Introduzione alla chemiometria

Chemiometria

Matematica

Statistica

Science dell’Informazione

In Chimica

Discipline simili

- Biometrics ±1900
- Psychometrics ±1930
- Econometrics ±1950
- Technometrics ±1960

Qualche dato storico

- **Nome proposto originariamente nei primi anni 1970 dal chimico organico svedese Svante Wold.**
- **International Chemometrics Society - 1970s.**
- **Meeting Internazionale - Cosenza 1983**
- **Riviste : 1986 (Chemometrics and Intelligent Laboratory Systems) and 1987 (J Chemometrics)**
- **Libri : metà anni 1980**
- **Corsi : nei tardi anni 1980 principalmente come formazione professionale continua.**

Cosa è la chemiometria

La chemiometria è un settore della chimica che studia l'applicazione dei metodi matematici o statistici ai dati chimici. La International Chemometrics Society (ICS) ne da la seguente definizione: *la chemiometria è la scienza di relazionare le misure effettuate su un sistema o su un processo chimico allo stato del sistema via applicazione di metodi matematici o statistici*.

http://www.gruppochemiometria.it

Altre definizioni…

”A chemical discipline that uses mathematical and statistical methods to: design/select optimal procedures and experiments, provide maximum chemical information by analysing data, give a graphical representation of this information, in other words… information aspects of chemistry” (D.L.Massart)

La chemiometria può essere definita come la branca della chimica che si serve di metodi matematici, statistici e logici per:

- progettare, selezionare ed ottimizzare procedure ed esperimenti;
- estrarre la massima informazione possibile sul sistema in esame attraverso l’analisi dei dati;
- fornire una rappresentazione grafica di questa informazione.

(modificato da
http://www.iupac.org/publications/pac/pdf/1983/pdf/5512x1861.pdf)

Appare chiaro come la chemiometria accompagni il processo chimico, ed in particolare chimico-analitico, lungo tutte le sue fasi a partire dal campionamento fino all'ottimizzazione.

Campi di applicazione della chemiometria

Tra i campi d'applicazione della chemiometria si possono citare:

- Controllo di qualità
- Monitoraggio e controllo di processo
- Tracciabilità degli alimenti
- QSAR/QSPR e REACH
- Genomica, proteomica e metabolomica
- Progettazione degli esperimenti e Ottimizzazione
- Progettazione di Farmaci e materiali (*Drug & material design*)
- Analisi di immagini
- Applicazioni in ambito industriale e ambientale

Analisi Multivariata dei Dati

- Dati campionati e progetti con molte risposte anche da:
  - Attività minerarie
  - Ospedali
  - Agricultura
  - Industria alimentare
  - Etc.

La Chemiometria è una disciplina per l'analisi dei dati che:

- Tratta dati **multivariati** (e "**multiway**")
- Si basa su modellizzazione **soft**
- Usa **metodi di proiezione** e il concetto di **variabili latenti**
- Considera i **dati** come **informazione + rumore**
- Considera il **rumore** come **informazione inutile**

