

ANALISI DELLE COMPONENTI PRINCIPALI (PCA)

Federico Marini

Introduzione

- Come detto, gli esseri umani sono particolarmente validi quando si tratta di riconoscere somiglianze e differenze tra gli oggetti.
- Ad esempio, sin da bambini si è educati a riconoscere le forme (un quadrato da una sfera,...)
- In chimica analitica, spesso il problema è analogo: riconoscere somiglianze e differenze tra gli oggetti sulla base di una serie di misure chimiche.
- Per introdurre l'analisi delle componenti principali, partiamo da un esempio semplice.
- Supponiamo di dover analizzare quattro campioni e di voler determinare quali di essi siano simili tra di loro, avendo misurato pH, temperatura e densità

Introduzione - 2

- I risultati ottenuti sui 4 campioni sono riportati in Tabella:

Sample	pH	Temperature (°C)	Density (g/mL)
1	5	20	1.1
2	7	80	0.8
3	7	80	0.8
4	5	20	1.1

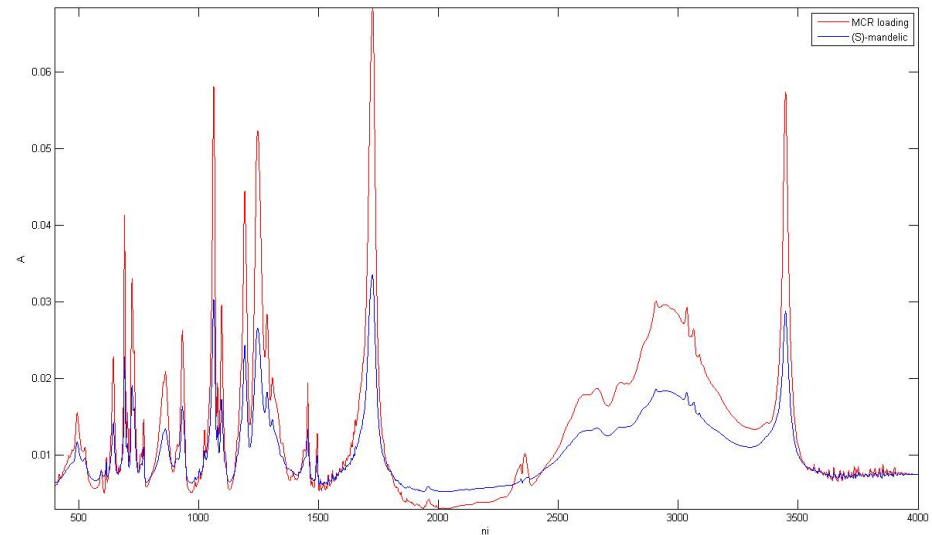
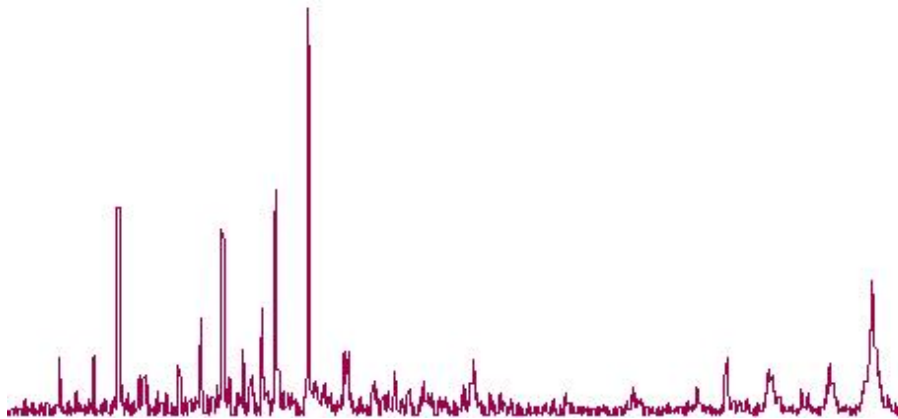
- Con un colpo d'occhio è facile affermare che i campioni 1 e 4 sono simili tra loro e differenti dai campioni 2 e 3
- Le capacità umane permettono di cogliere le differenze tra le righe (campioni) di semplici matrici di dati con pochi elementi.
- Questa capacità però è limitata al caso in cui si abbiano poche righe e/o colonne.
- Infatti, se al posto di questa tabella si considerasse quella riportata nel lucido seguente, nonostante i dati siano legati da una relazione matematica piuttosto semplice, l'occhio umano non sarebbe in grado di coglierla

Introduzione - 3

Sample Number	Measurement Variable 1	Measurement Variable 2	Measurement Variable 3
1	1.2036	0.3514	0.5336
2	1.5747	2.7024	1.3651
3	1.5610	0.8800	1.3329
4	2.4709	2.1856	1.3781
5	2.1031	3.3598	1.1370
6	3.1262	2.1694	2.4536
7	3.5547	3.5375	1.8349
8	2.8297	3.9520	2.0272
9	3.4619	2.0095	2.3668
10	4.1251	3.7992	3.2149
11	4.3519	4.7840	3.1682
12	4.3726	5.6372	3.1923
13	4.1362	5.5989	3.8251
14	5.1222	3.2704	3.2974
15	4.7702	4.3310	3.3547
16	4.9319	5.0789	3.7546
17	5.0776	5.6956	4.2358
18	4.5237	7.5085	3.8541
19	6.2747	3.1689	3.5633
20	6.2657	3.6446	3.9434
21	6.2770	4.7385	4.3812
22	5.6613	6.4239	4.7231
23	6.0396	6.6905	4.6699
24	7.1459	5.0001	4.4495
25	7.2072	6.5147	4.7595
26	6.5666	7.1129	5.3562
27	6.9216	8.5184	5.7890
28	8.3753	6.0726	5.5092
29	7.7007	7.1723	5.3151
30	8.5584	7.9795	5.6718
31	7.3148	9.0592	6.3253
32	8.7246	8.6834	6.5414
33	9.0110	8.7975	6.4387
34	9.7705	9.2219	7.0289
35	9.5646	9.9156	8.1231

“Vedere” attraverso il computer

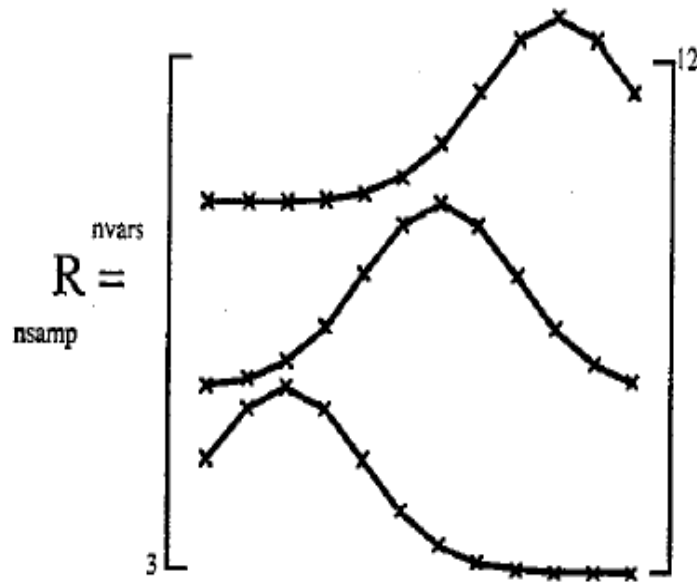
- Vista la capacità dell'occhio umano di cogliere facilmente le somiglianze e le differenze, non c'è da stupirsi che in chimica si usino spesso grafici per presentare e interpretare i dati.
- Ad esempio, spettri o cromatogrammi sono spesso rappresentati come curve continue piuttosto che come tabelle di numeri.
- Infatti, la presenza o assenza di picchi, l'eventuale sovrapposizione e altre informazioni sono colte molto più facilmente in questa forma che non guardando i numeri



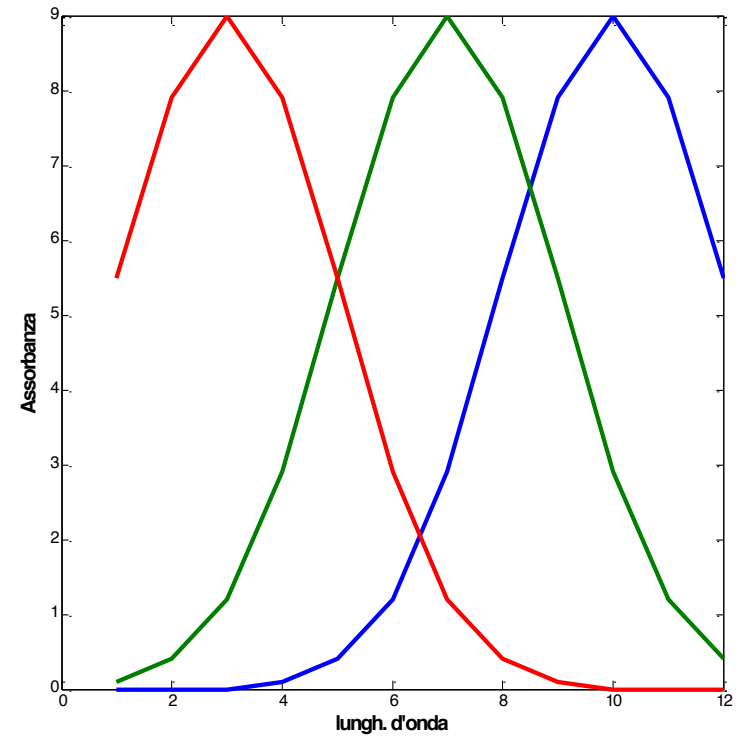
“Vedere” con il computer - 2

- Scopo di questa parte del corso sarà discutere come il computer può aiutare le capacità esplorative umane.
- Per fare questo descriviamo in maggior dettaglio alcuni concetti già parte dell'introduzione generale all'analisi multivariata.
- Immaginiamo di registrare lo spettro di 3 campioni a 12 lunghezze d'onda.
- Per quanto detto in precedenza, questi dati possono essere ordinati in una matrice 3×12 in cui le righe rappresentano gli spettri di ciascun campione e le colonne le assorbanze dei diversi campioni a ciascuna lunghezza d'onda.
- Allo stesso modo possiamo rappresentare gli spettri in maniera grafica.
- Le due rappresentazioni sono equivalenti

Grafici vs tabelle di dati



0.0	0.0	0.0	0.1	0.4	1.2	2.9	5.5	7.9	9.0	7.9	5.5
0.1	0.4	1.2	2.9	5.5	7.9	9.0	7.9	5.5	2.9	1.2	0.4
5.5	7.9	9.0	7.9	5.5	2.9	1.2	0.4	0.1	0.0	0.0	0.0

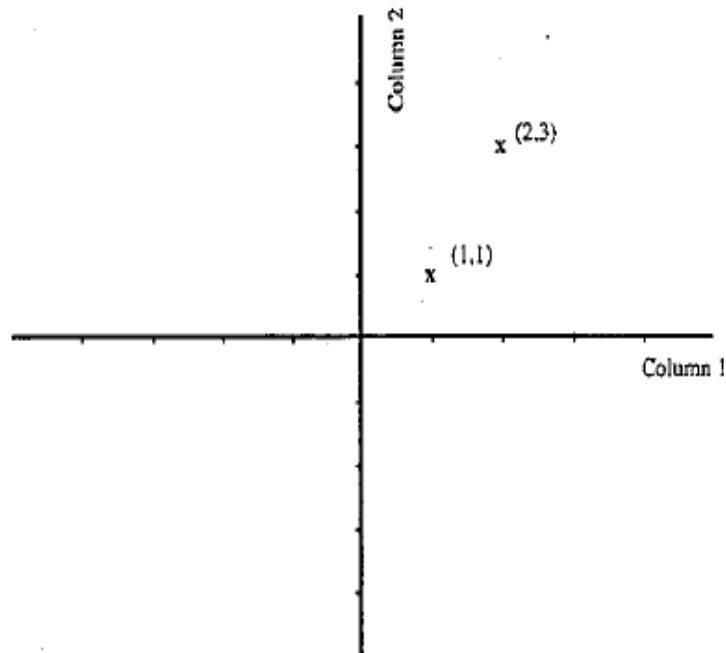


Lo spazio delle righe

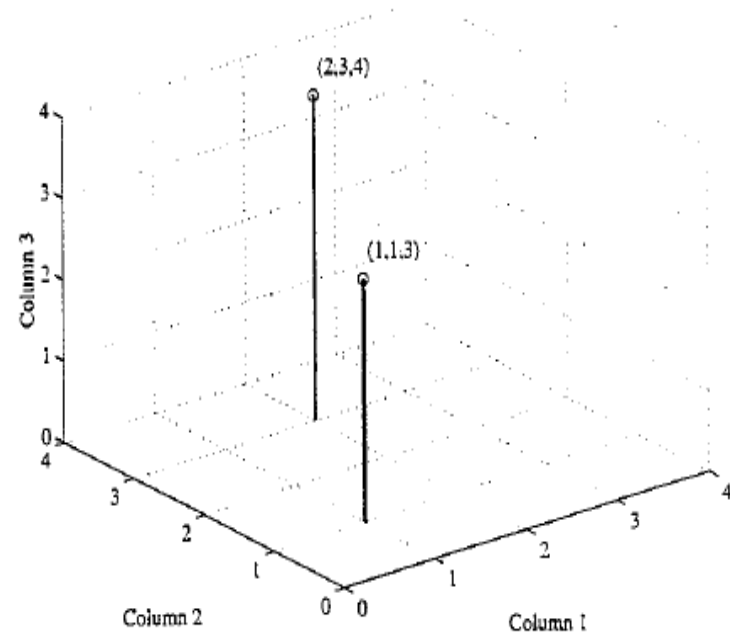
- Nel caso di dati spettrali (o cromatografici) la rappresentazione grafica cui siamo abituati è quella riportata nel lucido precedente, in cui sull'asse x si riportano le variabili e sulla y le intensità (A vs nm)
- Tuttavia, per rappresentare le relazioni tra i campioni può risultare più utile servirsi di un'altra rappresentazione.
- In questo altro tipo di rappresentazione, ogni riga della matrice dei dati è rappresentata come un punto in un sistema di coordinate i cui assi sono definiti dalle colonne.
- Si parla di rappresentazione nello spazio delle righe (**row space**) perché le righe della matrice dei dati (campioni) sono contenute in questo spazio

