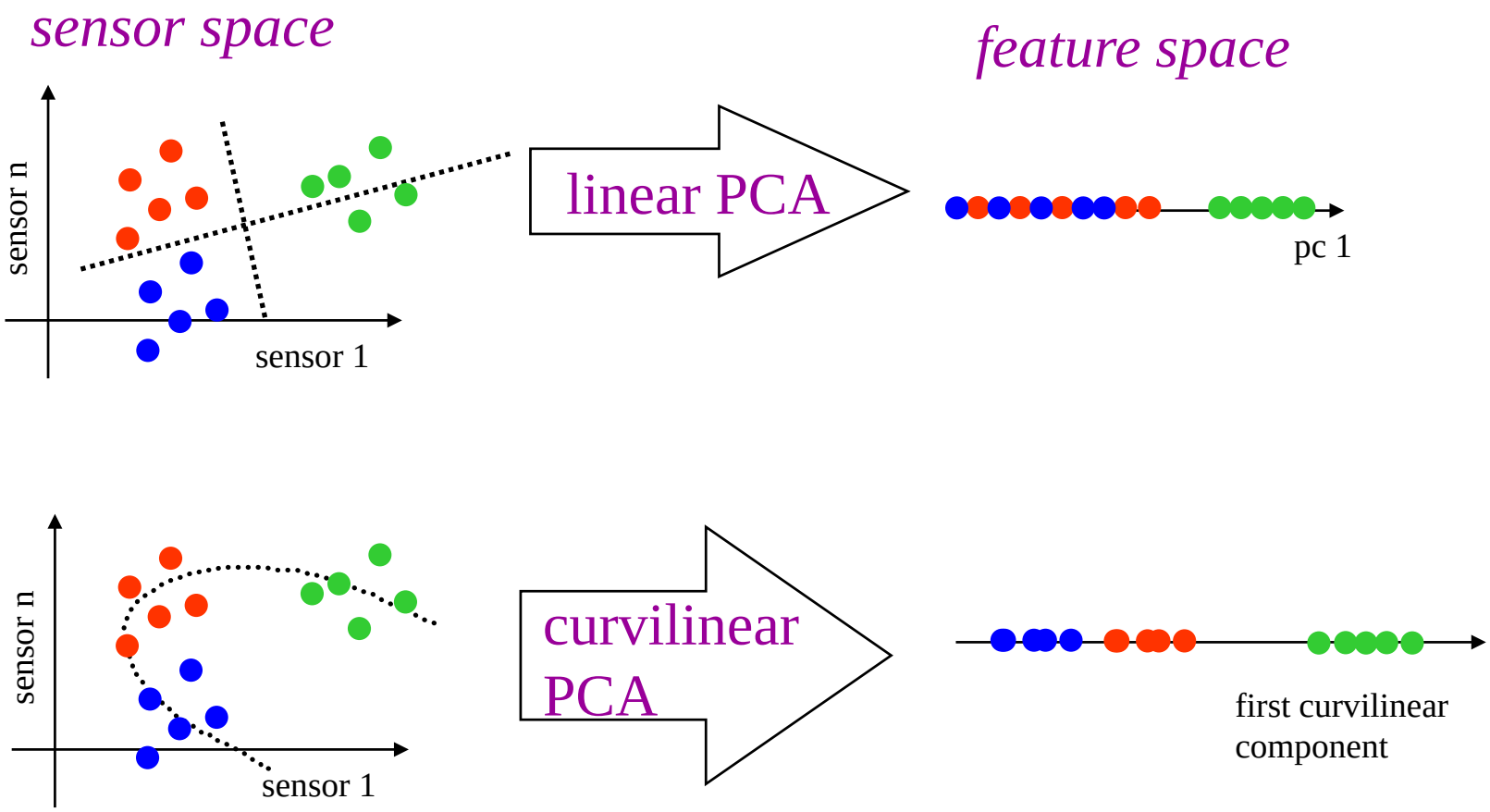
21 indicatori geografici, sociali, ed economici



PCA bi-plot


Densità
n. Lavoratori
Lav. Industria
Veicoli
brevetti
PIL pro capite
Diff. Stampa
Consumo energia /p.c.
telefoni
Lav. Agricoli
Tasso disoccupaz.
Incr. popolazione
Biplot: (o) normalized scores, (+) loads
PC 1 (35.56%)
PC 2 (22.28%)