Nomenclatura

- I campioni sono oggetti
- Ciò che è misurato su un oggetto è una variabile

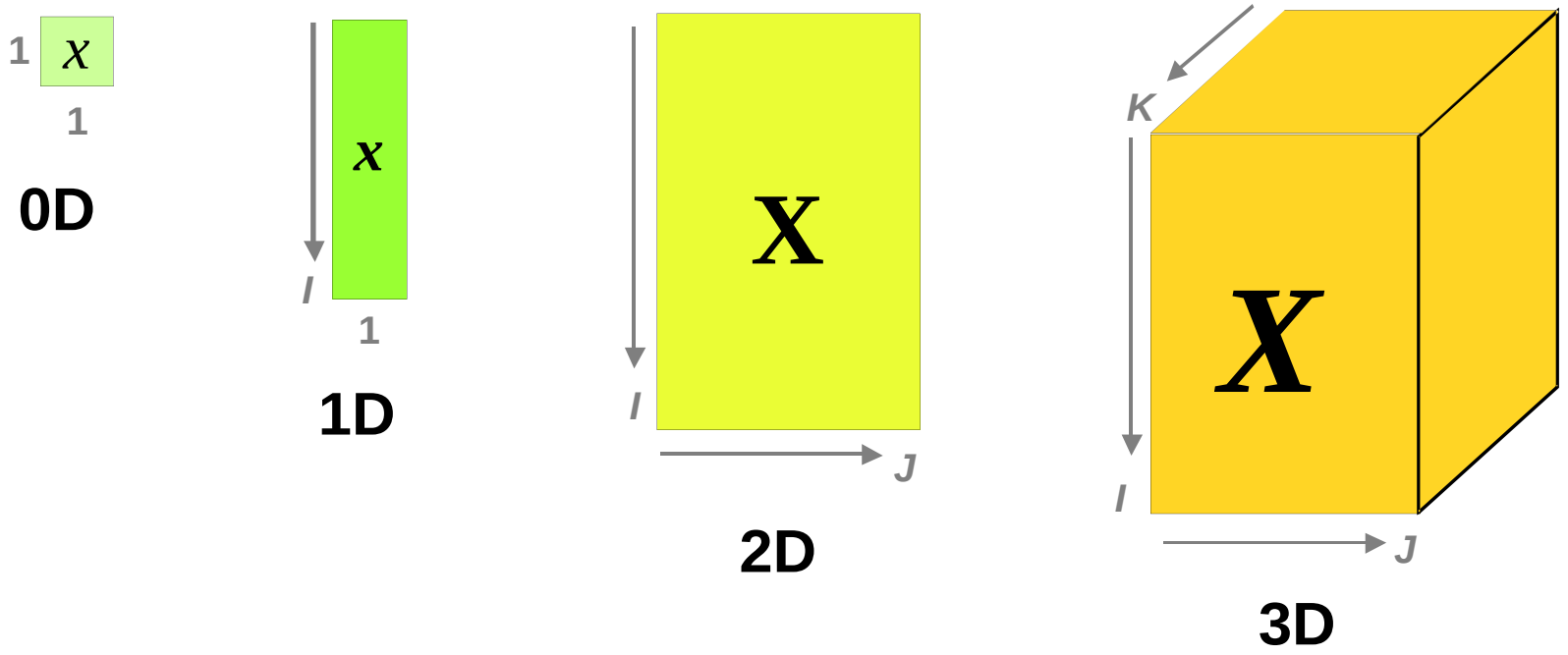
Dati multivariati

Variabili

Campioni

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
s1	-1.19E-01	7.28E-01	-2.15E-02	5.22E-01	7.06E-04	7.32E-01	3.10E-04	-6.13E-04	-5.92E-05	1.28E+00	1.32E+00	-7.03E-02	1.23E-03	4.67E-01	-4.21E-
s2	-1.37E-01	7.28E-01	-2.89E-02	6.08E-01	7.09E-04	7.02E-01	6.58E-04	-1.22E-03	-1.49E-04	1.35E+00	1.39E+00	-3.27E-01	2.48E-04	4.84E-01	-2.94E-
s3	2.51E-02	-9.15E-02	6.73E-03	-1.13E-01	-9.07E-05	-7.58E-02	-2.29E-04	4.10E-04	5.65E-05	-1.96E-01	-2.02E-01	1.49E-01	3.83E-04	-6.80E-02	1.43E-
s4	-1.14E-01	6.70E-01	-2.18E-02	5.04E-01	6.50E-04	6.65E-01	3.83E-04	-7.34E-04	-7.96E-05	1.20E+00	1.24E+00	-1.36E-01	8.59E-04	4.34E-01	-1.09E-
s5	-7.93E-02	4.14E-01	-1.69E-02	3.51E-01	4.04E-04	3.98E-01	3.96E-04	-7.35E-04	-9.05E-05	7.71E-01	7.94E-01	-2.02E-01	7.80E-05	2.76E-01	-1.83E-
s6	1.51E-02	-6.38E-02	3.74E-03	-6.75E-02	-6.28E-05	-5.67E-02	-1.15E-04	2.07E-04	2.78E-05	-1.29E-01	-1.33E-01	7.08E-02	1.40E-04	-4.52E-02	6.71E-
s7	7.44E-02	-5.24E-01	1.11E-02	-3.24E-01	-5.06E-04	-5.45E-01	-1.73E-05	7.92E-05	-1.07E-05	-8.87E-01	-9.13E-01	-1.02E-01	-1.47E-03	-3.26E-01	-1.19E-
s8	3.65E-02	-2.66E-01	5.12E-03	-1.59E-01	-2.56E-04	-2.78E-01	1.43E-05	-3.95E-07	-1.14E-05	-4.46E-01	-4.59E-01	-6.86E-02	-8.12E-04	-1.64E-01	-7.71E-
s9	1.36E-01	-7.06E-01	2.89E-02	-6.01E-01	-6.88E-04	-6.77E-01	-6.83E-04	1.26E-03	1.56E-04	-1.31E+00	-1.35E+00	3.50E-01	-1.12E-04	-4.71E-01	3.18E-
s10	-2.74E-02	3.60E-01	1.82E-03	1.12E-01	3.42E-04	4.12E-01	-4.31E-04	7.24E-04	1.22E-04	5.29E-01	5.43E-01	3.97E-01	2.27E-03	2.02E-01	4.03E-
s11	7.47E-02	-3.31E-01	1.80E-02	-3.34E-01	-3.25E-04	-2.99E-01	-5.30E-04	9.62E-04	1.28E-04	-6.54E-01	-6.74E-01	3.20E-01	5.44E-04	-2.31E-01	3.01E-
s12	-1.17E-01	7.02E-01	-2.16E-02	5.13E-01	6.81E-04	7.03E-01	3.40E-04	-6.63E-04	-6.76E-05	1.25E+00	1.28E+00	-9.79E-02	1.07E-03	4.52E-01	-7.02E-
s13	1.06E-01	-2.82E-01	3.23E-02	-4.82E-01	-2.85E-04	-1.87E-01	-1.25E-03	2.21E-03	3.14E-04	-7.01E-01	-7.25E-01	8.59E-01	2.71E-03	-2.36E-01	8.33E-
s14	7.39E-02	-5.28E-01	1.07E-02	-3.21E-01	-5.09E-04	-5.50E-01	2.49E-06	4.48E-05	-1.59E-05	-8.90E-01	-9.15E-01	-1.17E-01	-1.54E-03	-3.27E-01	-1.34E-
s15	-9.87E-03	1.02E-01	-3.21E-04	4.17E-02	9.75E-05	1.13E-01	-8.29E-05	1.36E-04	2.44E-05	1.57E-01	1.61E-01	8.35E-02	5.31E-04	5.92E-02	8.57E-
s16	-1.06E-01	7.68E-01	-1.52E-02	4.62E-01	7.41E-04	8.03E-01	-2.54E-05	-2.68E-05	2.88E-05	1.29E+00	1.33E+00	1.86E-01	2.30E-03	4.75E-01	2.11E-
s17	-4.76E-02	2.66E-01	-9.52E-03	2.10E-01	2.59E-04	2.61E-01	1.92E-04	-3.61E-04	-4.19E-05	4.84E-01	4.99E-01	-8.35E-02	2.29E-04	1.75E-01	-7.21E-
s18	9.54E-02	-6.55E-01	1.48E-02	-4.16E-01	-6.33E-04	-6.77E-01	-6.69E-05	1.79E-04	-1.61E-06	-1.12E+00	-1.15E+00	-9.35E-02	-1.71E-03	-4.10E-01	-1.16E-
s19	-1.32E-01	5.01E-01	-3.49E-02	5.94E-01	4.96E-04	4.22E-01	1.16E-03	-2.09E-03	-2.86E-04	1.06E+00	1.09E+00	-7.49E-01	-1.84E-03	3.68E-01	-7.17E-
s20	8.91E-02	-4.23E-01	2.05E-02	-3.97E-01	-4.14E-04	-3.94E-01	-5.56E-04	1.02E-03	1.32E-04	-8.15E-01	-8.40E-01	3.19E-01	3.49E-04	-2.90E-01	2.97E-
s21	-8.91E-02	5.08E-01	-1.75E-02	3.93E-01	4.94E-04	5.01E-01	3.34E-04	-6.35E-04	-7.18E-05	9.17E-01	9.45E-01	-1.36E-01	5.26E-04	3.31E-01	-1.15E-
s22	1.15E-01	-6.22E-01	2.39E-02	-5.10E-01	-6.06E-04	-6.04E-01	-5.25E-04	9.79E-04	1.18E-04	-1.14E+00	-1.18E+00	2.52E-01	-3.18E-04	-4.12E-01	2.24E-
s23	-4.08E-02	5.43E-01	2.94E-03	1.67E-01	5.17E-04	6.22E-01	-6.60E-04	1.11E-03	1.86E-04	7.96E-01	8.18E-01	6.06E-01	3.45E-03	3.04E-01	6.15E-
s24	9.92E-02	-6.00E-01	1.82E-02	-4.36E-01	-5.82E-04	-6.02E-01	-2.81E-04	5.49E-04	5.53E-05	-1.06E+00	-1.09E+00	7.64E-02	-9.46E-04	-3.86E-01	5.28E-
s25	1.08E-01	-5.37E-01	2.40E-02	-4.81E-01	-5.25E-04	-5.07E-01	-6.15E-04	1.13E-03	1.43E-04	-1.02E+00	-1.05E+00	3.37E-01	1.85E-04	-3.63E-01	3.11E-
s26	-6.95E-02	4.56E-01	-1.15E-02	3.04E-01	4.41E-04	4.67E-01	1.03E-04	-2.23E-04	-1.34E-05	7.88E-01	8.11E-01	2.38E-02	1.03E-03	2.88E-01	4.01E-
s27	4.90E-02	-1.23E-01	1.51E-02	-2.22E-01	-1.25E-04	-7.71E-02	-5.94E-04	1.05E-03	1.50E-04	-3.15E-01	-3.25E-01	4.12E-01	1.33E-03	-1.05E-01	4.00E-
s28	-1.65E-03	-7.79E-02	-3.41E-03	1.09E-02	-7.26E-05	-1.01E-01	2.35E-04	-4.05E-04	-6.30E-05	-8.88E-02	-9.08E-02	-1.91E-01	-8.92E-04	-3.68E-02	-1.90E-
s29	8.73E-02	-5.70E-01	1.46E-02	-3.82E-01	-5.52E-04	-5.83E-01	-1.37E-04	2.94E-04	1.91E-05	-9.87E-01	-1.02E+00	-2.30E-02	-1.27E-03	-3.61E-01	-4.36E-
s30	-6.93E-02	6.98E-02	-2.51E-02	3.19E-01	7.94E-05	-2.22E-02	1.11E-03	-1.96E-03	-2.85E-04	3.19E-01	3.31E-01	-8.06E-01	-2.95E-03	9.79E-02	-7.88E-
s31	-8.99E-02	3.66E-01	-2.28E-02	4.03E-01	3.61E-04	3.20E-01	7.21E-04	-1.30E-03	-1.76E-04	7.48E-01	7.72E-01	-4.52E-01	-9.80E-04	2.63E-01	-4.30E-
s32	-6.32E-02	2.05E-01	-1.79E-02	2.85E-01	2.05E-04	1.59E-01	6.44E-04	-1.15E-03	-1.61E-04	4.63E-01	4.78E-01	-4.31E-01	-1.23E-03	1.59E-01	-4.15E-
s33	-1.42E-01	6.98E-01	-3.20E-02	6.33E-01	6.83E-04	6.57E-01	8.32E-04	-1.53E-03	-1.95E-04	1.33E+00	1.37E+00	-4.62E-01	-3.32E-04	4.74E-01	-4.28E-
s34	1.32E-01	-6.89E-01	2.81E-02	-5.85E-01	-6.72E-04	-6.62E-01	-6.61E-04	1.23E-03	1.51E-04	-1.28E+00	-1.32E+00	3.38E-01	-1.27E-04	-4.60E-01	3.06E-
s35	-1.08E-01	4.80E-01	-2.61E-02	4.84E-01	4.72E-04	4.35E-01	7.66E-04	-1.39E-03	-1.84E-04	9.49E-01	9.79E-01	-4.61E-01	-7.76E-04	3.36E-01	-4.34E-
s36	2.13E-02	-2.11E-01	9.91E-04	-9.02E-02	-2.02E-04	-2.34E-01	1.57E-04	-2.55E-04	-4.66E-05	-3.28E-01	-3.38E-01	-1.62E-01	-1.06E-03	-1.24E-01	-1.67E-
s37	-2.39E-03	4.55E-03	-7.90E-04	1.09E-02	4.75E-06	1.94E-03	3.28E-05	-5.80E-05	-8.35E-06	1.36E-02	1.41E-02	-2.32E-02	-7.97E-05	4.43E-03	-2.26E-
s38	6.29E-02	-3.26E-01	1.35E-02	-2.79E-01	-3.18E-04	-3.13E-01	-3.21E-04	5.94E-04	7.36E-05	-6.09E-01	-6.27E-01	1.66E-01	-3.74E-05	-2.18E-01	1.51E-
s39	1.02E-01	-5.03E-01	2.26E-02	-4.51E-01	-4.91E-04	-4.74E-01	-5.80E-04	1.07E-03	1.35E-04	-9.53E-01	-9.83E-01	3.19E-01	1.85E-04	-3.40E-01	2.94E-
s40	1.00E-01	-4.22E-01	2.50E-02	-4.49E-01	-4.16E-04	-3.75E-01	-7.68E-04	1.39E-03	1.86E-04	-8.52E-01	-8.79E-01	4.75E-01	9.52E-04	-3.00E-01	4.51E-
s41	7.05E-02	-3.24E-01	1.66E-02	-3.14E-01	-3.17E-04	-2.97E-01	-4.70E-04	8.56E-04	1.12E-04	-6.31E-01	-6.51E-01	2.77E-01	3.94E-04	-2.24E-01	2.60E-
s42	-7.27E-02	3.21E-01	-1.76E-02	3.24E-01	3.15E-04	2.90E-01	5.17E-04	-9.38E-04	-1.25E-04	6.35E-01	6.55E-01	-3.12E-01	-5.35E-04	2.25E-01	-2.94E-
s43	-3.34E-02	2.26E-01	-5.31E-03	1.46E-01	2.18E-04	2.33E-01	3.24E-05	-7.79E-05	-1.84E-06	3.87E-01	3.98E-01	2.54E-02	5.64E-04	1.42E-01	3.33E-
s44	-1.36E-01	6.25E-01	-3.21E-02	6.08E-01	6.13E-04	5.73E-01	9.11E-04	-1.66E-03	-2.18E-04	1.22E+00	1.26E+00	-5.37E-01	-7.70E-04	4.32E-01	-5.04E-
s46	2.41E-02	-1.53E-01	4.17E-03	-1.06E-01	-1.48E-04	-1.55E-01	-4.92E-05	1.01E-04	8.29E-06	-2.67E-01	-2.75E-01	2.90E-03	-3.05E-04	-9.74E-02	-2.79E-
s47	-9.11E-02	4.29E-01	-2.11E-02	4.06E-01	4.21E-04	3.98E-01	5.77E-04	-1.05E-03	-1.37E-04	8.29E-01	8.55E-01	-3.33E-01	-3.91E-04	2.95E-01	-3.11E-
s49	5.84E-02	-2.37E-01	1.49E-02	-2.62E-01	-2.34E-04	-2.07E-01	-4.71E-04	8.49E-04	1.15E-04	-4.85E-01	-5.00E-01	2.96E-01	6.47E-04	-1.70E-01	2.82E-
s50	-4.05E-02	2.15E-01	-8.50E-03	1.79E-01	2.10E-04	2.08E-01	1.93E-04	-3.58E-04	-4.36E-05	3.98E-01	4.10E-01	-9.53E-02	7.91E-05	1.43E-01	-8.56E-

dati **multivariati** (e “**multiway**”)