Spazio delle righe 2 e 3D

$$R = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}$$



$$R = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 1 & 3 \end{bmatrix}$$



Estendendo il concetto

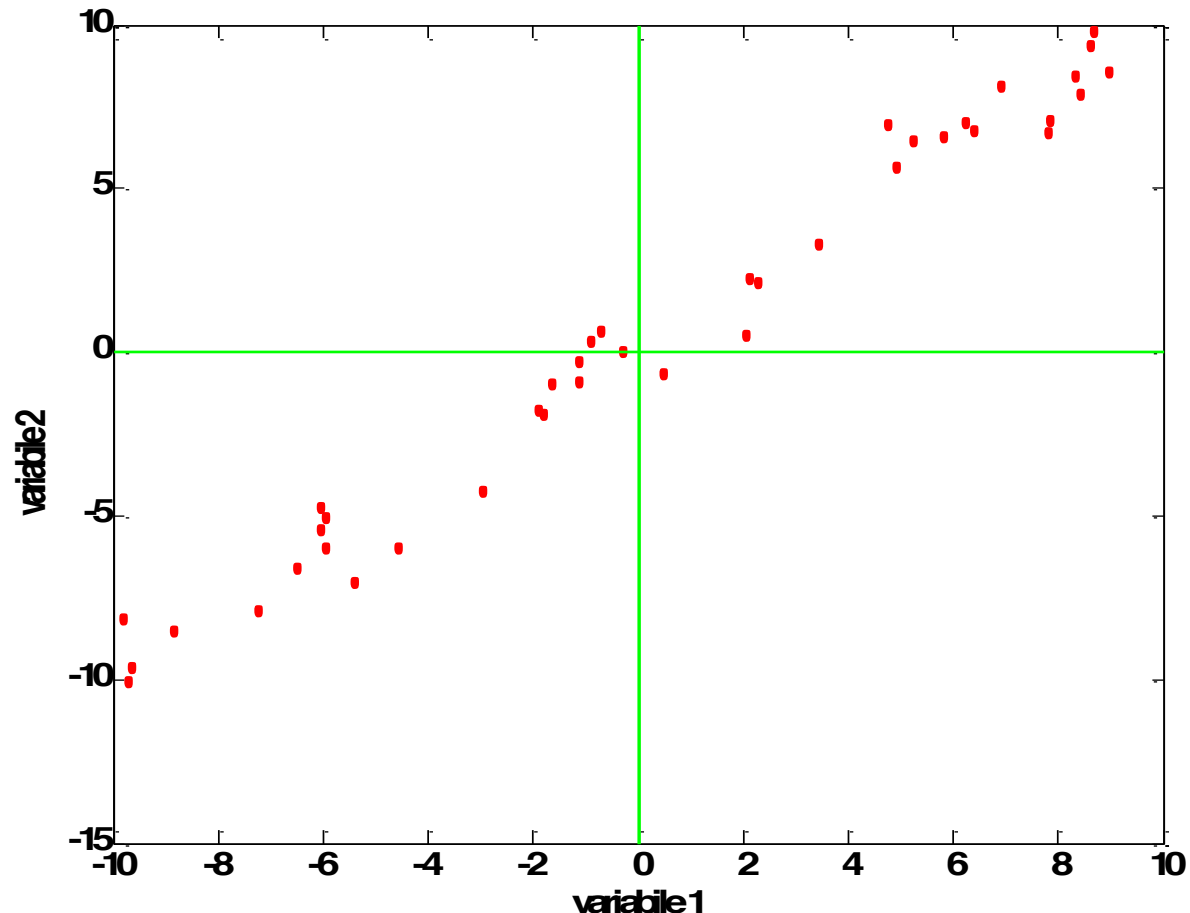
- A differenza dell'esempio, i problemi chimici spesso coinvolgono un gran numero di misure fatte su un sufficiente numero di campioni
- Ad es, un lavoro di spettroscopia applicata può prevedere la determinazione di 800-1000 variabili su almeno 30 campioni
- Generalizzando quanto detto prima, lo spazio di questo problema consiste di una trentina di punti rappresentati in uno spazio a 800 o 1000 dimensioni
- Seppure non sia possibile nella pratica costruire questo grafico, concettualmente non è altro che un estensione dei grafici precedenti.
 - All'aumentare dei campioni aumenta il numero di punti
 - All'aumentare delle variabili misurate aumenta il numero di coordinate

PCA – perché?

- Riassumendo: esaminare lo spazio delle righe di una matrice è una maniera efficace di studiare le relazioni tra i campioni
- Questo è fattibile solo quando il numero delle variabili misurate è minore di 3
- L'Analisi delle Componenti Principali (PCA) è un trattamento matematico della matrice dei dati il cui scopo è rappresentare la variazione presente nelle tante variabili utilizzando un numero molto più piccolo di “fattori” o “componenti principali”
- Si costruisce un nuovo spazio su cui rappresentare i campioni ridefinendo gli assi utilizzando le componenti principali al posto delle variabili originali.
- L'uso di questi nuovi assi – le componenti principali (PC) – permette di rappresentare la vera natura multivariata dei dati in un numero relativamente piccolo di dimensioni e di usare questa rappresentazione per identificare la struttura dei dati stessi

PCA – come funziona?

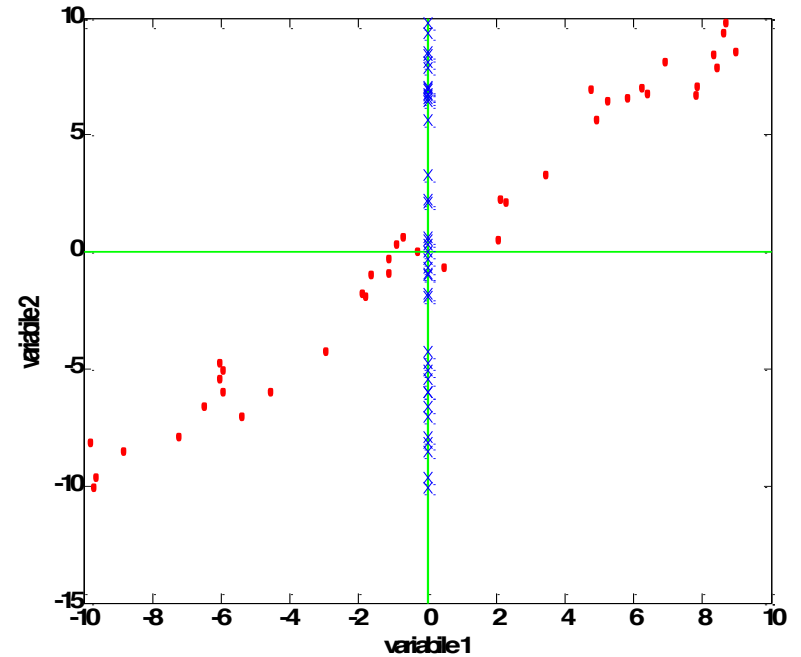
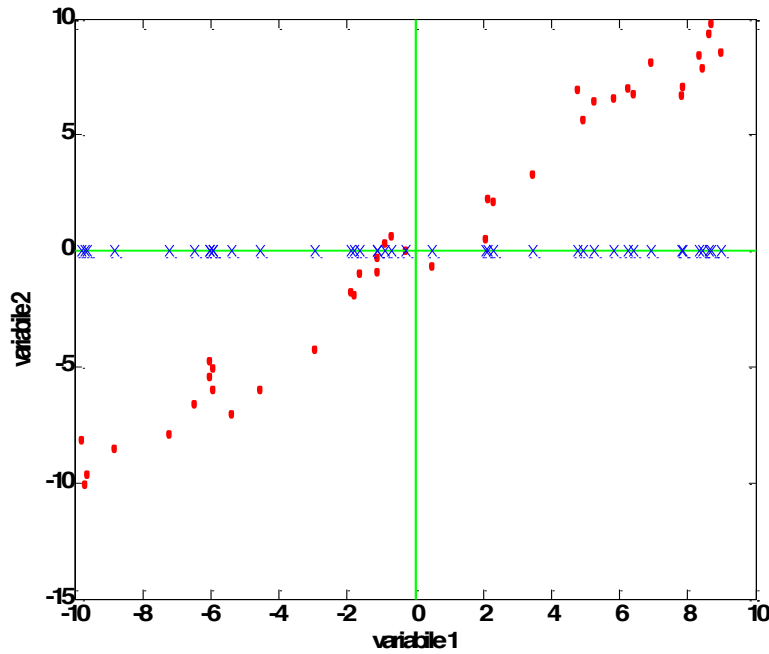
- Il funzionamento della PCA può essere compreso attraverso un esempio 2D
- Immaginiamo di aver misurato il valore di due variabili su 40 campioni



PCA – come funziona? - 2

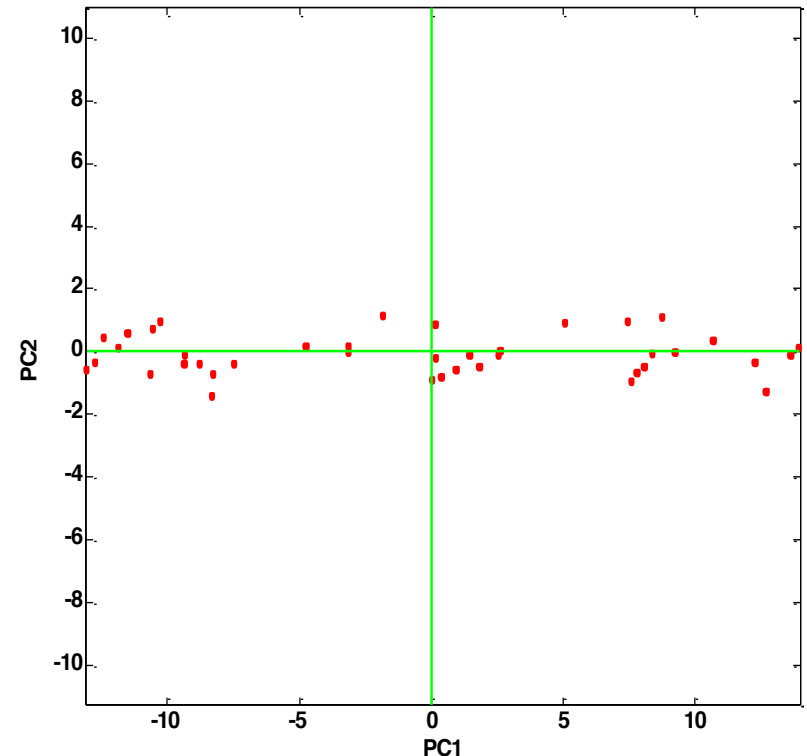
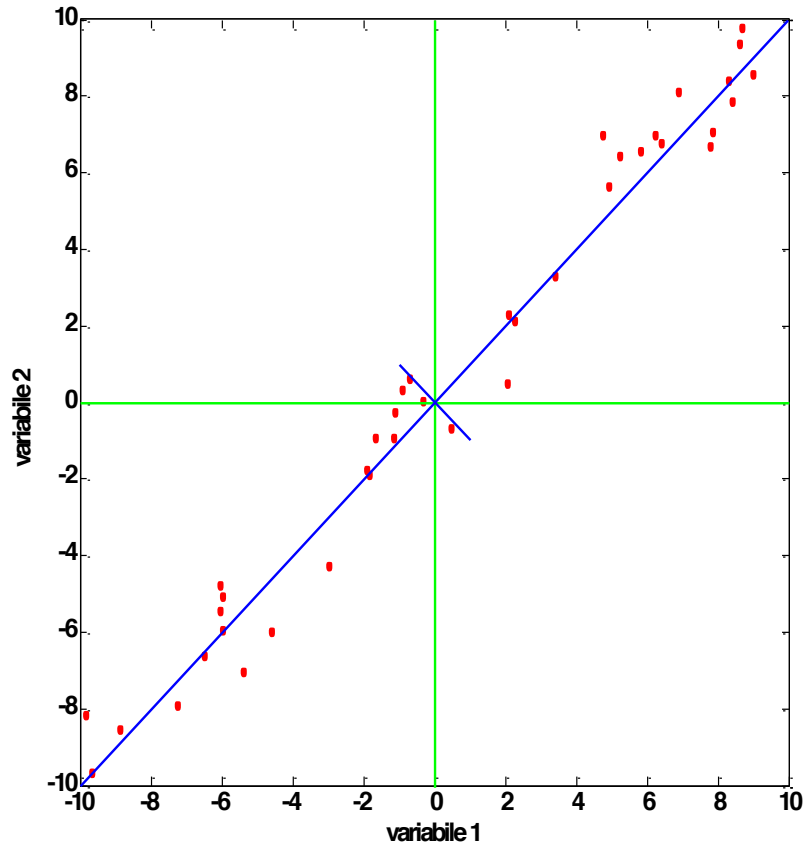
- Siamo interessati a studiare le relazioni tra i campioni nello spazio delle righe
- Le distanze tra i campioni sono usate per definire similarità e differenze
- In termini matematici: scopo della PCA è di descrivere le distanze fra i punti (distribuzione, variabilità) utilizzando il minor numero di dimensioni possibili
- Questo scopo si raggiunge costruendo assi che si allineano coi dati
- Infatti, nessuna delle variabili originali descrive completamente la variabilità all'interno dei dati stessi

Proiettando sulle variabili originali



PCA

- Tuttavia, la prima componente principale è calcolata in maniera tale da descrivere una quantità della variabilità originale, maggiore di quella spiegata da ciascuna delle variabili misurate presa singolarmente



PCA - 2

- La prima componente principale spiega la massima percentuale della variabilità presente nei dati rappresentabile in una sola dimensione
- Messa in un'altra maniera: è la direzione lungo cui si registra la massima dispersione dei dati.
- Inoltre, questa percentuale di variabilità spiegata può essere calcolata attraverso la **varianza**.
- La varianza è infatti un indice della dispersione dei dati lungo una particolare direzione.
- Inoltre, essa è indipendente dal sistema di riferimento: una rotazione degli assi mantiene inalterata la varianza totale all'interno dei dati (somma delle varianze lungo tutte le direzioni, e misura della variabilità presente nel data set).