PCA: esempio di ra



Limiti della rappresentazione PCA

- La rappresentazione offerta dalla PCA è guidata dalle caratteristiche della matrice di covarianza dei dati
  - Se i dati non sono distribuiti normalmente la matrice di covarianza non esaurisce il contenuto statistico dei dati stessi, quindi la rappresentazione PCA risulta formalmente non corretta

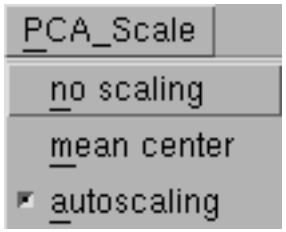
- Lo score plot della PCA è una proiezione lineare da uno spazio a dimensione N ad uno spazio a dimensione 2 o 3. Sono possibili effetti di falsa proiezione che comportano errori di classificazione
  - Effetto costellazioni della volta celeste

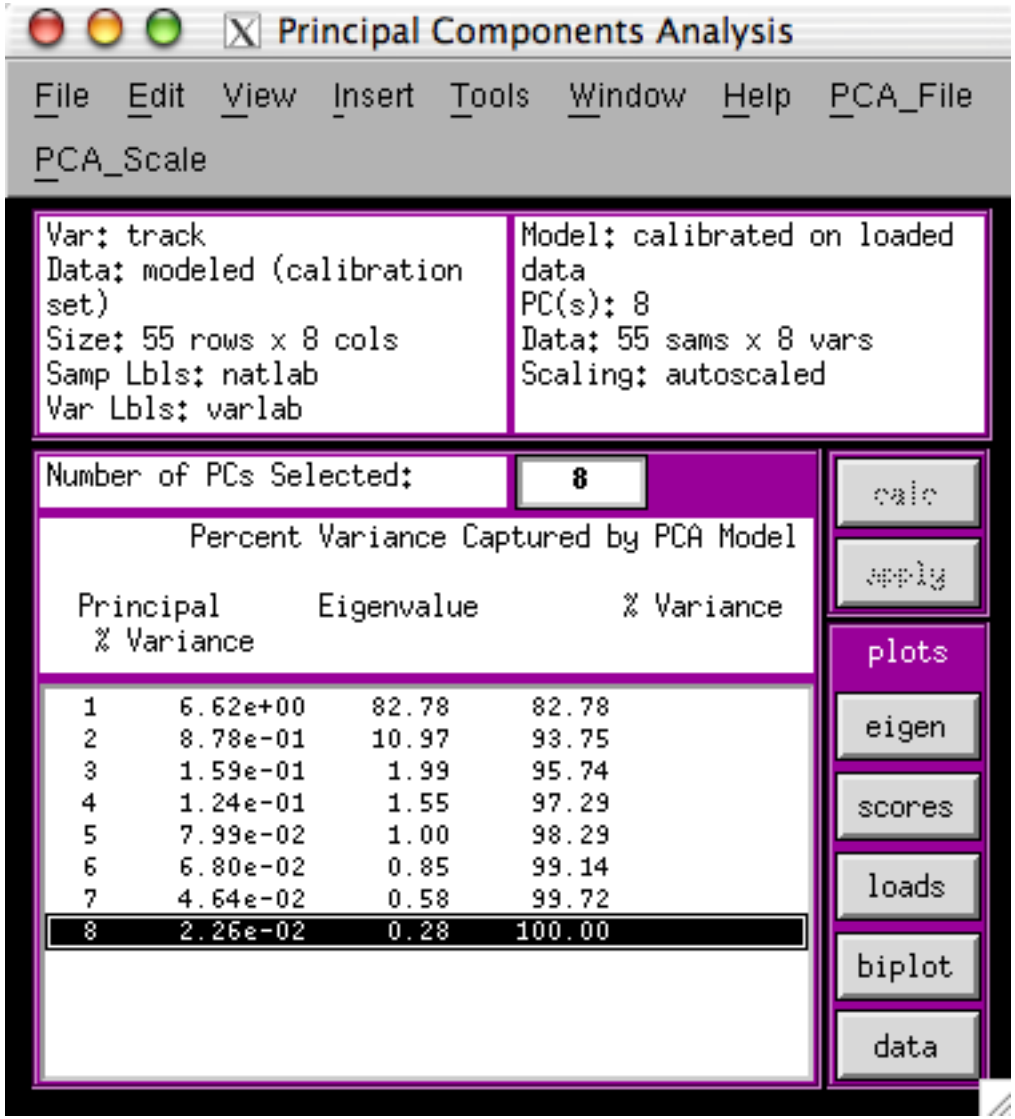
Matlab PLS Toolbox
pcagui

- Il comando apre un interfaccia Guided User Interface a finestre e bottoni che consente dio eseguire la PCA di una matrice gia presente nel workspace di matlab
- I dati possono essere normalizzati (mean center e autoscale)