Molti inputs inducono un effetto

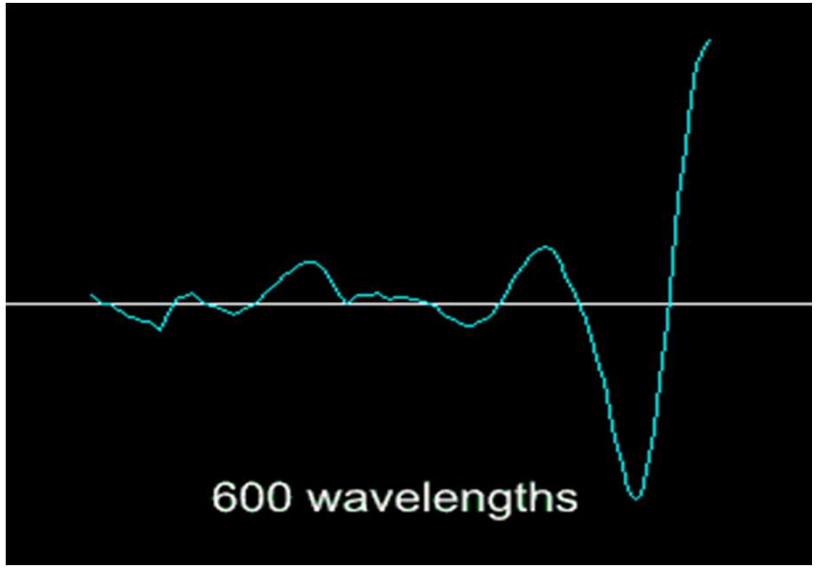
Molti effetti sono derivati da un input from

etc

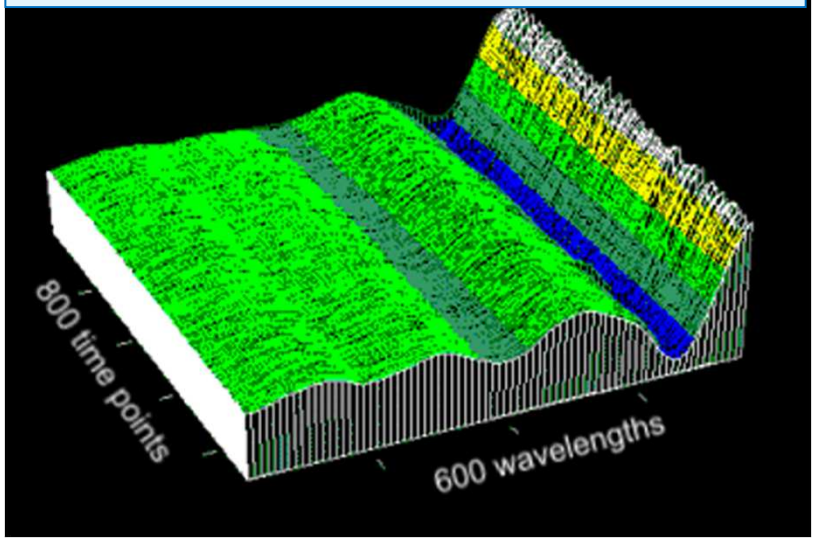
dati **multivariati** (e “**multiway**”)

Molte variabili e molti campioni

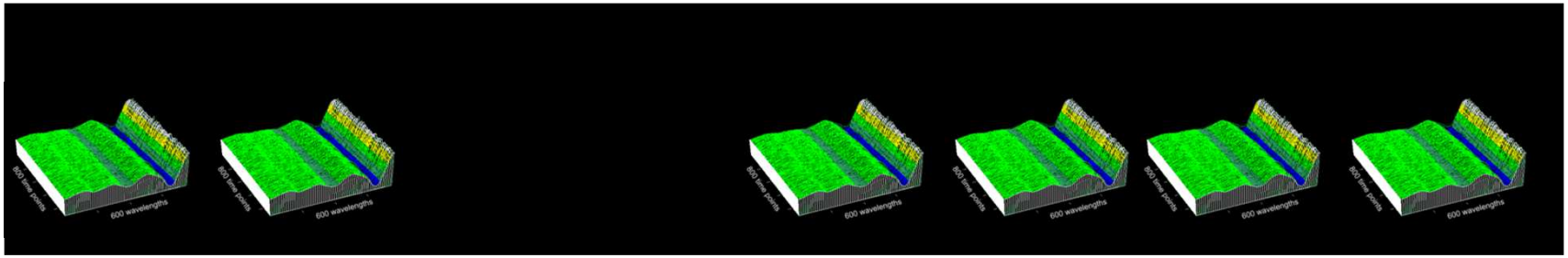
Una misura – spettro (600 punti)



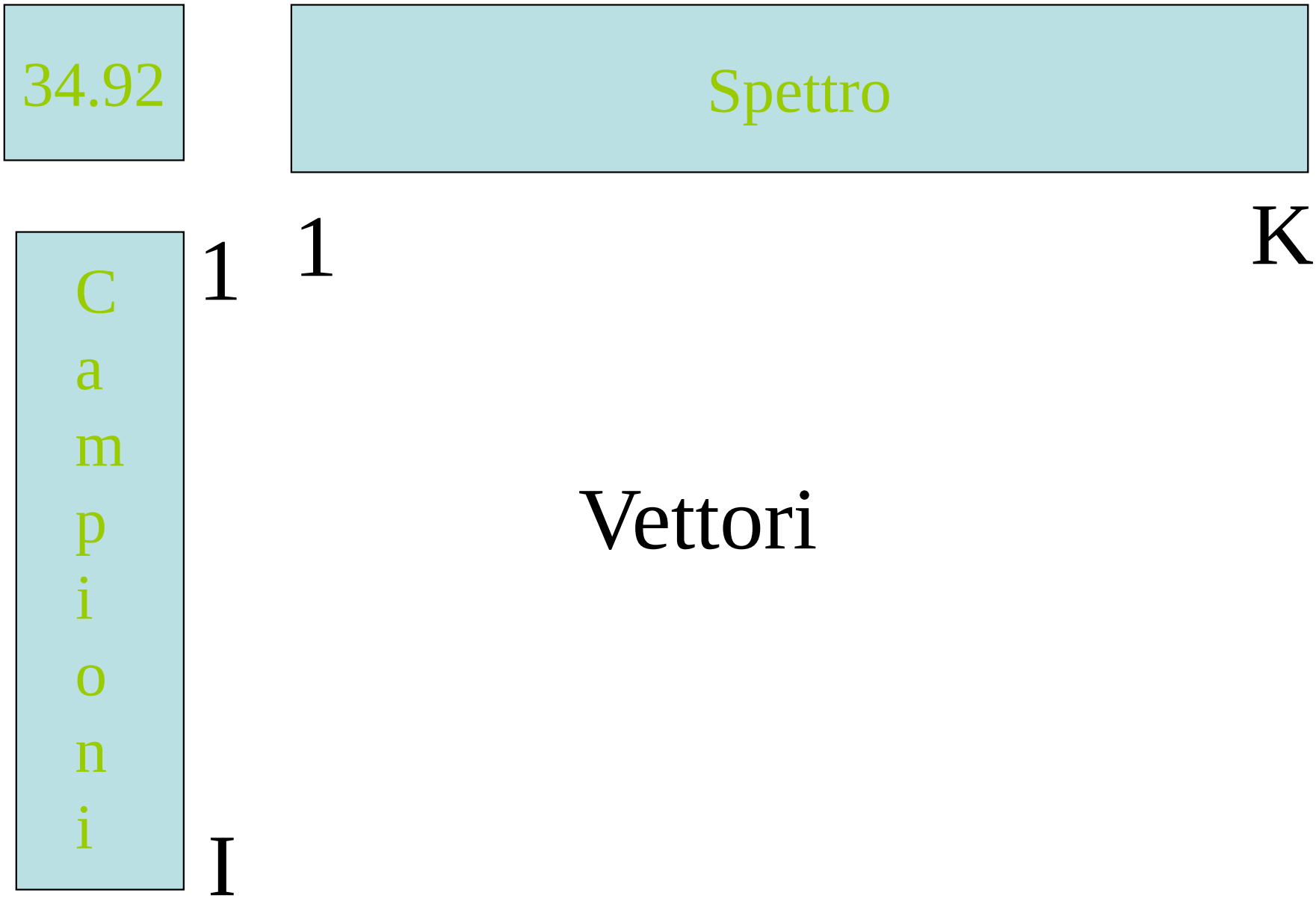
Un "batch" – 800 spettri (suddivisioni temporali)



Un set di dati – 200 campioni (batches)



dati **multivariati** (e "**multiway**")


34.92
Spettro
1
1
K
Campioni
Vettori
I

Un vettore è una raccolta di numeri.

E’ sempre un vettore colonna

12 3.6 11.1 5.9 34 0.5 1.4 17

La trasposta di un vettore è un vettore riga.
I simboli per indicare la trasposizione sono ’ o ^{T}.
Es. **a**’ o **a**^{T}.