PCA e varianza

Variabili originali			Componenti principali		
Var.	Varianza	Varianza%	PC	Varianza	Varianza %
1	36.54	48.90%	1	74.59	99.83%
2	38.18	51.10%	2	0.43	0.17%

- Nell'esempio descritto, la prima componente principale cattura praticamente tutta la variabilità presente nei dati (99.83%)
- La seconda descrive la rimanente variazione (0.17%).
- Questa considerazione può essere generalizzata: le componenti principali successive spiegano una sempre minore percentuale della variabilità originale.
- Seguendo questo principio è possibile dire che le ultime componenti principali descrivono principalmente “rumore” ovvero il contributo degli errori di misura o informazioni irrilevanti

PCA e varianza - 2

- Conoscere la percentuale di variabilità spiegata quando si interpretano i grafici delle componenti principali è essenziale
- Ad esempio, se la percentuale di varianza catturata dalle prime due o tre componenti principali è relativamente alta, allora il grafico che si ottiene può essere efficacemente utilizzato per interpretare i dati.
- Se invece le prime due o tre componenti principali rappresentano una percentuale non troppo elevata della variabilità dei dati, le conclusioni che si possono trarre dai dati stessi ne dovranno tenere conto

Costruire le PC

- Come detto, ogni campione può essere descritto da nuove coordinate rispetto allo spazio delle PC.
- Queste coordinate prendono il nome di **scores**
- Dal punto di vista matematico, quanto descritto prima graficamente corrisponde a dire che le componenti principali sono costruite come combinazioni lineari delle variabili originali:

$$t_{i1} = p_{11}x_{i1} + p_{21}x_{i2} + p_{31}x_{i3} + \dots + p_{m1}x_{im} = \mathbf{x}_i\mathbf{p}_1$$

$$t_{i2} = p_{12}x_{i1} + p_{22}x_{i2} + p_{32}x_{i3} + \dots + p_{m2}x_{im} = \mathbf{x}_i\mathbf{p}_2$$

- In queste equazioni t_{i1} e t_{i2} rappresentano rispettivamente le coordinate del campione i-esimo sulla prima e seconda PC.
- \mathbf{x}_i è invece il vettore riga corrispondente alle misure effettuate sul campione i-esimo
- I coefficienti delle combinazioni lineari sono indicati come p_{kl} e sono organizzati nei vettori colonna $\mathbf{p}_1, \mathbf{p}_2, \dots$

Costruire le PC - 2

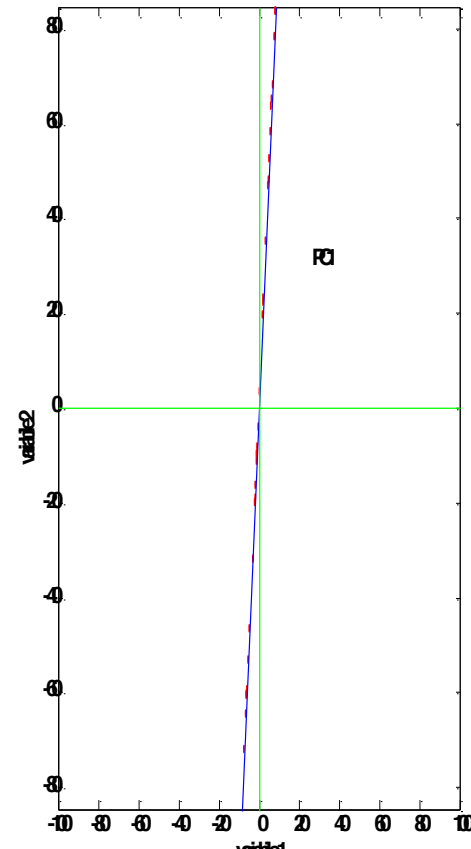
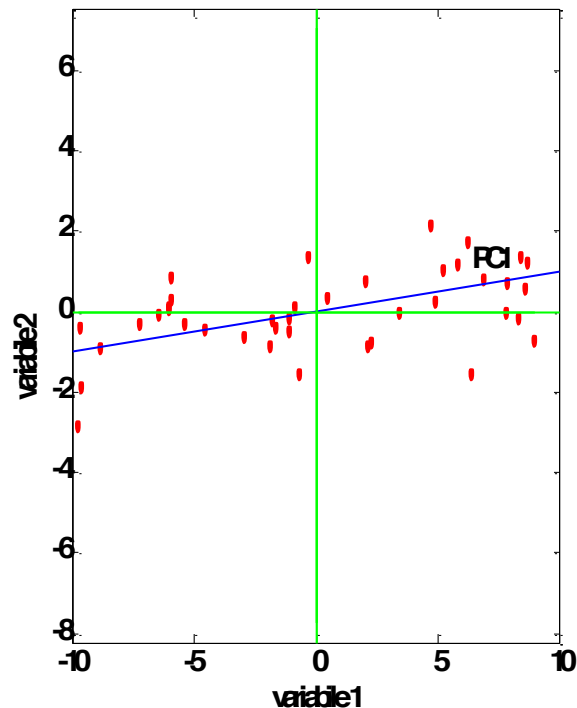
- Le equazioni descritte nella slide precedente possono essere riunite in una rappresentazione in forma di matrice:

$$\underset{n \times f}{\mathbf{T}} = \underset{n \times m}{\mathbf{X}} \underset{m \times f}{\mathbf{P}}$$

- In questo caso \mathbf{T} è la matrice degli scores, ovvero una matrice che racchiude le coordinate dei campioni nello spazio delle componenti principali.
- Ogni riga di \mathbf{T} rappresenta quindi le coordinate di un campione lungo tutte le PC, mentre ogni colonna rappresenta le coordinate di tutti i campioni lungo una particolare PC.
- Allo stesso modo, i coefficienti delle combinazioni lineari che descrivono le componenti principali in termini delle variabili sperimentali sono organizzati nella matrice \mathbf{P} detta dei **loadings**.
- Ogni colonna di \mathbf{P} descrive i coefficienti per una particolare PC.

PC e variabili

- Per poter interpretare le componenti principali, è importante sapere quali variabili contribuiscano di più alla definizione delle PC
- Ad esempio, nel caso a sinistra la prima PC è molto più simile alla variabile 1 che alla 2, mentre accade il contrario nel caso a destra.



PC e variabili - 2

- In termini matematici, il contributo di ciascuna variabile alla PC è il coseno dell'angolo tra le due:
 - Se una PC punta esattamente nella stessa direzione di una variabile, l'angolo tra le due è 0° ed il coseno è 1.
 - Se punta in direzione opposta l'angolo è 180° e il coseno è -1
 - Se la PC è perpendicolare ad una variabile l'angolo che si forma è 90° ed il coseno è 0.
- Questi coseni non sono altro che i loadings descritti in precedenza e raccolti nella matrice **P**.
- Per quanto detto, quindi, i loadings possono variare tra -1 e 1.
- Inoltre, le PC sono costruite in maniera da essere ortonormali
- Questo implica che la somma dei quadrati dei loadings corrispondenti a ciascuna componente principale è unitaria:

$$\sum_{j=1}^m p_{jf}^2 = 1$$

PCA e “rumore”

- Escludere le componenti principali non significative può servire a “filtrare il rumore” presente nei dati
- Infatti per costruzione le prime PC spiegano la maggior parte della variabilità all'interno dei dati
- Il “rumore” sarà quindi concentrato nelle ultime PC
- Non includere queste ultime PC permette di avere dei dati più puliti, con un rapporto ~~segnale~~ ^{segnale}/rumore più alto.
- In linea di principio il massimo numero di PC che può essere calcolato è il minimo tra il numero di righe e il numero di colonne della matrice dei dati (**rango** del problema).
- È possibile però includere nel modello solamente le componenti principali che si ritengono significative ed ottenere una riduzione della dimensionalità del problema

PCA e riduzione di dimensionalità

- La PCA è particolarmente utile quando la dimensionalità dello spazio delle misure è particolarmente elevata (molte colonne) ma i campioni si trovano in un sottospazio di dimensioni significativamente ridotte
- In molti fenomeni chimici la dimensionalità intrinseca del problema è significativamente più piccola del numero di variabili misurate
- Questo perché la dimensionalità intrinseca del problema è legato alle fonti di variabilità in gioco al momento della misura
- Nel linguaggio della PCA, la dimensionalità intrinseca del problema è il numero di PC necessarie per spiegare la variabilità non legata al rumore
- Uno degli obiettivi della PCA quindi è quello di determinare il numero di componenti principali significative

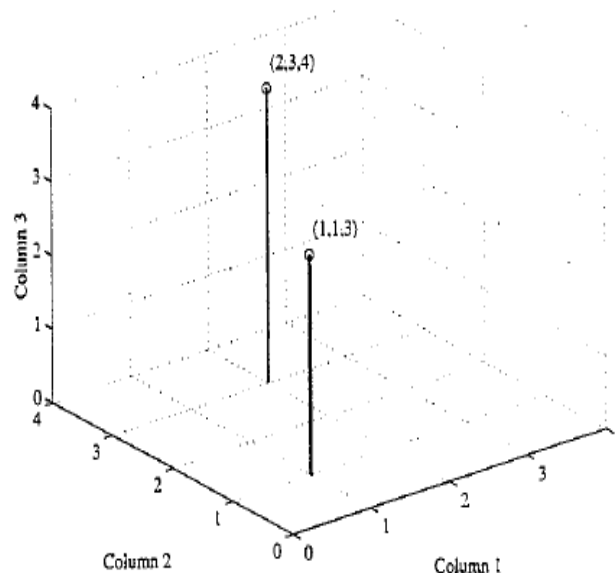
PCA e riduzione delle variabili - 2

- Questo è equivalente a dire che si vuole determinare la quantità di informazione rilevante contenuta nel data set
- Tuttavia, questa quantità è strettamente dipendente dal problema che si sta studiando
 - Se la percentuale di variabilità legata al rumore è lo 0.1%, spiegare il 99% della varianza lascia da parte una porzione di informazione
 - Lo stesso 99% diventa un valore troppo elevato se l'effetto del rumore corrisponde al 10% della variabilità totale
- Scegliere il numero di componenti principali opportuno è importante per visualizzare solo l'informazione rilevante
- Includere una quantità troppo elevata di “rumore” considerando troppe PC o escludere informazione rilevante includendone poche può avere effetti negativi sull'interpretazione

Fattori che limitano la dimensionalità

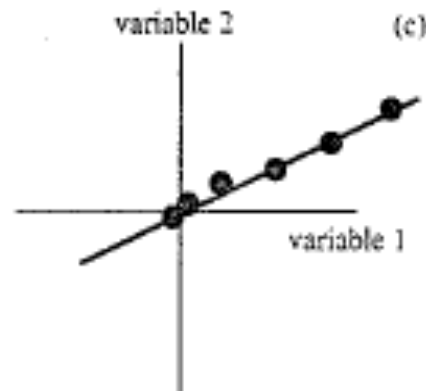
- Alcuni vincoli possono limitare la dimensionalità intrinseca di un set di dati
- Se ripensiamo all'esempio visto in precedenza, se si analizzano solo due campioni, (anche se lo spazio è 3D) questi non possono occupare più di due dimensioni
- In realtà, visto che per due punti passa una retta, il problema è 1D