- Il modello può essere salvato in matlab
- La gui genera I seguenti grafici: autovalori, scores, loadings, bi-plot
- Si possono visualizzare I residui

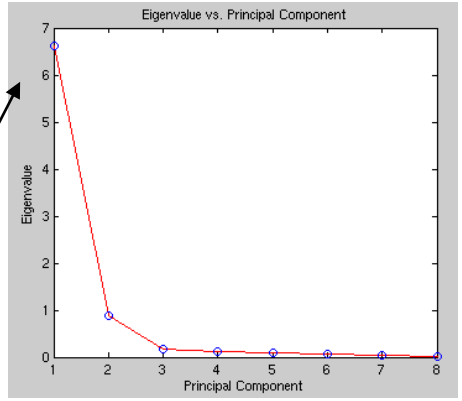
- Esempio: tracks records

Esempio di *pcagui*
dati: tracks records





Esempio di *pcagui* dati: tracks records



Principal Components Analysis

File Edit View Insert Tools Window Help PCA_File
PCA_Scale

Var: track
Data: modeled (calibration set)
Size: 55 rows x 8 cols
Samp Lbls: natlab
Var Lbls: varlab

Model: calibrated on loaded data
PC(s): 8
Data: 55 sams x 8 vars
Scaling: autoscaled

Number of PCs Selected: 8

Percent Variance Captured by PCA Model

Principal	Eigenvalue	% Variance	% Variance
1	6.62e+00	82.78	82.78
2	8.78e-01	10.97	93.75
3	1.59e-01	1.99	95.74
4	1.24e-01	1.55	97.29
5	7.99e-02	1.00	98.29
6	6.80e-02	0.85	99.14
7	4.64e-02	0.58	99.72
8	2.26e-02	0.28	100.00

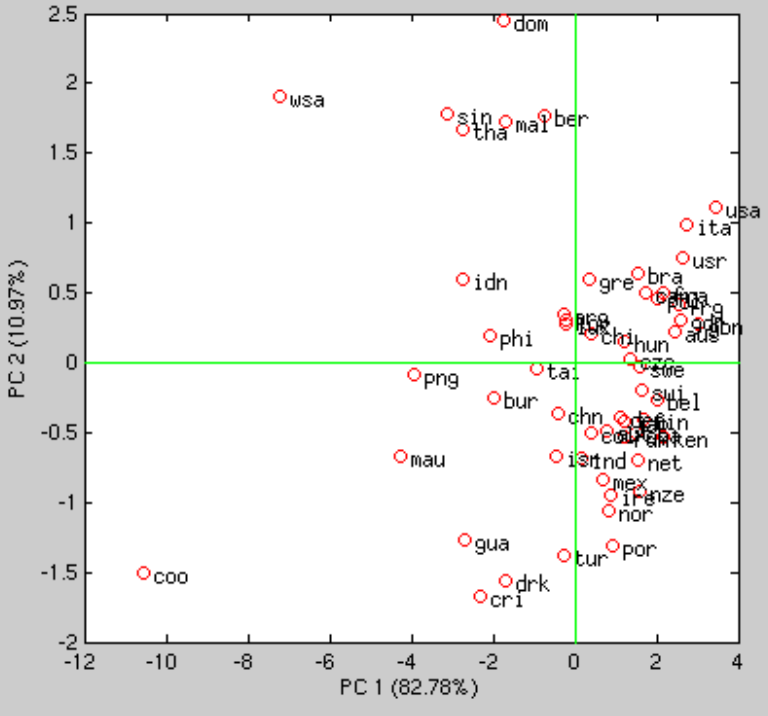
calc
apply
plots
eigen
scores
loads
biplot
data