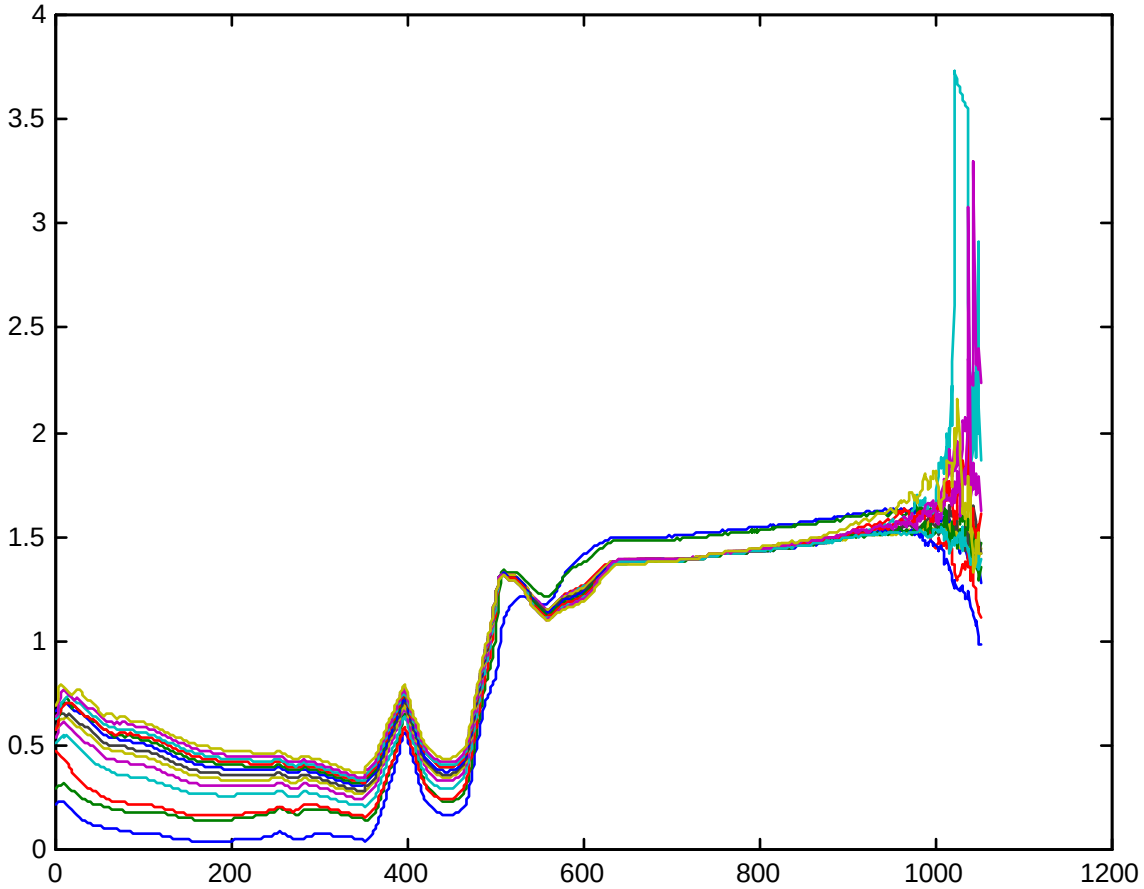
La matrice dei dati

K

Una matrice di dati
è un vettore di vettori

I

Tempi in una reazione batch



Lunghezze d’onda NIR

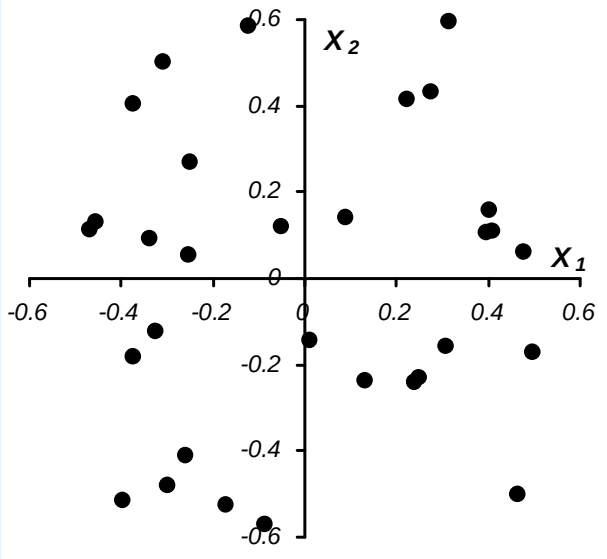
Modelli “hard” e “soft”

	Modelli “hard”	Modelli “soft”
Origine	Da conoscenza *a priori*	Dai dati
Formula	$\mathbf{y}=f(\mathbf{x},\mathbf{a})+\boldsymbol{\varepsilon}$	$\mathbf{y}=\mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon}$
Parametri	Hanno significato fisico esplicito	Non c’è significato fisico “esplicito”
Problema	formulazione di modelli	Analisi dei dati
Scopo	estrapolazione	interpolazione
Example	Beer-Lambert	ANOVA

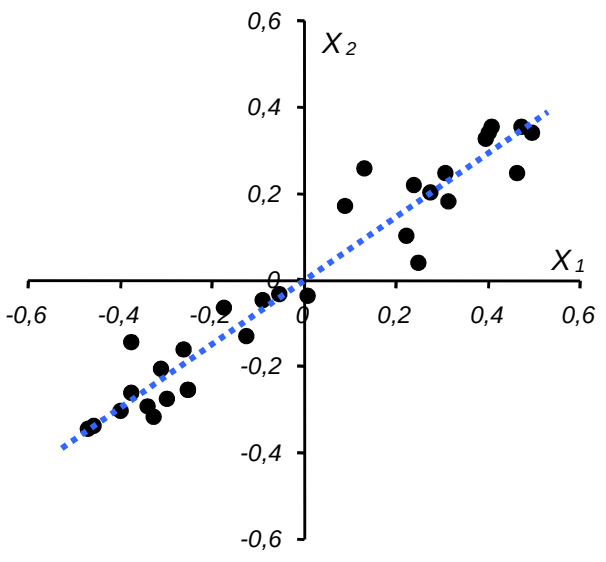
Es. L'**analisi della varianza** (**ANOVA**) è un insieme di tecniche statistiche che permettono di confrontare due o più gruppi di dati confrontando la variabilità *interna* a questi gruppi con la variabilità *tra* i gruppi. Si confrontano medie di due o più campioni tenendo conto contemporaneamente di più variabili.

Metodi di proiezione e variabili latenti

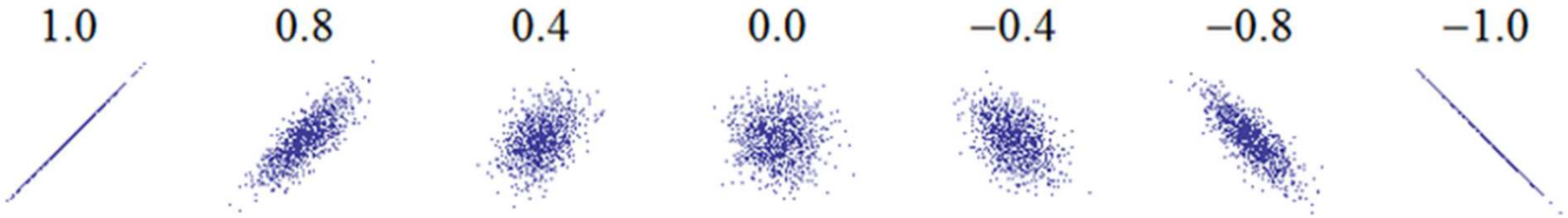
Dati senza struttura



Dati con una struttura nascosta



In statistica per **correlazione** si intende una relazione tra due variabili tale che a ciascun valore della prima variabile corrisponda con una certa regolarità un valore della seconda. Non si tratta necessariamente di un rapporto di causa ed effetto, ma semplicemente della tendenza di una variabile a variare in funzione di un'altra.



$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \leq +1$$

Grandi ”aree” della Chemiometria

1. Progettazione degli esperimenti (*Design of Experiments - DOE*)
2. Analisi esplorativa dei dati (*Exploratory Data Analysis*)
3. Classificazione (*Classification*)
4. Regressione e calibrazione (*Regression and Calibration*)

In ciascuna ”area” ci sono molti metodi

http://www.difar.unige.it/images/Chimica_Analitica/depliant_Scuola_2016_feb.pdf

1) **Progettazione degli esperimenti**

(*Design of Experiments o Experimental Design)*

Di estrema importanza, da applicare ove possibile

Impiega:

- ANOVA
- F-test
- t-test
- Diagrammi
- Superfici di risposta

RECUPERARE E LEGGERE

Riccardo Leardi

«Experimental design in chemistry: A tutorial»

Analytica Chimica Acta Volume 652, Issues 1-2, 2009, Pages 161-172

Progettazione degli esperimenti

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_Kx_K + b_{11}x_1^2 + b_{22}x_2^2 + ... + b_{KK}x_K^2 + b_{12}x_1x_2 + ... + \varepsilon$$

I Fattori x_1, x_2,...x_K possono essere modificati sistematicamente

La Risposta y è misurata e modellata

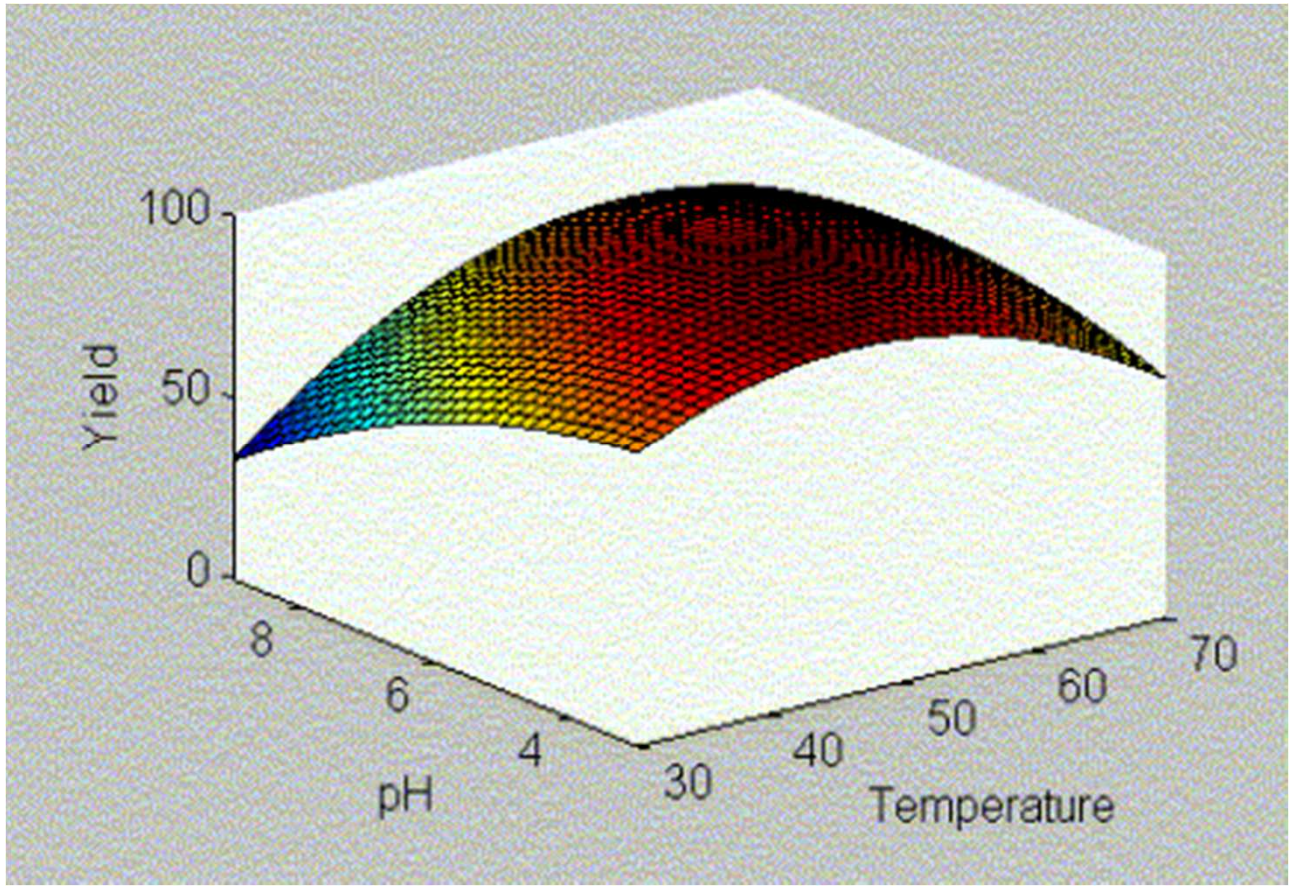
Perché progettare gli esperimenti

- *Screening* (per capire quali sono le variabili importanti nel determinare il valore di una risposta)
- *Saving time* (per risparmiare tempo)
- *Quantitative modelling* (per costruire un modello quantitativo dell'esperimento)
- *Optimisation* (per massimizzare rese di reazione, ottimizzare tempi, consumo di reagenti …)

Perché progettare gli esperimenti?