$$R = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 1 & 3 \end{bmatrix}$$



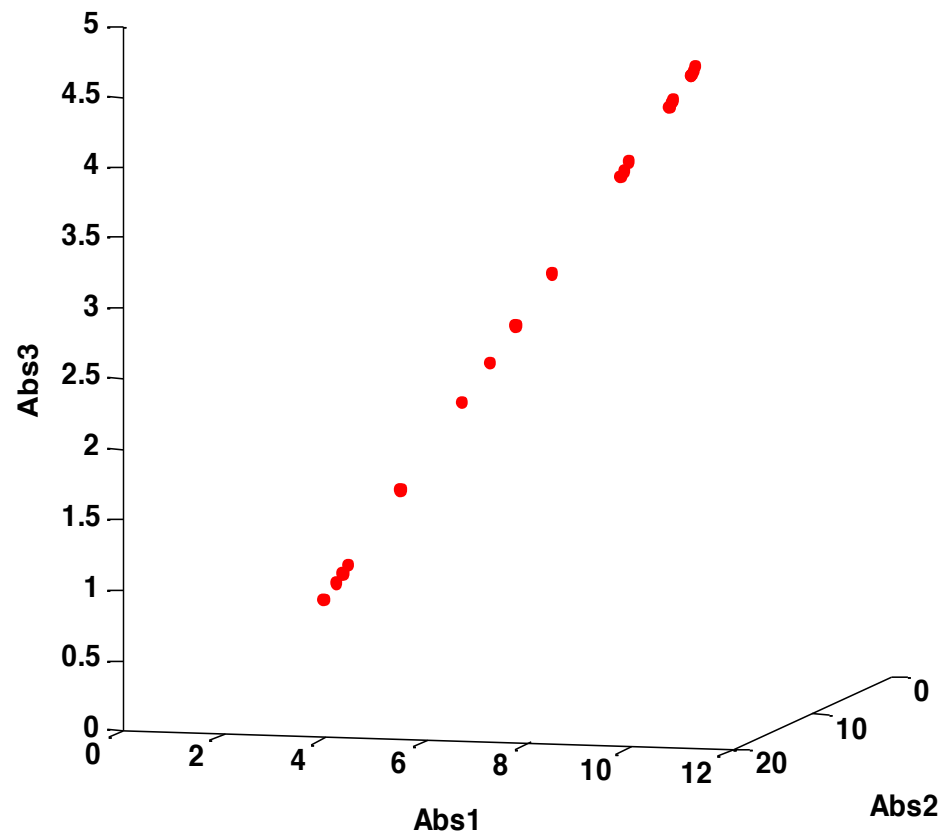
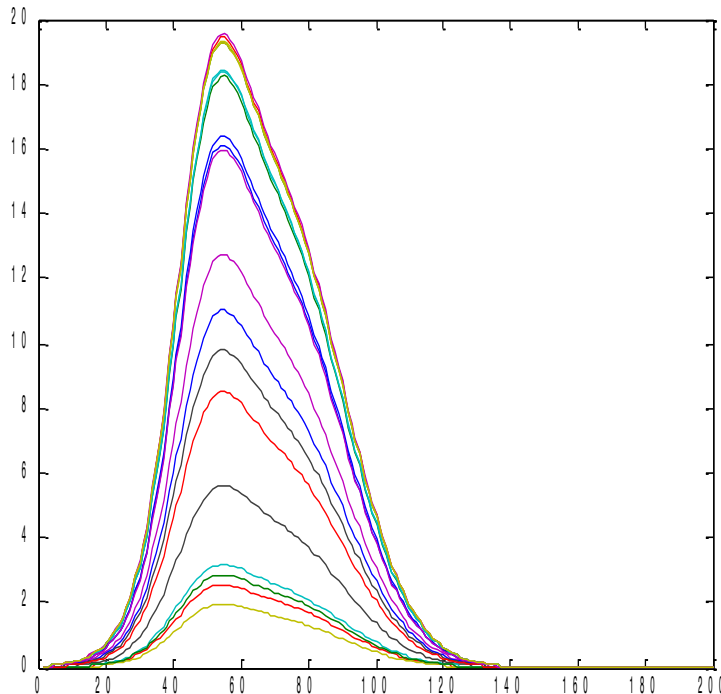
Fattori che limitano la dimensionalità - 2

- Il numero di campioni può essere quindi un fattore limitante la dimensionalità intrinseca quando questo sia minore del numero delle variabili misurate
- Anche il numero delle variabili misurate costituisce un limite: la dimensionalità intrinseca del problema non può eccedere la dimensionalità “misurata”
- Nella chemiometria, comunque, in genere quello che limita la dimensionalità del problema è la chimica
- Immaginiamo di misurare due variabili su un campione in cui sia presente una sola specie chimica:



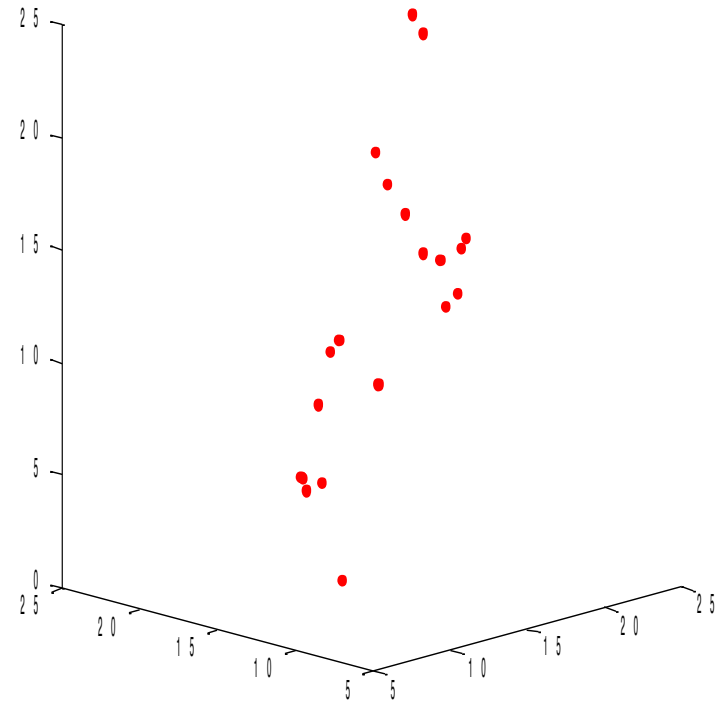
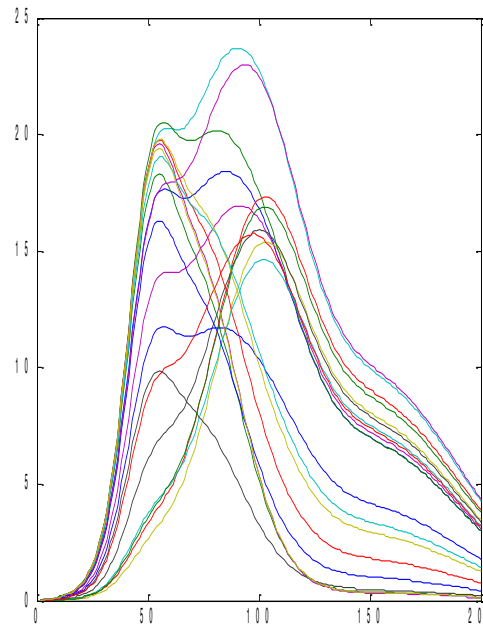
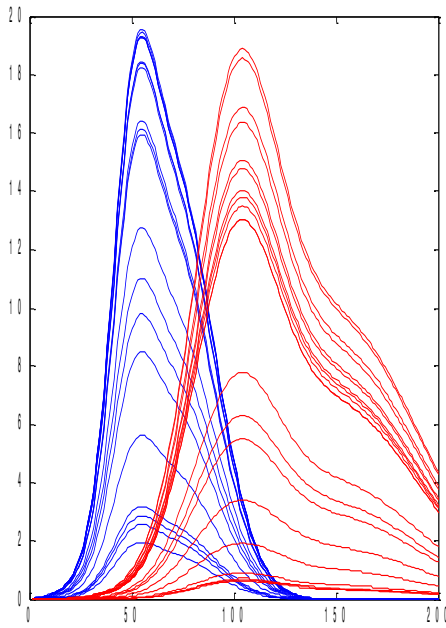
Fattori che limitano la dimensionalità - 3

- Il fatto che si misuri un sistema in cui una sola specie è “responsabile” di tutti i segnali che si registrano fa sì che questi segnali siano completamente correlati e che la dimensionalità del problema sia 1.
- Se si aumenta il numero di variabili misurate (ad es spettro) la situazione non cambia.



Fattori che limitano la variabilità - 4

- Analogamente, nel caso sia presente una seconda specie chimica, la dimensionalità del problema sarà due, perché si avranno solamente due fonti di variabilità all'interno dei dati



Fattori che influenzano la variabilità – 5

- Queste considerazioni possono essere generalizzate ad includere ogni fonte di variabilità
- Se i dati non fossero affetti da errore, la scelta del numero di componenti principali rilevanti non sarebbe un problema
- Infatti, in assenza di errore negli esempi precedenti la variabilità spiegata con 1 o 2 dimensioni sarebbe sempre del 100%.
- La presenza di errore all'interno dei dati costituisce un ulteriore fonte di variabilità che rende il **rango sperimentale** del problema (il numero di PC necessarie a spiegare il 100% della variabilità all'interno dei dati) maggiore del **rango chimico** (numero di fonti di variabilità rilevanti/informative)

PCA in pratica – 1

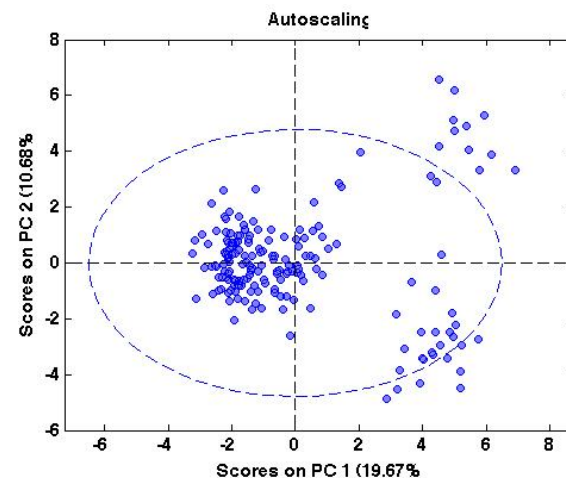
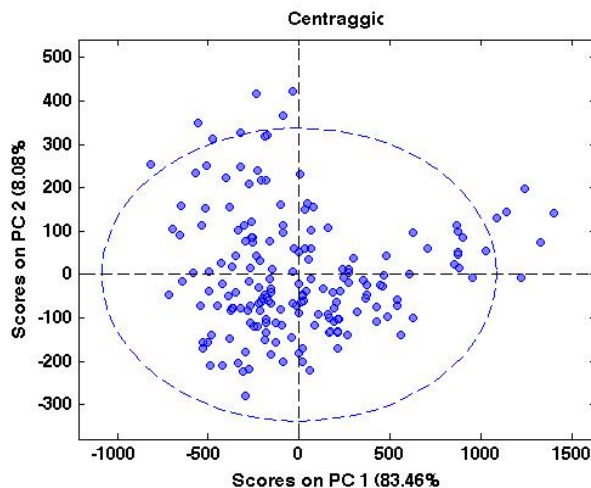
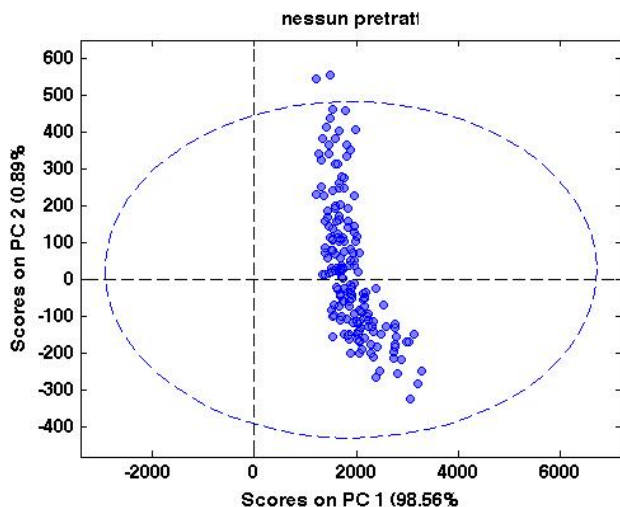
- Una volta visto il retroterra teorico dell'analisi delle componenti principali, vediamone l'utilità con un esempio.
- Per far questo ci serviamo di un data set costituito dalle analisi chimiche effettuate su campioni di vini DOC.
- In particolare sono stati considerati 180 campioni provenienti da 7 differenti denominazioni (Chianti, Pinerolese, Sagrantino, Montepulciano, Nero d'Avola, Solopaca, Terrano).

PCA in pratica - 2

- Su ciascun vino sono state effettuate 35 misure; Grado alcolico (% v/v), acidità totale, SO₂ (mg/L), Cu (mg/L), Zn (mg/L), Pb (ppb), polifenoli total (mg/L), acido gallico (mg/L), acido protocatechico (mg/L), tirosolo (mg/L), acido vanillico (mg/L), acido siringtonico (mg/L), acido caffeico (mg/L), acido ferulico (mg/L), acido p-coumarico (mg/L), procianidina B1 (mg/L), procianidina B2 (mg/L), (+)-catechina (mg/L), (-)-epicatechina (mg/L), etilgallato (mg/L), rutina (mg/L), isoquercetina (mg/L), isoramnetina-3-O-glucoside (mg/L), kaempferolo-3-O-glucoside (mg/L), miricetina (mg/L), quercetina (mg/L), kaempferolo (mg/L), isoramnetina (mg/L), ramnetina (mg/L), trans-resveratrolo (mg/L), cis-resveratrolo (mg/L), trans-piceide (mg/L), cis-piceide (mg/L), prolina (mg/L), antociani totali (mg/L).