Plot Scores

max PC 8
x PC 1
y PC 2
z none
labels
lim 95 %
plot

samples
spawn info
Q con T con
data delete

zoom
in out
home

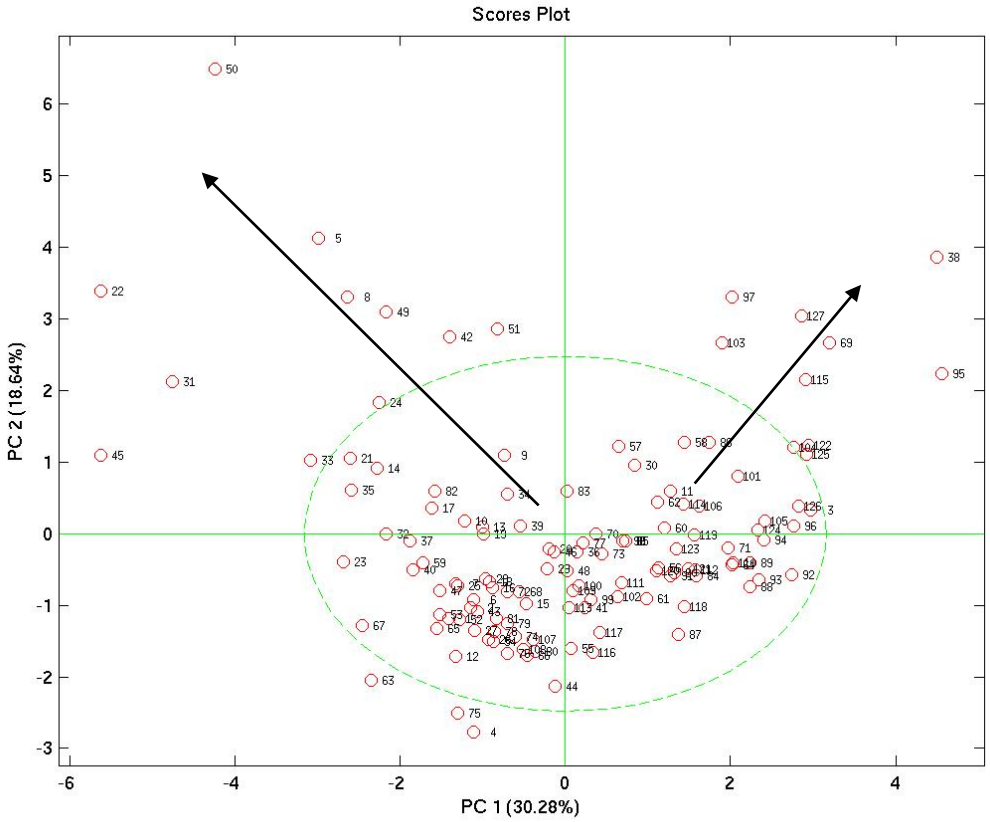


Esempio PCA: outliers detection
studio della composizione di acque dolci

- Acque campionate in 127 località in Texas
- 12 quantità misurate:
  - U, As, B, Ba, Mo, Se, V, Solfati, alcalinità totale, bicarbonati, conducibilità, pH
- Analisi con PCA
- Ricerca delle anomalie:
  - Campioni statisticamente differenti
  - Probabili sorgenti inquinate

Score plot

- L'ipotesi fondamentale della PCA è che I dati siano distribuiti normalmente, in questo caso le superfici di iso-proabilità sono ellissi che nella base delle componenti principali sono rappresentate in forma canonica.
- Sullo score plot vengono quindi rappresentate le curva di isoproabilità, in particolare graficando il contorno di probabilità al 95% si evidenziano gli outliers
- Lo score plot evidenzia due direzioni di deviazione dalla popolazione normale



Bi-plot

- Le due direzioni di deviazione dalla popolazione normale sono orientate verso le seguenti variabili
  - 1: + solfati, B, conducibilità
  - - Ba
  - 2: U, Mo, V, As