Un problema : Ottimizzazione di una resa di reazione con pH e temperatura.

Possiamo trovare la combinazione di pH e temperatura che producono la resa migliore della reazione?

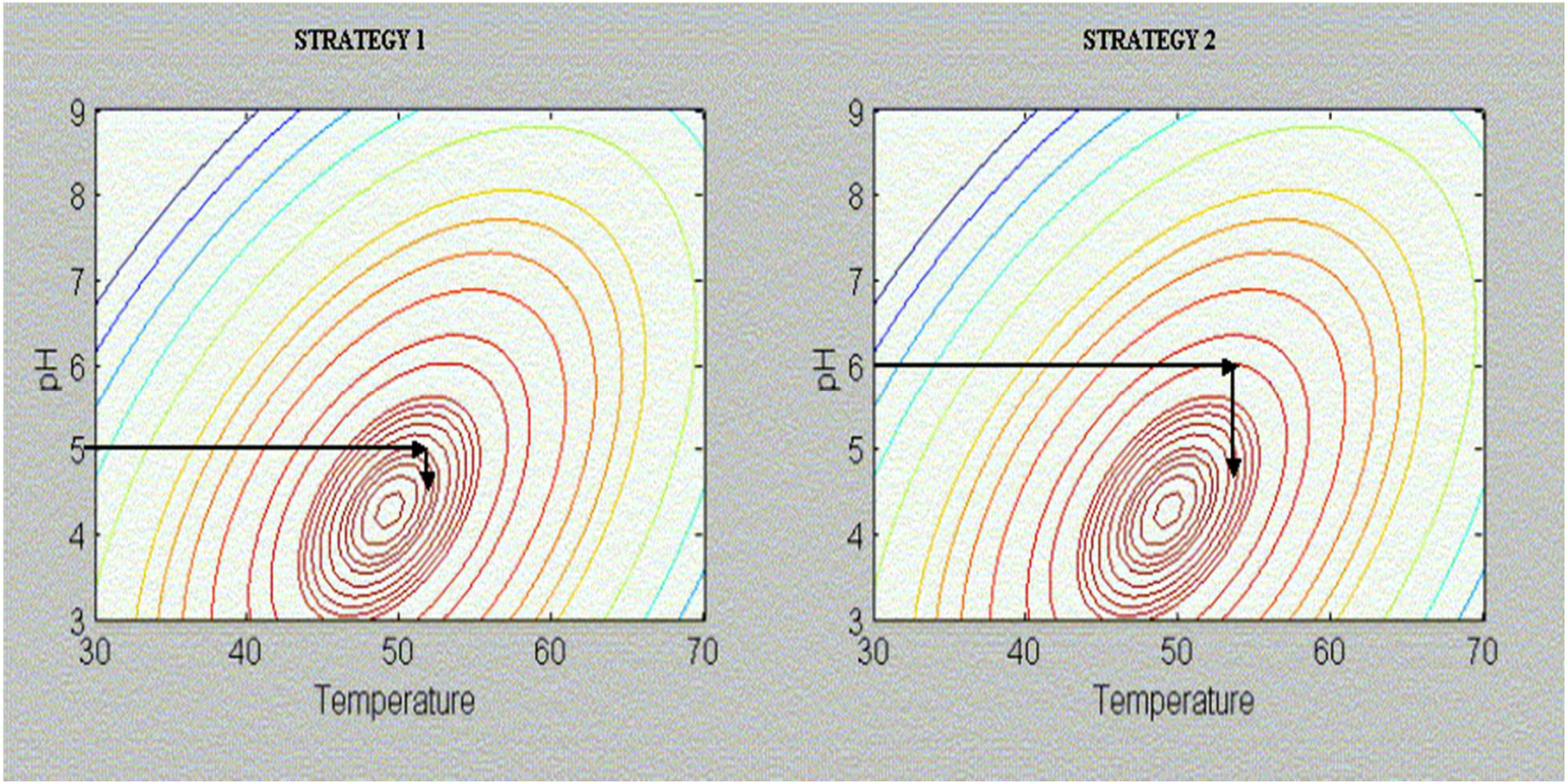


Progettazione degli esperimenti

La strategia di variare un fattore alla volta

(*One Variable At Time*):

può mancare di cogliere l' "ottimo"



DIFFICOLTA’

Interazioni – la risposta per ciascun fattore non è indipendente

La temperatura ottimale a pH 5 è diversa da quella a pH 6.

Come affrontare il problema? Forza bruta?

- **Una griglia di esperimenti (*Grid search*).**
10 pHs, 10 temperatures, 100 experiments.

- **Si inizia con una griglia a maglia larga.**
Poi a maglia più stretta.

Controindicazioni

- **Dispendioso in termini di tempo e denaro.**
- **Molti esperimenti vengono condotti in aree del dominio sperimentale che sono quasi sicuramente “non vicine” a un ottimo (quindi una perdita di tempo e danaro)**
- **Come stimare riproducibilità ed errore sperimentale? (Altri esperimenti, replica dei precedenti ?!?)**

Che facciamo?

Abbiamo bisogno di regole !

La progettazione formale degli esperimenti

Analytica Chimica Acta Volume 652, Issues 1-2,
12 October 2009, Pages 161-172
Experimental design in chemistry: A tutorial
Riccardo Leardi,
http://www.sciencedirect.com/science/article/pii/S0003267009008058

Divisione di Chimica Analitica
Società Chimica Italiana

GRUPPO DIVISIONALE DI CHEMIOMETRIA

UNIVERSITA' DEGLI STUDI DI GENOVA

The Section of Chemistry and Food and Pharmaceutical Technologies, Research Group of Analytical Chemistry and Chemometrics, of the Department of Pharmacy of the University of Genoa organizes a SCHOOL OF EXPERIMENTAL DESIGN (September 15-19, 2014).

Module 1 (15 September, 2.00 pm – 17 September, 1.00 pm)
Program:
Full Factorial Designs, Screening Designs, Fractional Factorial Designs, Response Surface Methodology (Central Composite Design).

Module 2 (17 September, 2.00 pm – 19 September, 1.00 pm)
Program:
Doehlert Design, D-Optimal Designs, "Multicriteria Decision Making", Designs with qualitative variables having more
Mixture Designs, Composite Designs (mixture + process variables).

be made by theoretical lessons and example illustrations, with hands-on-computer sessions in Excel
ed free software.

eats for each of the modules is limited to 12.
: English.
ake place at the Department of Pharmacy, Via Brigata Salerno 13, I-16147 GENOVA.

Progettazione degli esperimenti

Screening

- **Factorial designs**
- **Partial factorials and Plackett-Burman designs**

Modelling and optimisation

- **Response surface designs**
- **Mixture designs**

2) **Analisi Esplorativa dei Dati**

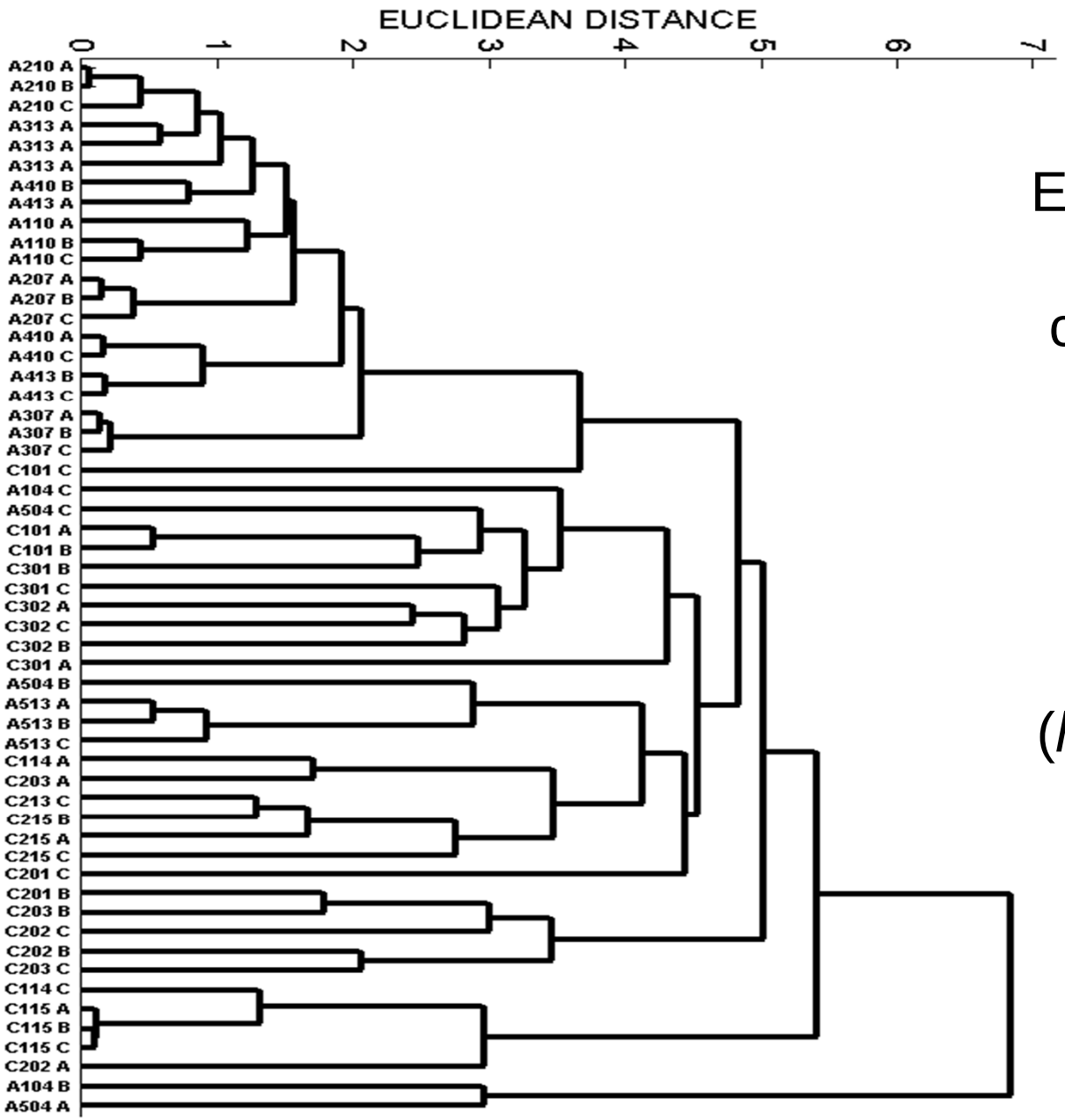
- Non ci è stato possibile progettare

Vogliamo:

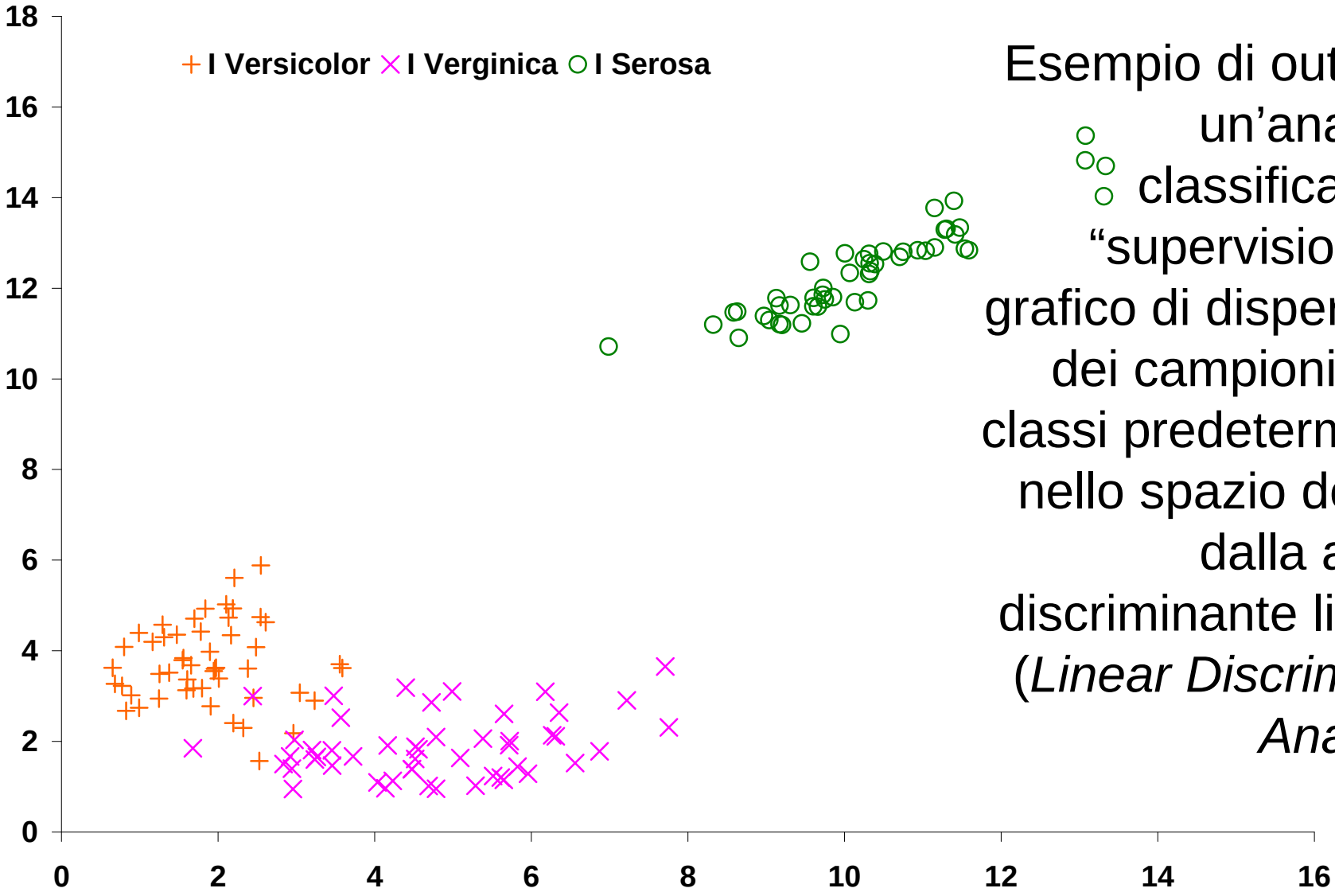
- Trovare strutture
- Trovare raggruppamenti
- Trovare dati anomali (outliers)

3) **Metodi di Classificazione**

- Ricerca di raggruppamenti (di campioni, molecole, etc.)= UNSUPERVISED classification
- I raggruppamenti sono noti = SUPERVISED classification
- Visualizzare i raggruppamenti
- Classificare
- Testare/validare la classificazione



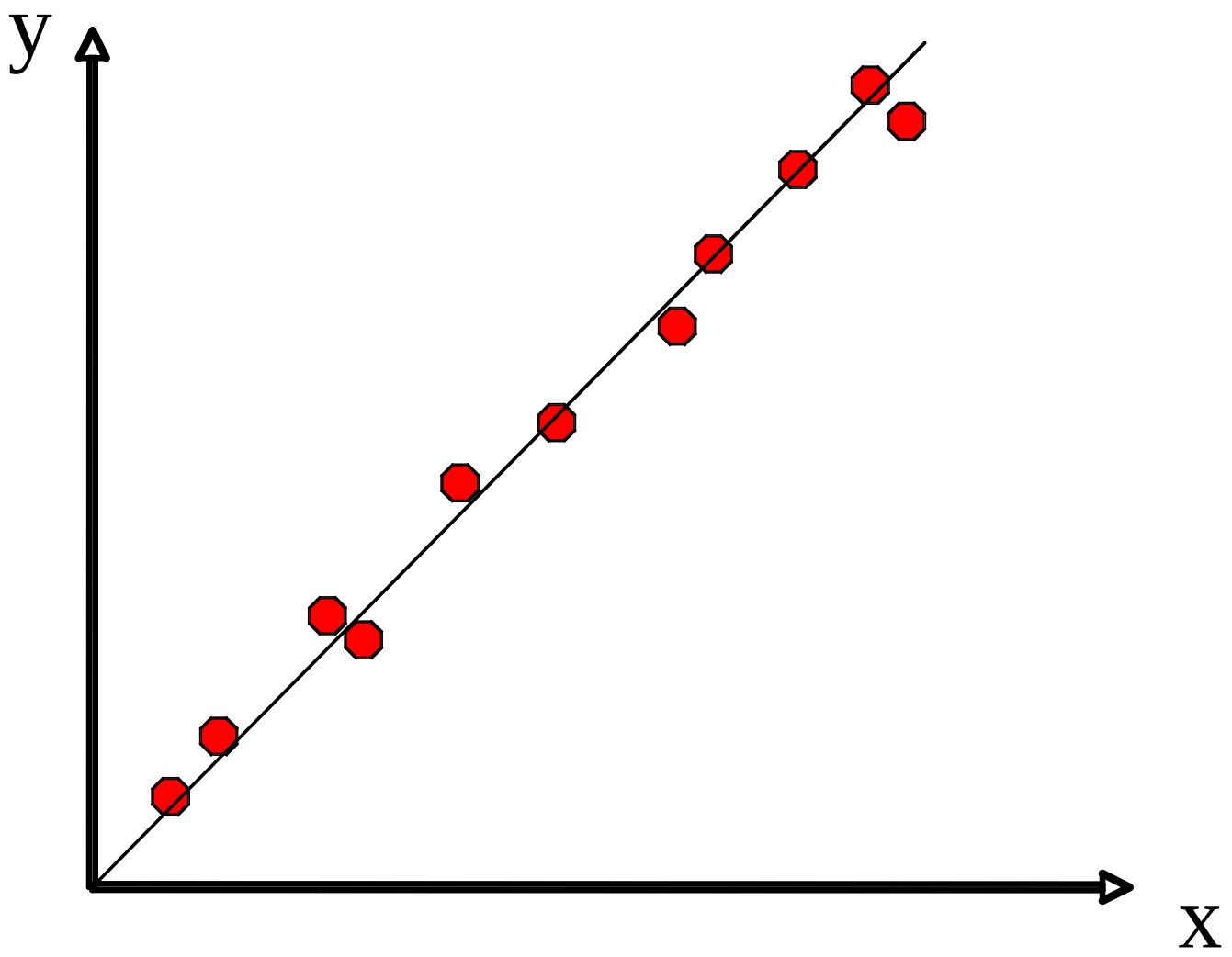
Esempio di output di un'analisi di classificazione "non supervisionata": un dendrogramma di un'analisi di raggruppamento gerarchico (*hyerarchical cluster analysis*)



Esempio di output di un'analisi di classificazione "supervisionata": grafico di dispersione dei campioni di tre classi predeterminate nello spazio definito dalla analisi discriminante lineare (*Linear Discriminant Analysis*)

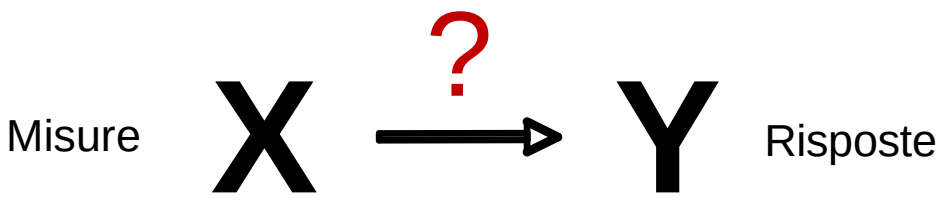
4) **Regressione / Calibrazione**

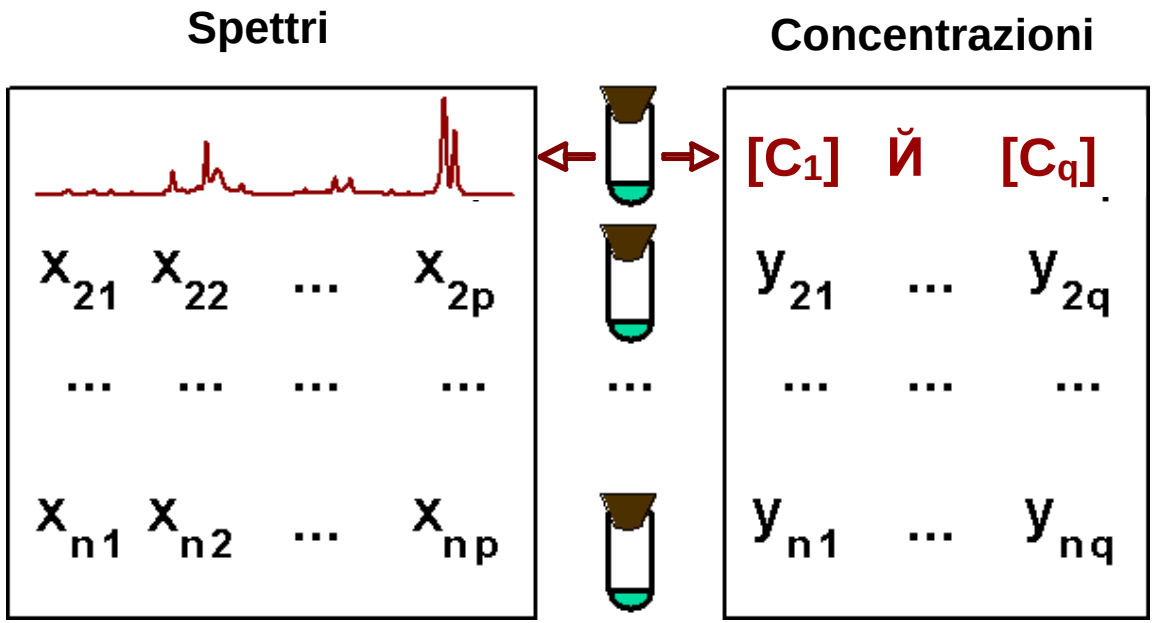
- Due tipi of variabili X / y
- Relazioni lineari / nonlineari
- Modelli
- Analisi diagnostica sulla bontà del modello