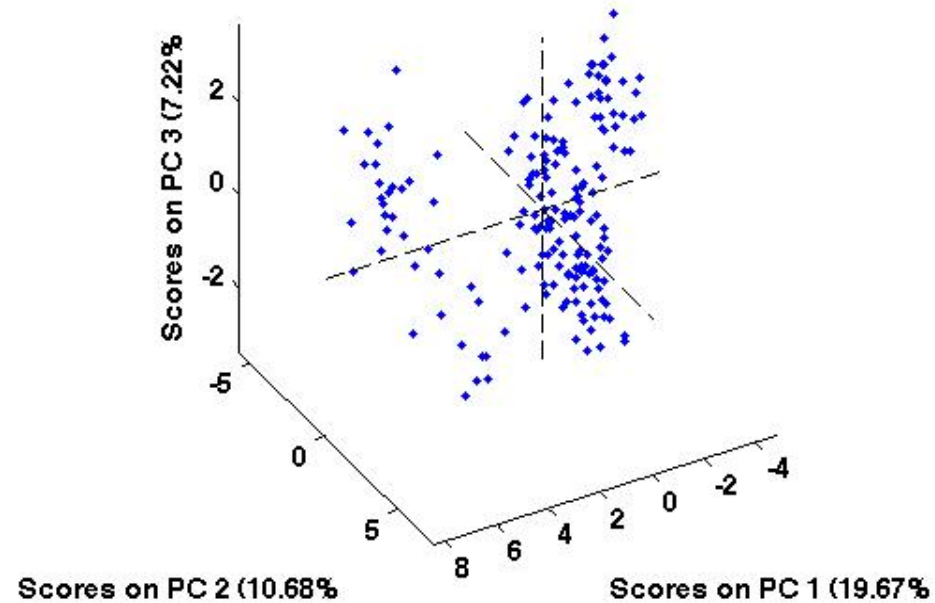
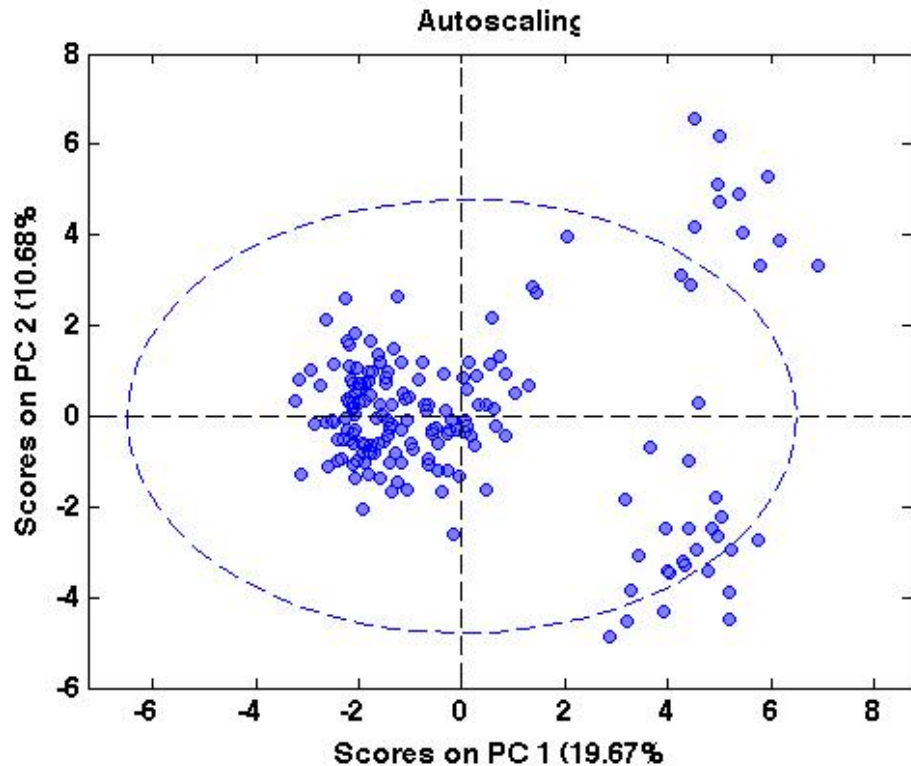
Pretrattamento

- Le variabili sono di natura differente. Per quanto detto in precedenza, il pretrattamento più opportuno dovrebbe essere l'autoscaling.



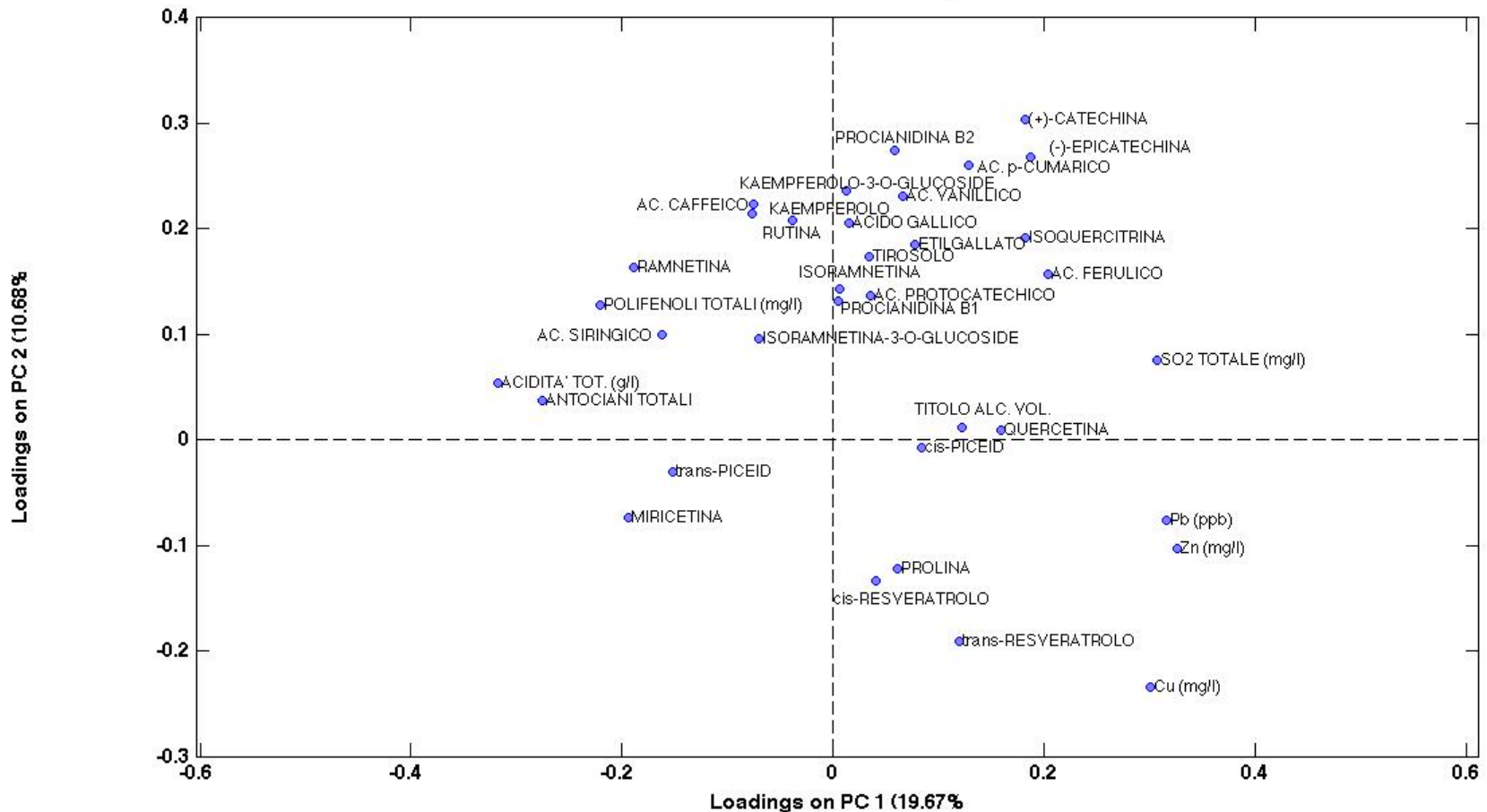
- Se si va a vedere l'effetto dei diversi pretrattamenti si ha una prima conferma del fatto che l'autoscaling sia effettivamente il più adatto.

PCA - scores



- Il grafico sulle prime tre componenti principali mostra l'esistenza di alcuni gruppi di campioni all'interno dei dati.
- Se ne identificano chiaramente 3 sulle prime 2 PC e un quarto in tre dimensioni.

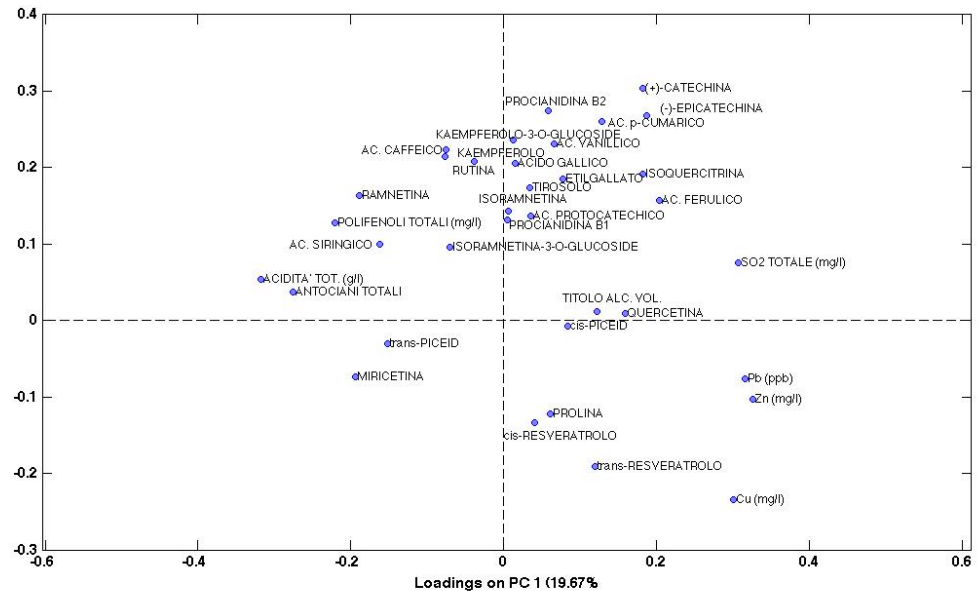
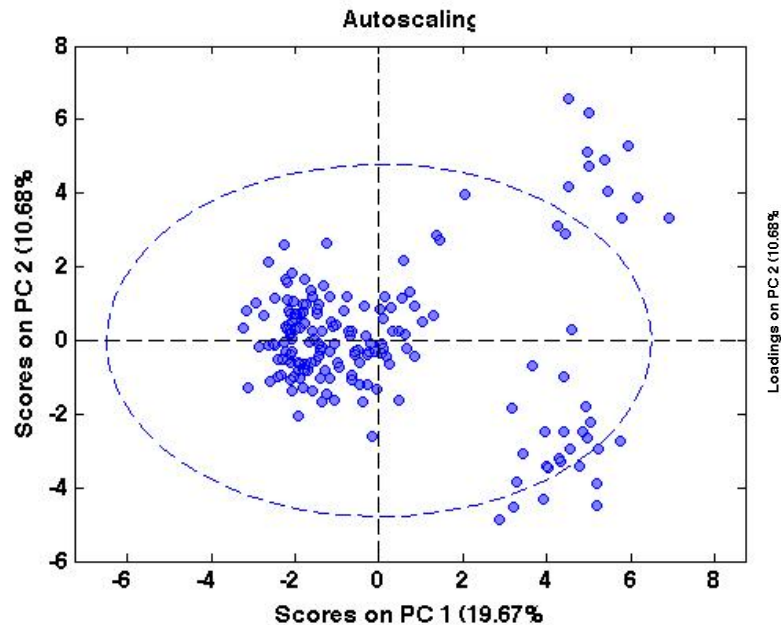
PCA - loadings



- Come detto l'analisi dei loadings sulle componenti principali ci permette di determinare il contributo delle variabili originali al modello PC.

PCA: interpretazione

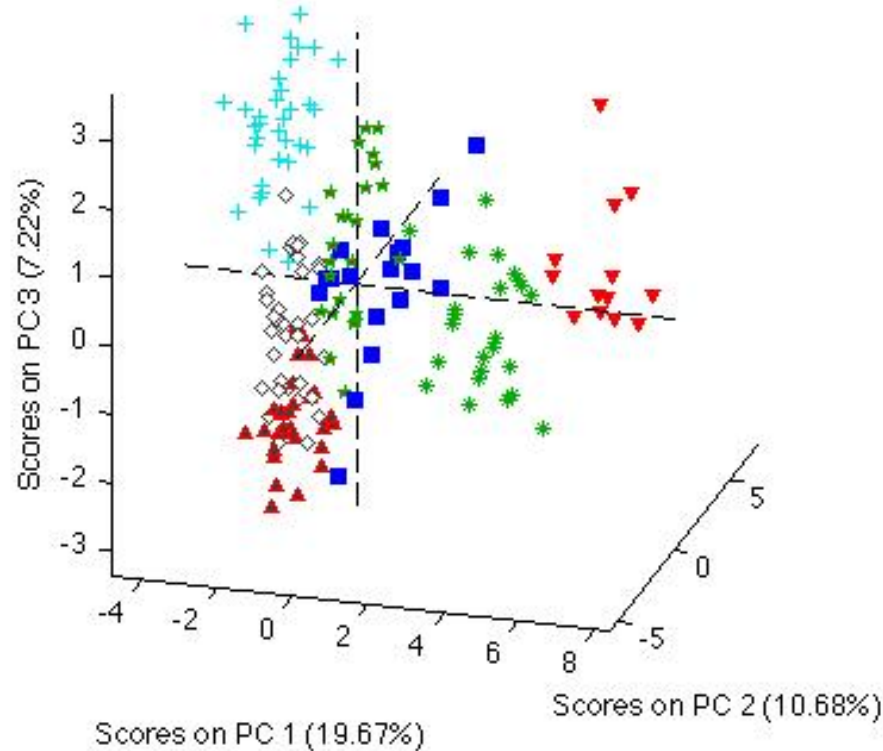
- Dal confronto tra il grafico degli scores e quello dei loadings si può procedere per l'interpretazione dei risultati:



- Ad es: i vini del gruppo in basso a destra si caratterizzano per un più elevato contenuto in metalli e in resveratrolo. I due gruppi a destra inoltre hanno un grado alcolico più elevato.

PCA e informazione aggiuntiva

- L'interpretazione dei dati può essere arricchita se si disponga di informazioni aggiuntive. In questo caso, sul tipo di DOC dei campioni:

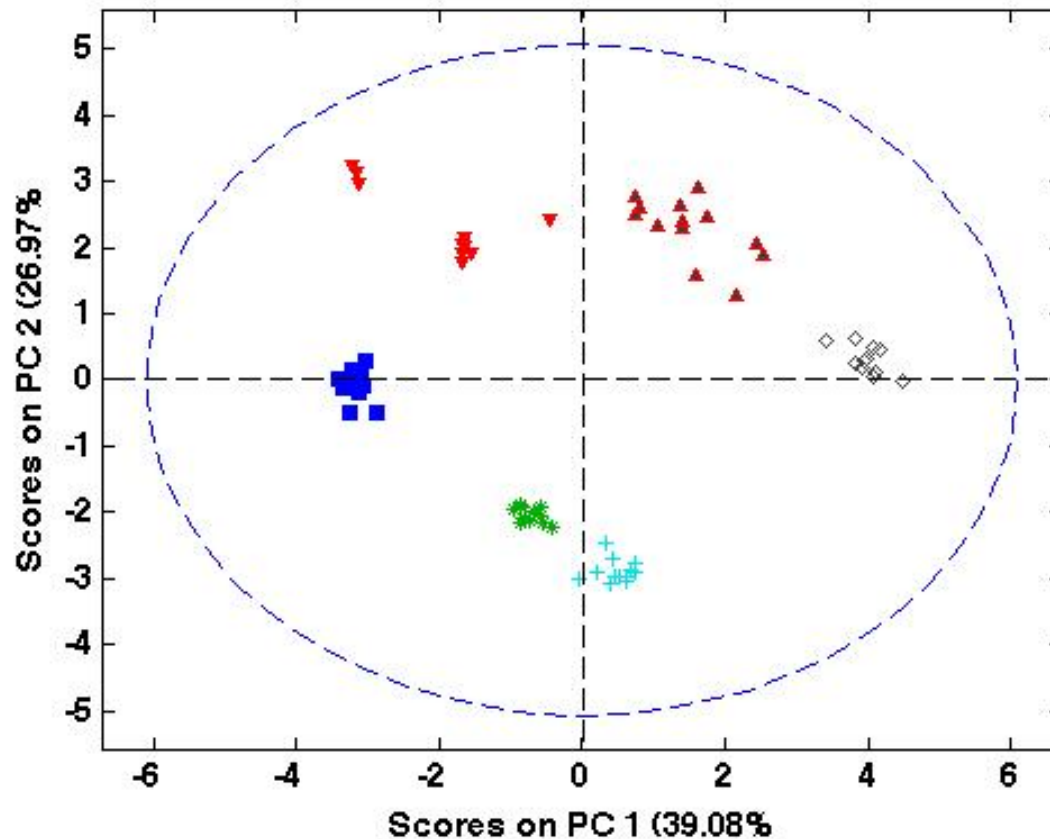


PCA: miele

- Utilizziamo un altro data set alimentare per presentare un altro esempio di applicazione dell'analisi delle componenti principali.
- 73 campioni di miele di diversa provenienza botanica (in particolare 6 origini; melata, millefiori, eucalipto, sulla, erica e castagno).
- 15 variabili misurate su ciascun campione
- Anche in questo caso, dato che le variabili sono di natura differente, è opportuno scegliere l'autoscaling come metodo di pretrattamento

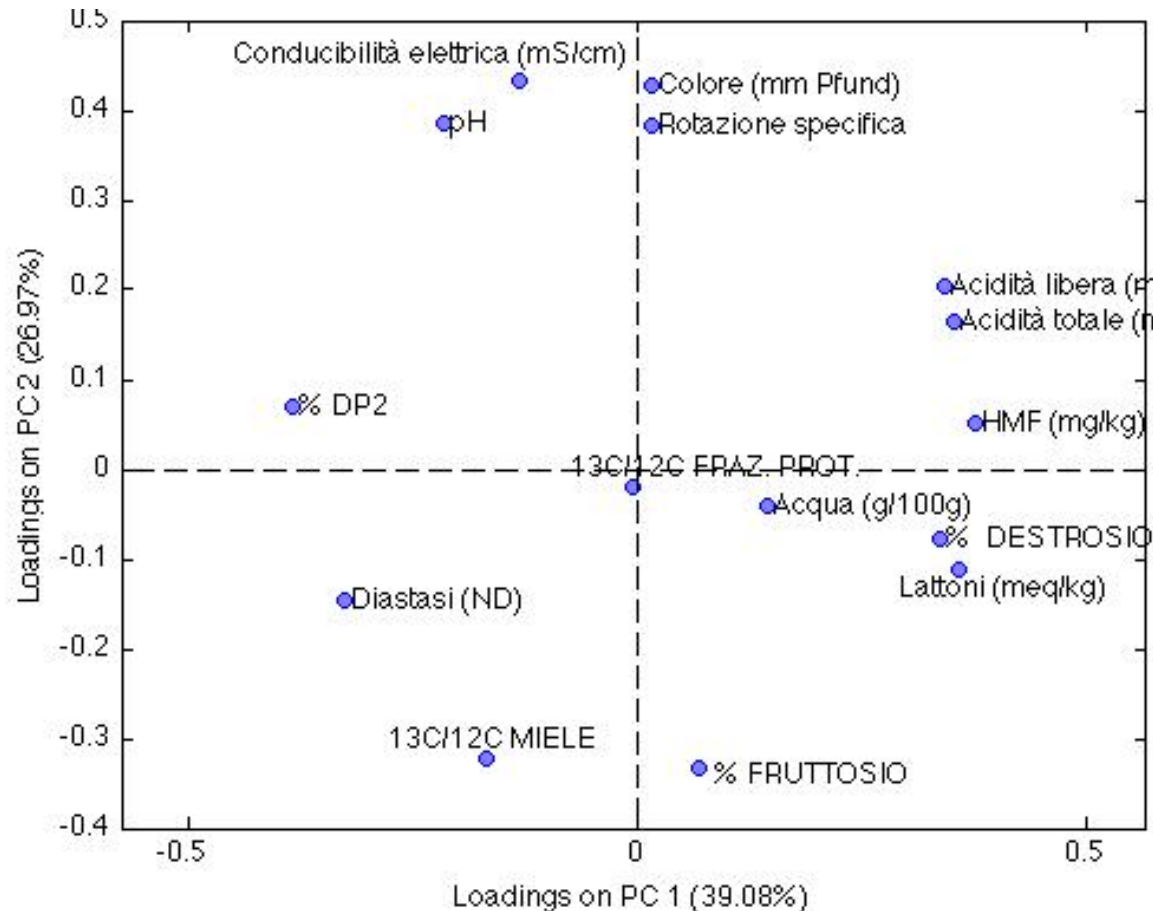
Miele: scores

- Il grafico degli scores sulle prime due componenti principali permette già di identificare la presenza di tutti e 6 i diversi tipi di miele.



Miele: loadings

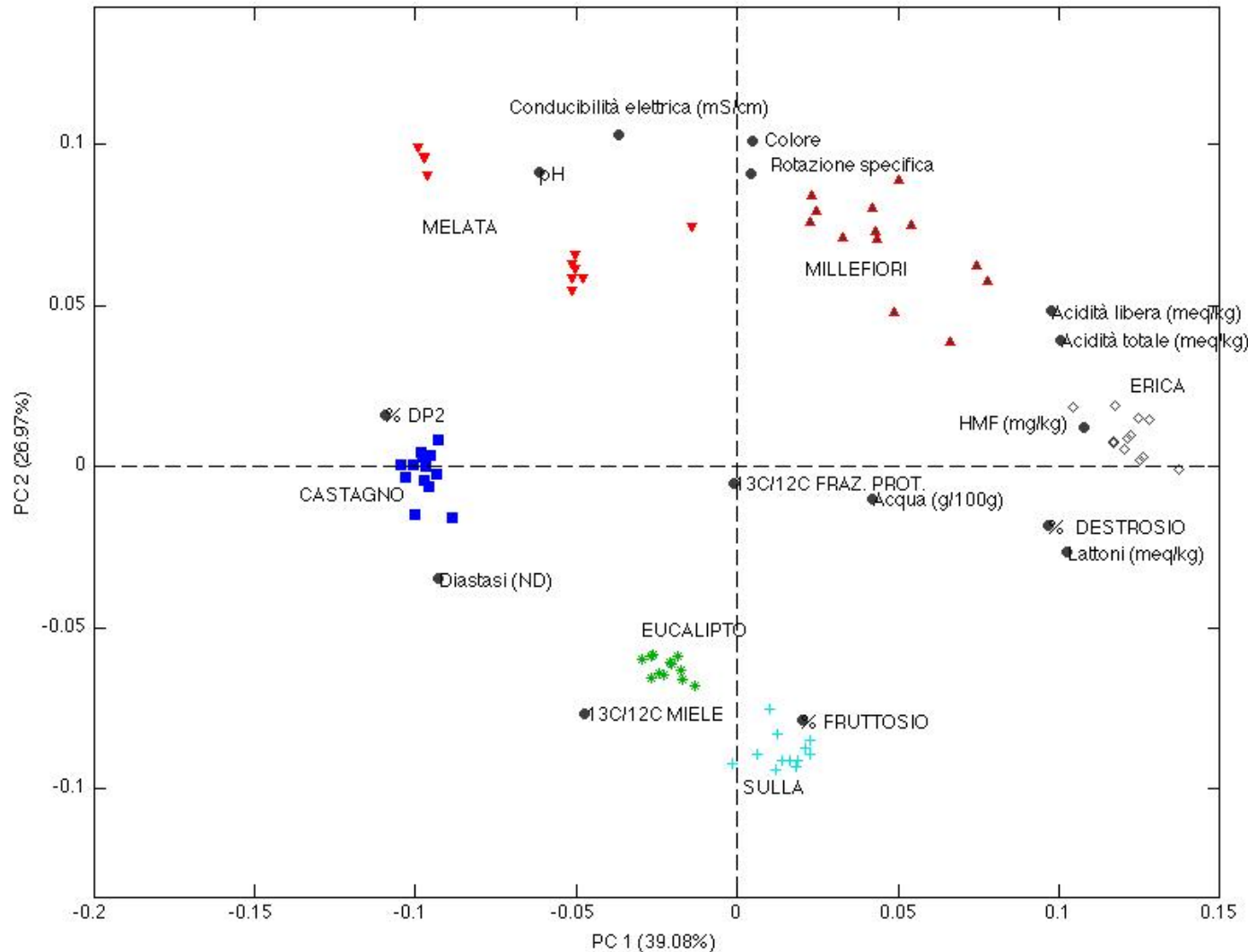
- Anche in questo caso, l'interpretazione dei risultati avviene attraverso l'analisi dei loadings:



- Ad esempio, la variabile C-13/C-12 nella frazione proteica, non contribuisce a spiegare alcuna variabilità (loading ca. 0)

Biplot

- Per semplificare l'interpretazione, i grafici degli scores e dei loadings possono essere riuniti insieme in un grafico detto biplot



PC: quante?

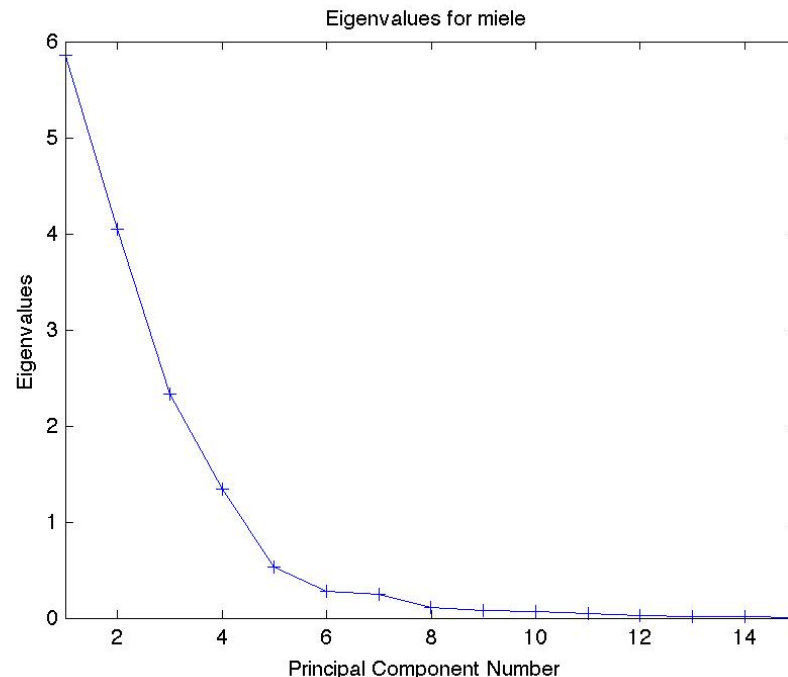
- Finora abbiamo visto le componenti principali come metodo di analisi esplorativa principalmente basato sulla rappresentazione grafica
- Tuttavia le componenti principali rappresentano un modello dei dati e affinché questo modello sia accurato è necessario conoscere quante PC spieghino l'informazione e quali siano quelle che non sono significative.
- Per fare questo esistono diversi metodi, che spesso – basandosi su concetti differenti – danno come risultato numeri non sempre coincidenti.
- Noi vedremo due di quelli più semplici e, in seguito, un terzo:
 - Criterio dell'autovalore medio
 - Criterio della percentuale di varianza spiegata
 - Cross-validation

Criterio dell'autovalore medio

- Un primo criterio di scelta del numero di componenti principali si basa sul concetto di varianza.
- Abbiamo detto come la varianza rappresenti un indice della variabilità catturata dalle componenti principali.
- Nel linguaggio delle PC, la varianza lungo una particolare componente prende anche il nome di autovalore (I)
- Il primo criterio che consideriamo seleziona come significative tutte quelle PC che hanno un autovalore maggiore dell'autovalore medio.
- Questo significa considerare come significative tutte quelle PC che spiegano una percentuale di variabilità maggiore di quanto, in media, ne spieghi ciascuna delle variabili sperimentali