Regressione / Calibrazione

Calibrazione multivariata



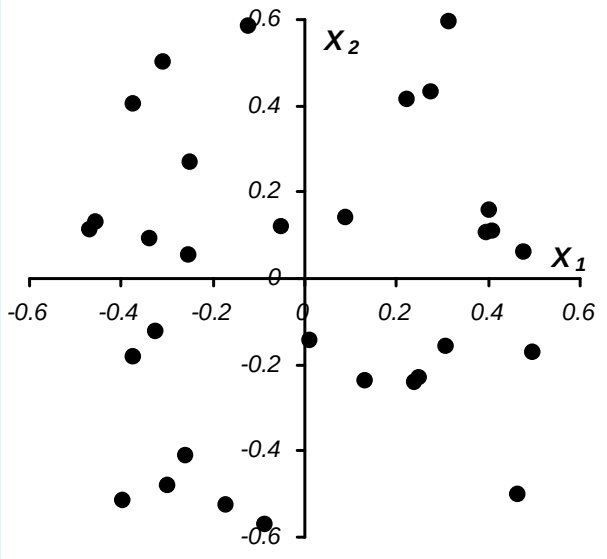


Un “*working horse*” per l’analisi esplorativa, la compressione dell’informazione e la visualizzazione di dati multivariati:

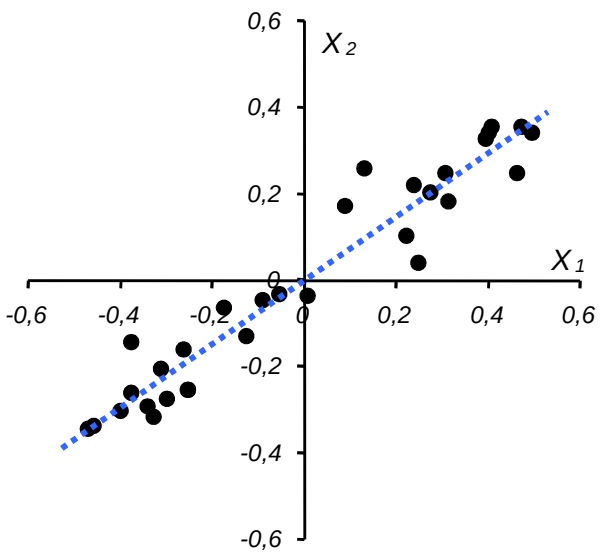
l’Analisi delle Componenti Principali (Principal Component Analysis - PCA)

Metodi di proiezione e variabili latenti

Dati senza struttura



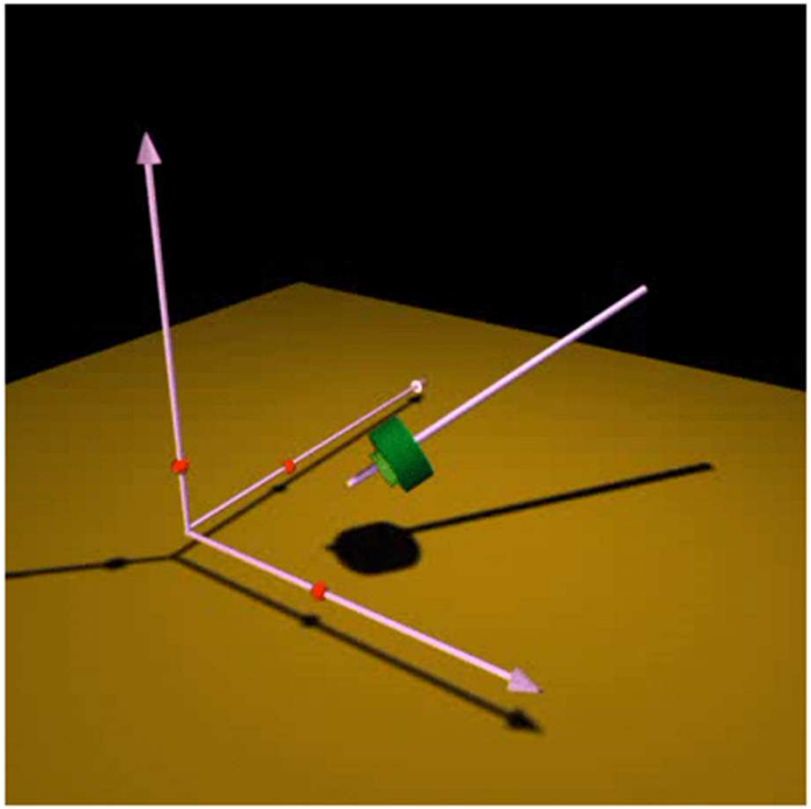
Dati con struttura nascosta



Metodi di proiezione e variabili latenti

Dimensioni formali – numero di variabili

Dimensioni effettive – numero di variabili latenti che coprono tutta la variabilità dei dati



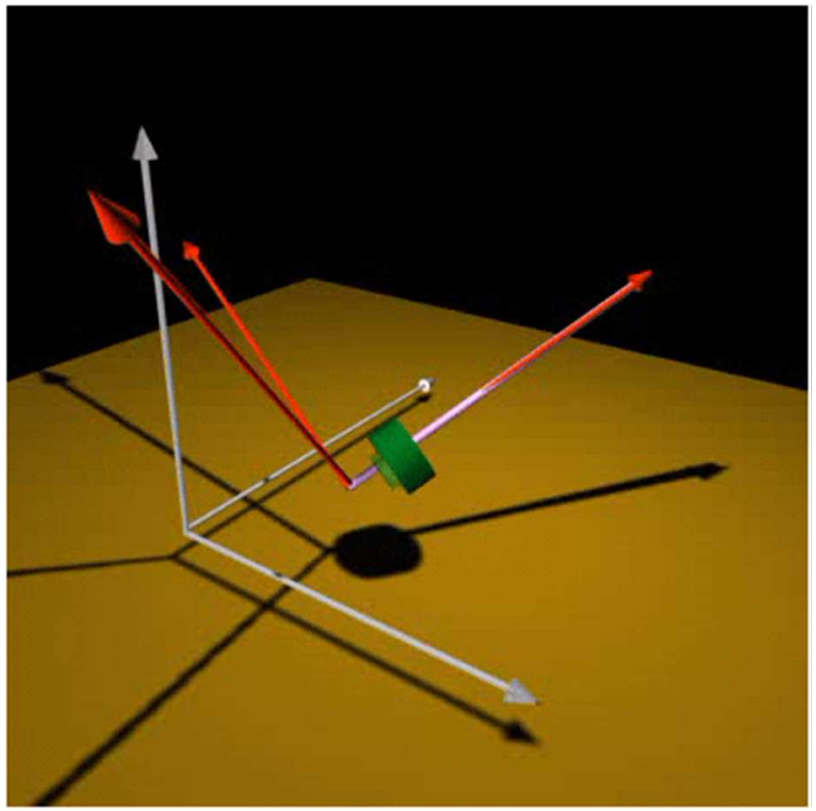
Dimensione formale = **3**

X_1	X_2	X_3
0,121	0,095	0,259
0,834	0,951	0,901
1,548	1,807	1,543
2,261	2,663	2,185
2,974	3,519	2,827
3,687	4,375	3,469
4,401	5,231	4,111
5,114	6,087	4,753
5,827	6,943	5,394

Metodi di proiezione e variabili latenti

Dimensioni formali – numero di variabili

Dimensioni effettive – numero di variabili latenti che coprono tutta la variabilità dei dati



Dimensione effettiva = **1**

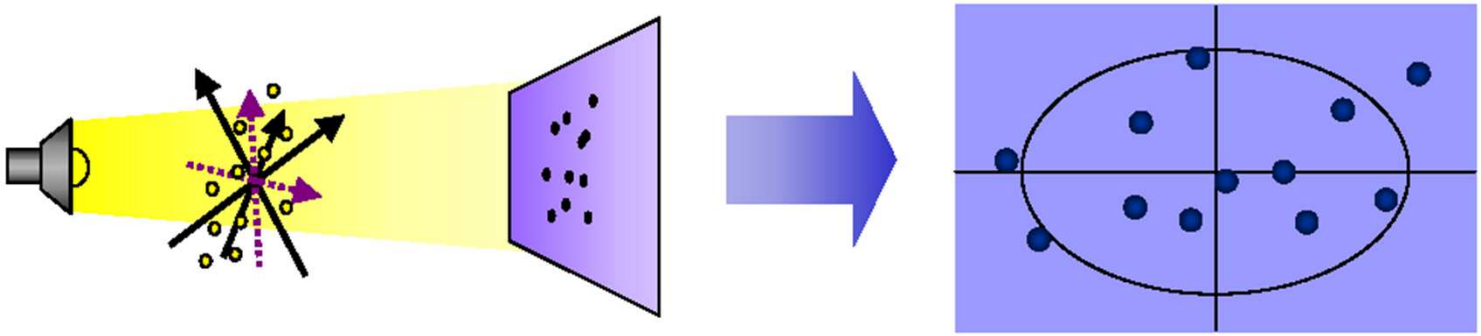
X_1'	X_2'	X_3'
0,1	0,0	0,0
0,2	0,0	0,0
0,3	0,0	0,0
0,4	0,0	0,0
0,5	0,0	0,0
0,6	0,0	0,0
0,7	0,0	0,0
0,8	0,0	0,0
0,9	0,0	0,0