Criterio dell'autovalore medio - 2

- Se i dati vengono (auto)scalati, ciascuna variabile originale viene trasformata in modo da avere varianza unitaria.
- Di conseguenza, in quei casi il criterio si trasforma in: sono considerate come significative tutte quelle PC il cui autovalore sia maggiore di 1.
- Ad esempio nel caso del miele questo significherebbe dire che il modello deve includere solo le prime 4 PC:

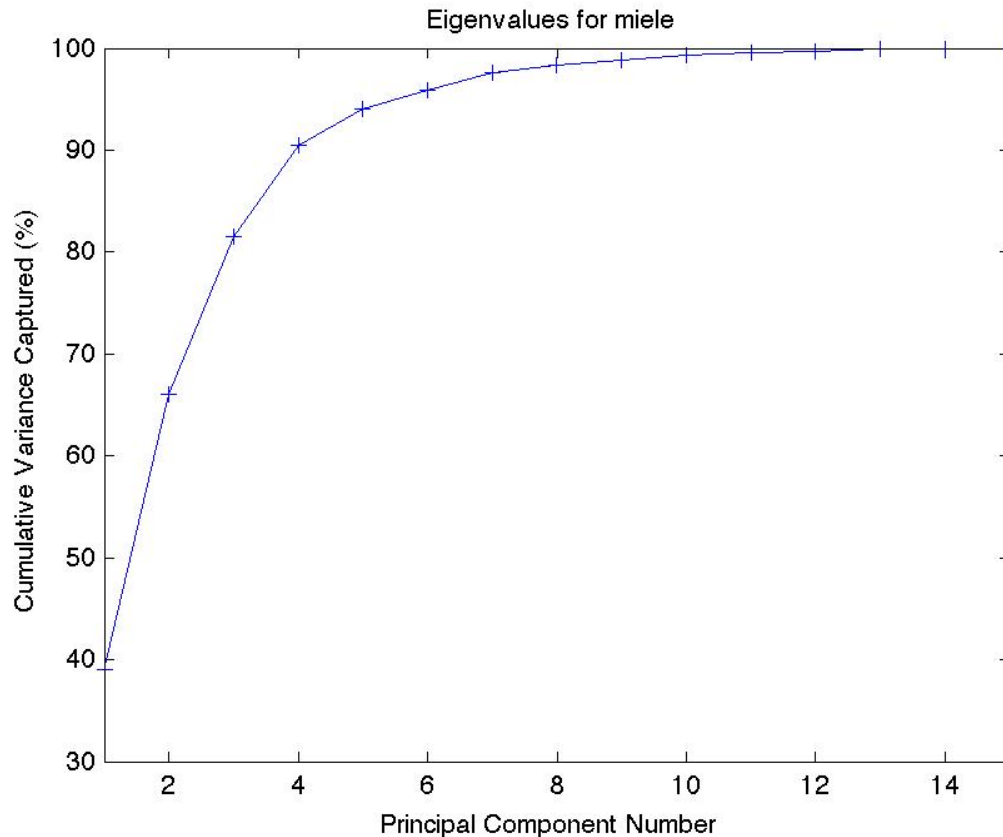


Criterio della percentuale di varianza spiegata

- In questo criterio, che si basa sempre sulla varianza, si include nel modello un numero di componenti principali sufficienti a spiegare almeno una certa percentuale della variabilità presente nei dati (ad es. l'80%, il 90%...)
- Questo criterio ha il vantaggio di essere semplice ma lo svantaggio di essere particolarmente arbitrario (abbiamo infatti visto come la varianza totale spiegata non sempre sia un buon indice di significatività delle PC).

Percentuale di varianza spiegata - 2

- Nel caso del data set del miele, si dovrebbero includere:
 - 3 PC per spiegare almeno l'80% della varianza
 - 4 PC per spiegare almeno il 90% della varianza



PCA come modello

- Abbiamo visto come la PCA costituisca una rappresentazione alternativa dei dati in un altro sistema di coordinate.

$$\underset{n \times f}{\mathbf{T}} = \underset{n \times m}{\mathbf{X}} \underset{m \times f}{\mathbf{P}}$$

- Se il numero di componenti principali che si considerano è il massimo possibile, le due rappresentazioni sono perfettamente equivalenti, e dalla rappresentazione in componenti principali si può ritornare ai dati originali riottenendo la matrice di partenza:

$$\mathbf{X} = \mathbf{TP}^T$$

- Tuttavia, nella maggior parte delle applicazioni della PCA, il numero di dimensioni dello spazio delle PC è significativamente minore del numero delle dimensioni originali.
- La rappresentazione in componenti principali costituisce quindi un'approssimazione dei dati stessi
- In particolare costituisce la migliore approssimazione f -dimensionale (se f è il numero di PC)

PCA come modello - 2

- Anche a partire da questa rappresentazione è possibile ritornare indietro alla matrice dei dati, solo che questa volta, la trasformazione inversa non sarà esatta:

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^T \quad \mathbf{X} - \hat{\mathbf{X}} = \mathbf{E}$$

- La matrice \mathbf{E} raccoglie le differenze tra i dati misurati e i valori approssimati secondo il modello PCA (**scarti** o **residui**)
- Per come sono costruite le PC, è possibile ottenere un'approssimazione dei dati originali con qualsiasi numero di

PC:

$$\hat{\mathbf{X}}(1) = \mathbf{t}_1\mathbf{p}_1^T$$

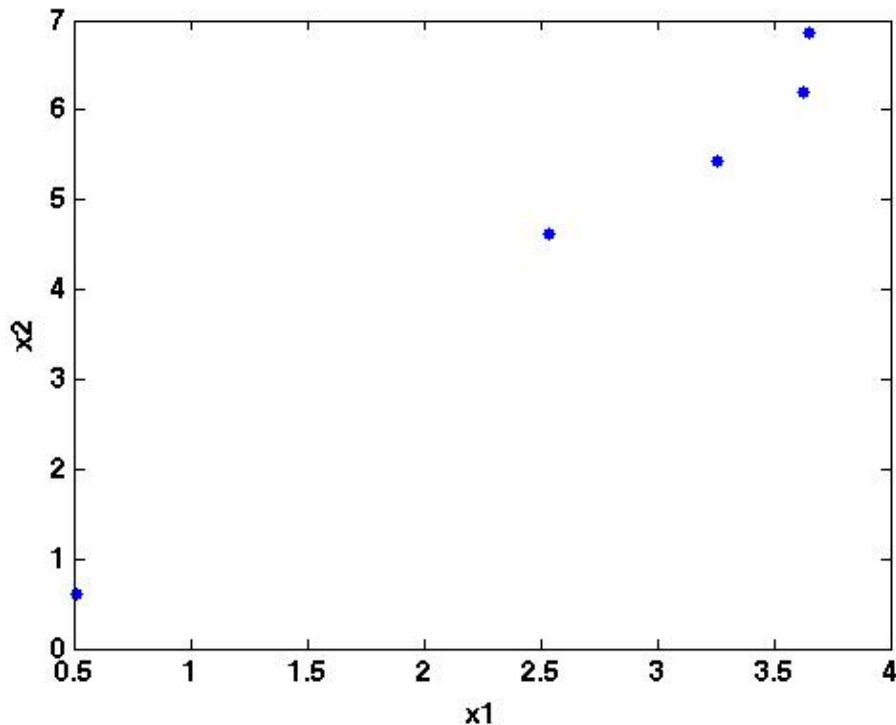
$$\hat{\mathbf{X}}(2) = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T$$

\vdots

$$\hat{\mathbf{X}}(k) = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \dots + \mathbf{t}_k\mathbf{p}_k^T$$

PCA come modello - 3

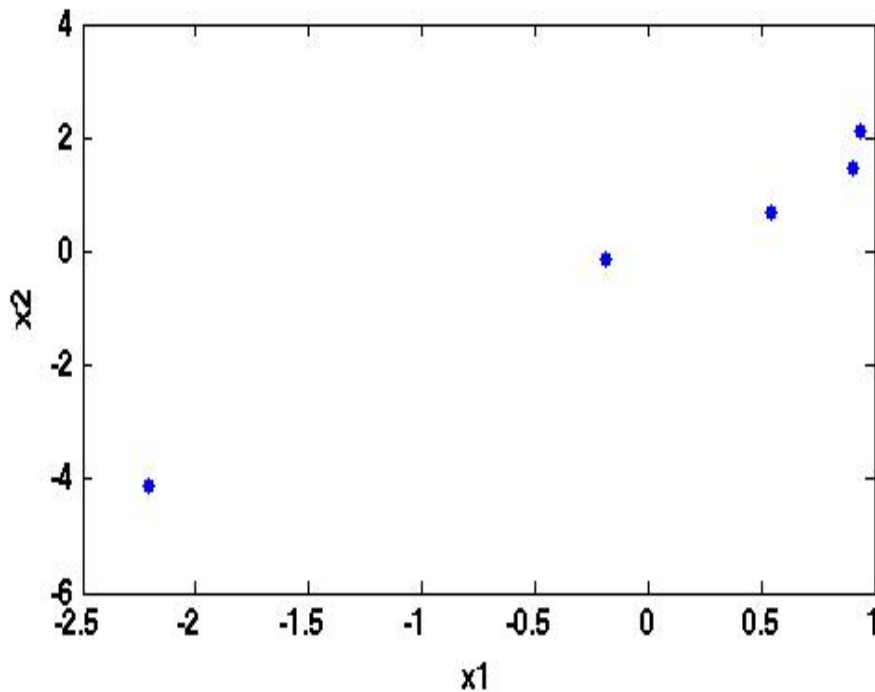
- Per rappresentare quanto detto consideriamo l'esempio in due dimensioni:



$X = \begin{bmatrix} 3.2589 & 5.4358 \\ 3.6232 & 6.1967 \\ 0.5079 & 0.6118 \\ 3.6535 & 6.8706 \\ 2.5294 & 4.6258 \end{bmatrix}$

PCA come modello - 4

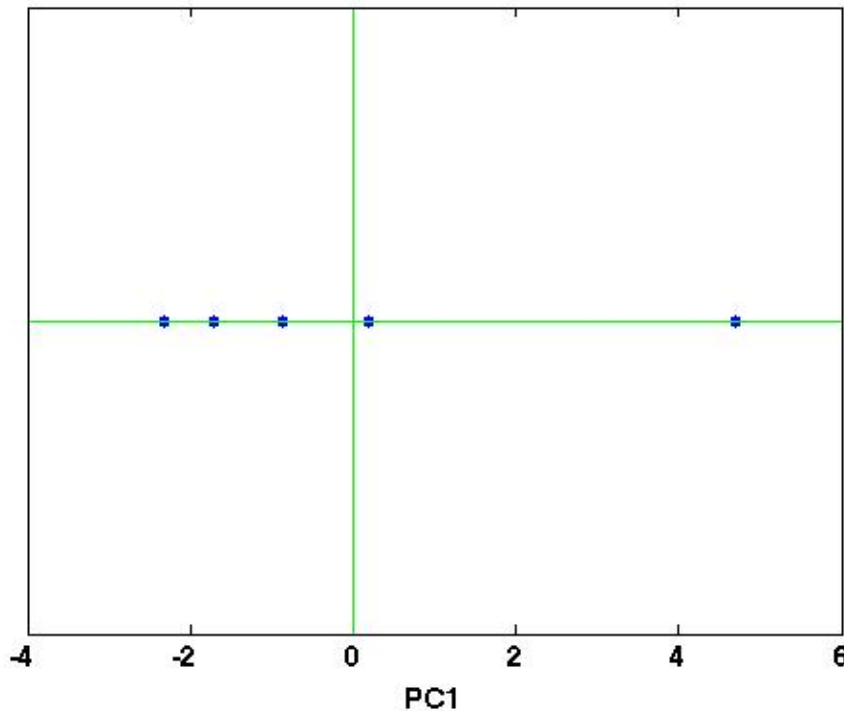
- Per quanto detto in precedenza, prima di procedere con la PCA centriamo i dati, sottraendo a ciascuna variabile la media :



$X_c = \begin{bmatrix} 0.5443 & 0.6877 \\ 0.9086 & 1.4486 \\ -2.2066 & -4.1364 \\ 0.9389 & 2.2125 \\ -0.1852 & -0.1224 \end{bmatrix}$

PCA come modello - 5

- Su questa matrice procediamo con l'analisi delle componenti principali.
- La prima componente principale è caratterizzata da questi valori degli scores e loadings:

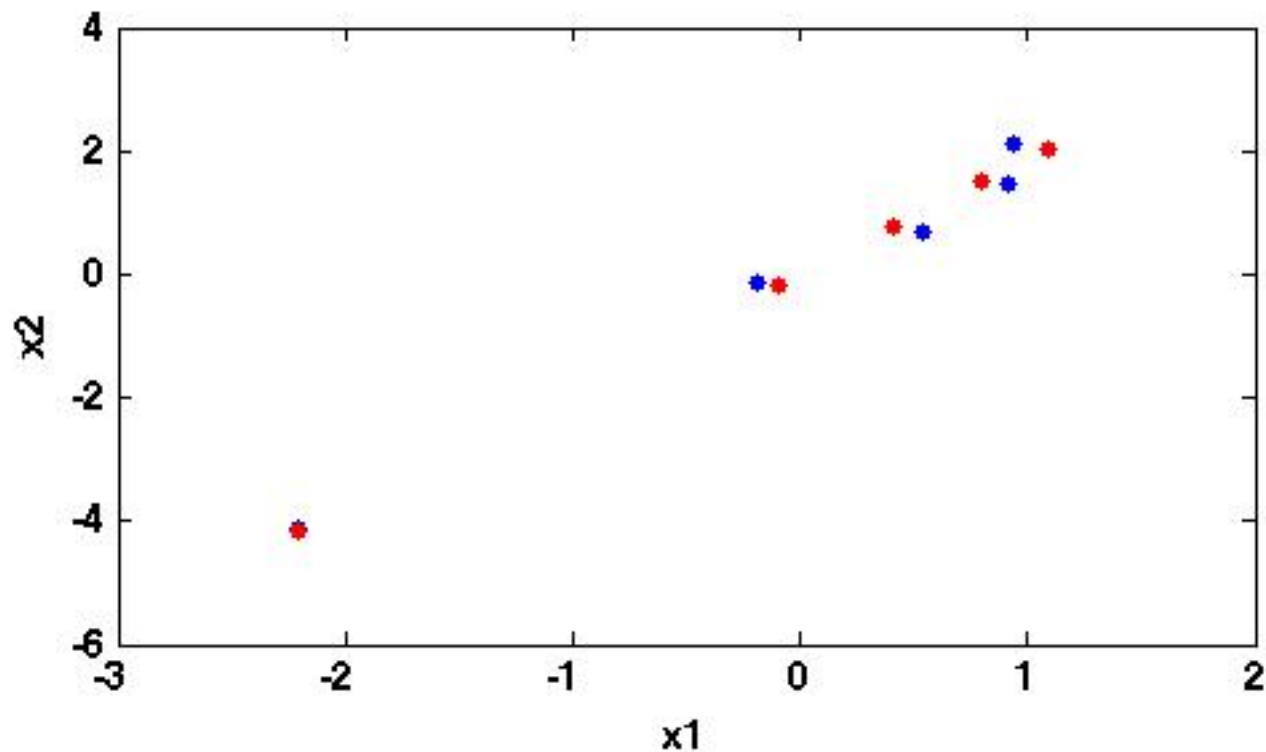


$$t_1 = \begin{bmatrix} -0.8628 \\ -1.7056 \\ 4.6881 \\ -2.3184 \\ 0.1950 \end{bmatrix}$$

$$p_1^T = [-0.4697 \ -0.8828]$$

PCA come modello - 6

- Sulla base di questa rappresentazione sulla prima PC è possibile ricostruire un'approssimazione dei dati originali (in rosso).
- Le distanze tra i dati originali (blu) e l'approssimazione fatta a partire da una sola PC rappresentano i residui:



PCA come modello - 7

- E numericamente:

$$X_c = \begin{bmatrix} 0.5443 & 0.6877 \\ 0.9086 & 1.4486 \\ -2.2066 & -4.1364 \\ 0.9389 & 2.2125 \\ -0.1852 & -0.1224 \end{bmatrix}$$

$$X(1) = t_1 p_1^T = \begin{bmatrix} 0.4052 & 0.7617 \\ 0.8011 & 1.5057 \\ -2.2020 & -4.1388 \\ 1.0872 & 2.0436 \\ -0.0916 & -0.1722 \end{bmatrix}$$

$$E = X_c - X(1) = \begin{bmatrix} 0.1391 & -0.0740 \\ 0.1075 & -0.0572 \\ -0.0047 & 0.0025 \\ -0.1483 & 0.0789 \\ -0.0936 & 0.0498 \end{bmatrix}$$

PCA e nuovi dati

- Se si vuole rappresentare nuovi campioni nello spazio delle componenti principali, è sufficiente applicare la stessa trasformazione di coordinate (ovvero moltiplicare la matrice – o il vettore, se c'è un solo campione – dei nuovi dati per il loadings):

$$\mathbf{T}_{new} = \mathbf{X}_{new} \mathbf{P}$$

- Anche in questo caso è possibile ritornare ai dati nello spazio delle variabili facendo la trasformazione inversa:

$$\hat{\mathbf{X}}_{new} = \mathbf{T}_{new} \mathbf{P}^T$$

- E, analogamente a quanto visto nel caso dei dati utilizzati per calcolare il modello, i residui rappresentano una stima della bontà dell'approssimazione:

$$\mathbf{X} - \hat{\mathbf{X}}_{new} = \mathbf{E}_{new}$$

PCA e dati nuovi - 2

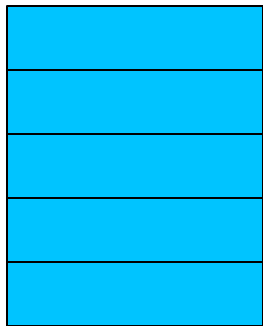
- Dal momento che le ultime PC modellano il contributo dell'errore e che questo contributo è diverso da campione a campione, nel caso di dati non utilizzati per costruire il modello il minimo valore dei residui non si ha utilizzando tutte le PC possibili.
- Il minimo valore degli scarti si ha quando il numero di componenti principali è quello che permette di spiegare tutta la variabilità sistematica (e quindi informativa) e di lasciar fuori solo il “rumore” legato all'errore sperimentale.
- Questo concetto è alla base del metodo della Cross-Validation

Cross-Validation

- Nella Cross-validation (CV) si divide la matrice dei dati in un opportuno numero di segmenti, contenenti uno o più campioni alla volta.
- A turno, ciascuno di questi segmenti è rimosso dalla matrice dei dati e trattato come un set di campioni incogniti.
- I restanti campioni vengono utilizzati per calcolare un modello (in questo caso un modello PCA).
- Il modello viene applicato ai campioni lasciati fuori come incogniti
- Si calcola una stima dell'errore che si compie applicando il modello ai campioni incogniti
- Nel caso della PCA, questo metodo è utilizzato per scegliere il numero di componenti principali ritenute significative

Cross-Validation & PCA

- Nel caso della PCA, la cross-validation segue lo stesso schema:
 - Si divide la matrice dei dati in un certo numero di “segmenti”



- Si seleziona un segmento alla volta come set di campioni incogniti e si rimuove dalla matrice dei dati.



modello



“incogniti”

Cross-validation & PCA - 2

- Si calcolano diversi modelli PCA utilizzando i dati rimasti, che comprendano da 1 al massimo numero possibile di PC
- Si applica ciascuno di questi modelli al set di dati lasciato fuori e si calcola il valore dei residui in funzione del numero di componenti principali:

$$\mathbf{E}(1), \mathbf{E}(2), \mathbf{E}(3) \dots \mathbf{E}(f)$$

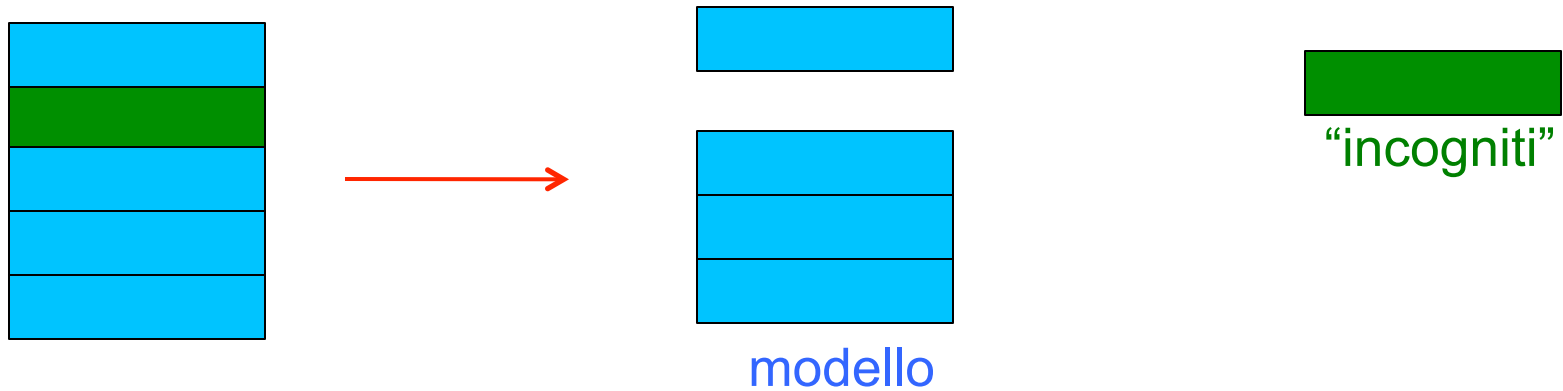
- In particolare, l'errore che si considera è chiamato PRESS (PREdictive Sum of Squares), che non è altro che la somma del quadrato degli scarti calcolati sui dati "incogniti":

$$PRESS(f) = \sum_{i,j} \left(x_{ij} - \hat{x}_{ij}(f) \right)^2$$

$$PRESS(f) = tr(\mathbf{E}^T(f) \mathbf{E}(f))$$

Cross-Validation & PCA - 3

- A questo punto si ripete l'intera procedura su un altro segmento:



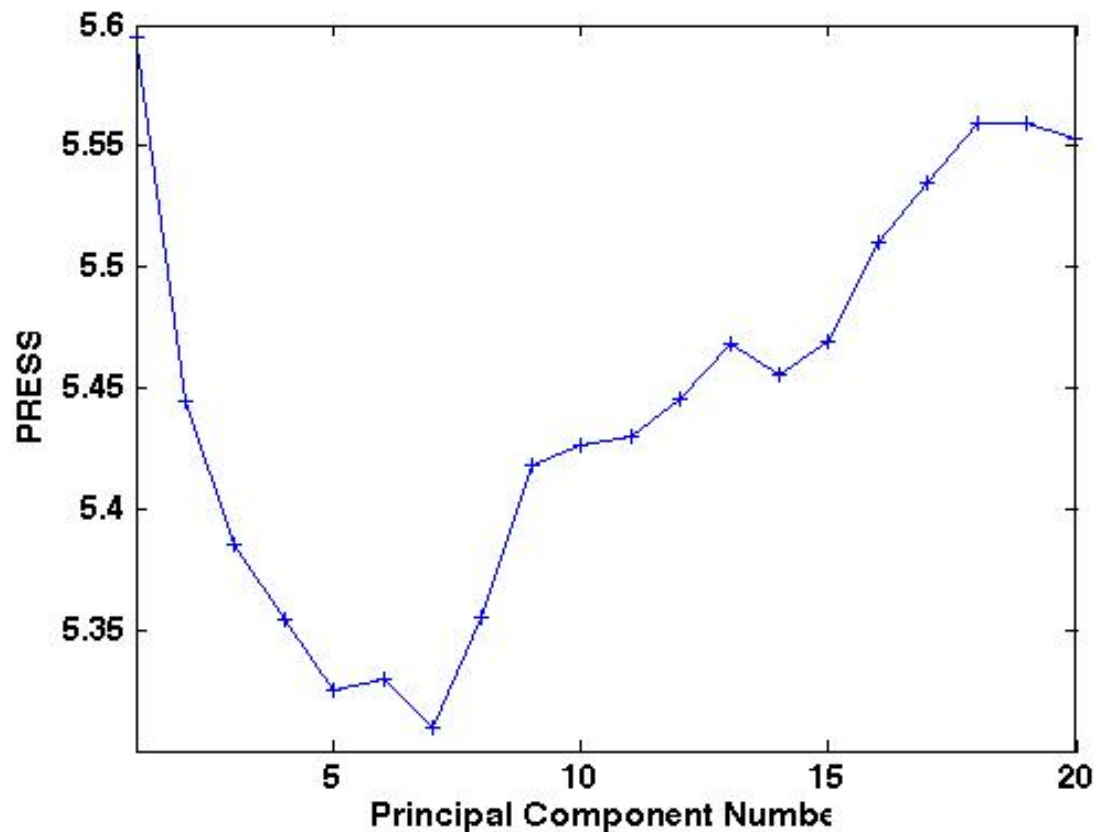
- Ottenendo una nuova stima dell'errore in funzione del numero di componenti principali, che si andrà a sommare alla precedente:

$$PRESS(f) = PRESS(f)_{segm1} + PRESS(f)_{segm2}$$

- Si continua finché ogni segmento non è stato trattato almeno una volta come set di campioni incogniti.

Cross-Validation & PCA

- A questo punto si riporta in grafico il valore del PRESS in funzione del numero di componenti principali e si sceglie il valore corrispondente al minimo dell'errore



PCA e dati anomali

- Un altro dei campi dell'analisi esplorativa in cui la PCA risulta particolarmente utile è l'identificazione di dati anomali (**outliers**)
- La presenza di questi outliers all'interno del set di dati può inficiare la qualità dei modelli che a partire dai dati stessi vengono costruiti.
- Un dato anomalo può essere in primo luogo identificato graficamente osservando il grafico degli scores
- Tuttavia non sempre questa identificazione risulta immediata, soprattutto quando il numero di componenti principali significative è maggiore di 3.
- Esistono dei criteri matematici che permettono di identificare i dati anomali sulla base di un modello PC

PCA e dati anomali - 2

- Un modello di componenti principali costituisce una rappresentazione dei dati su un sottospazio di dimensionalità minore.
- Sulla base di questo concetto, un dato può essere anomalo per due motivi:
 - Perché particolarmente distante dagli altri dati, nello spazio del modello (ovvero nello spazio delle PC significative).
 - Perché particolarmente distante dallo spazio del modello (ovvero perché il modello PC non è in grado di spiegare bene la sua variabilità).
- Entrambe queste distanze possono essere descritte in termini di variabili statistiche (chiamate T^2 e Q) per le quali possono essere calcolati dei valori critici sulla base dei dati.
- Un grafico bidimensionale che riporti, per ciascun campione, i valori di queste due variabili rappresenta un valido strumento per l'identificazione degli outliers e della loro natura

T^2 vs Q plot