Metodi di proiezione e variabili latenti

Proiezioni nel sottospazio delle variabili latenti

–Consente di ridurre la dimensionalità dei problemi

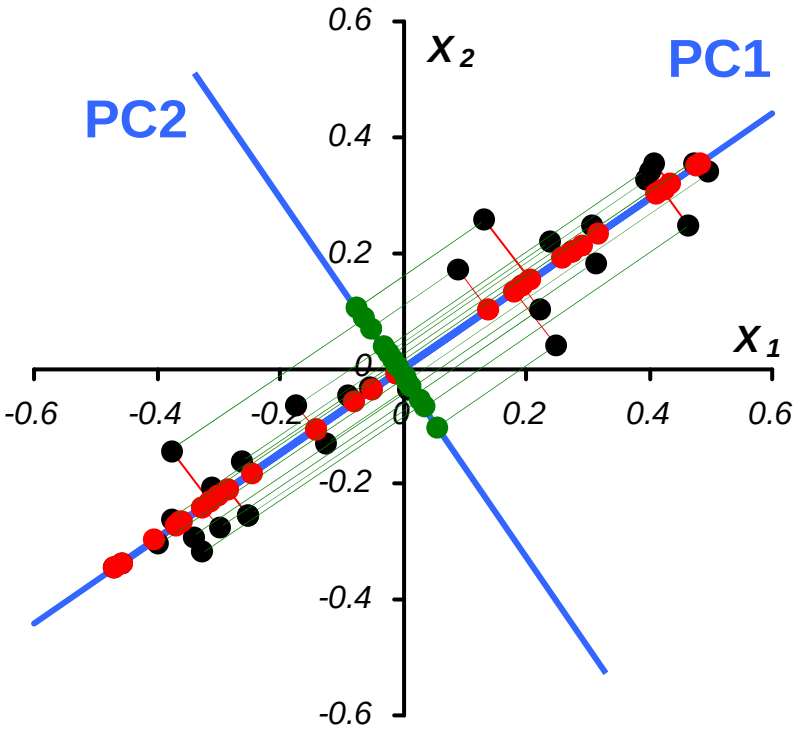
–Fornisce la possibilità di un'analisi visuale dei dati



Come trovare le variabili latenti?

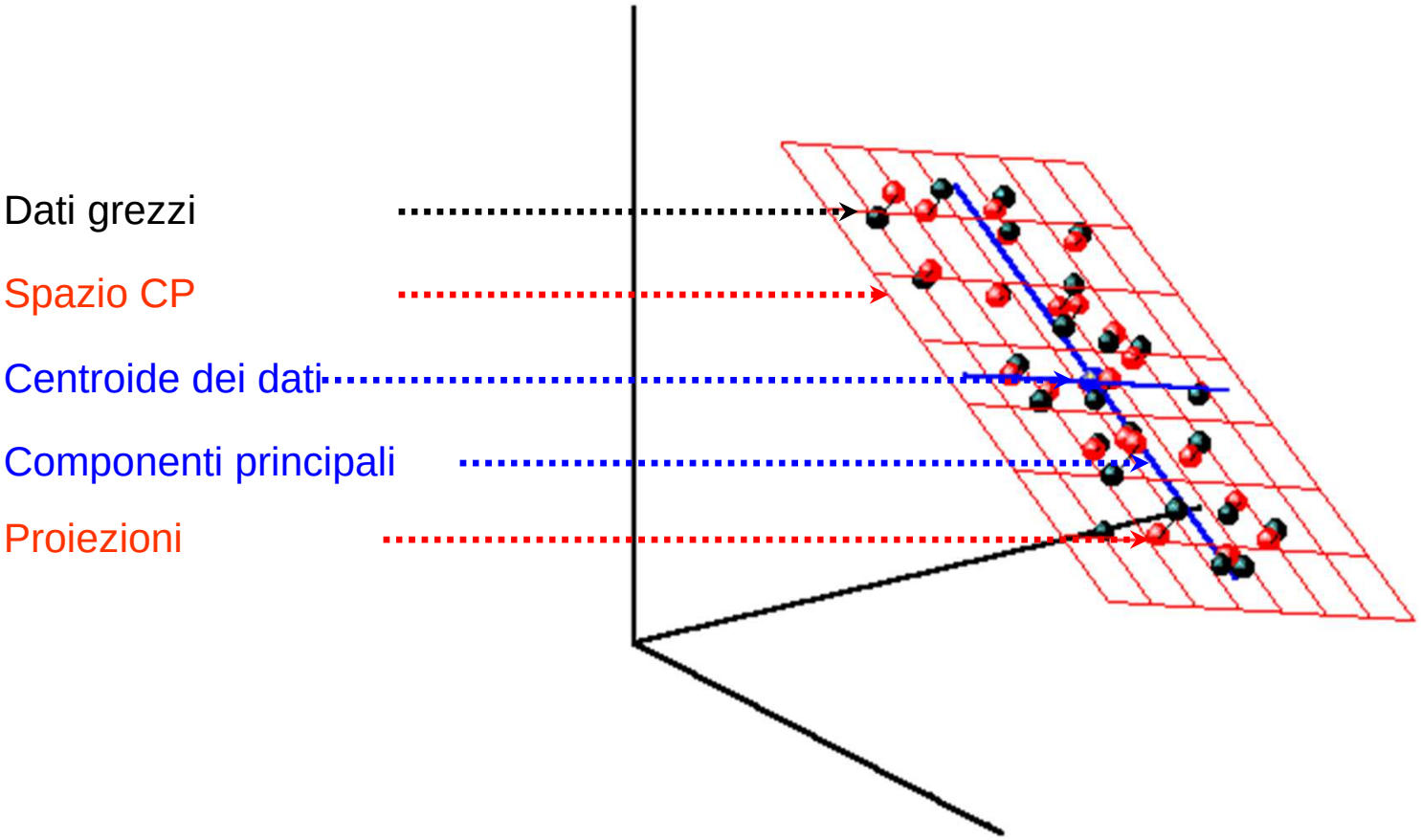
Spazio delle Componenti Principali

- Si individua la variabile latente – (prima **componente principale**, **PC1**) lungo la direzione di massima varianza
- Si proiettano tutti i campioni su **PC1**
- Rimane della **varianza residua**
  - considerata come noise/ rumore (informazione inutile)
  - modellabile con **PC2**

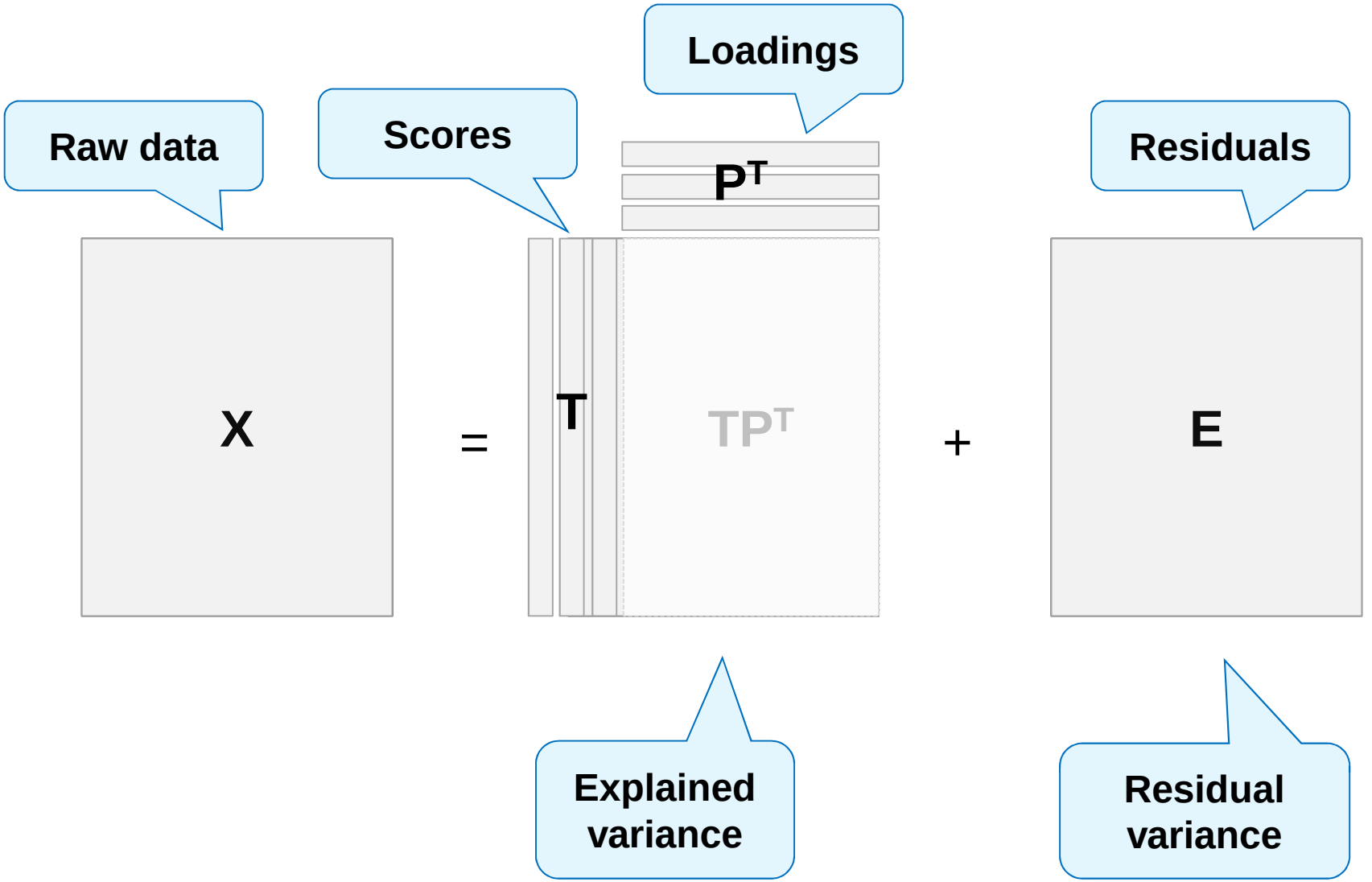


$$X_2 = aX_1 + E$$

Spazio delle Componenti Principali

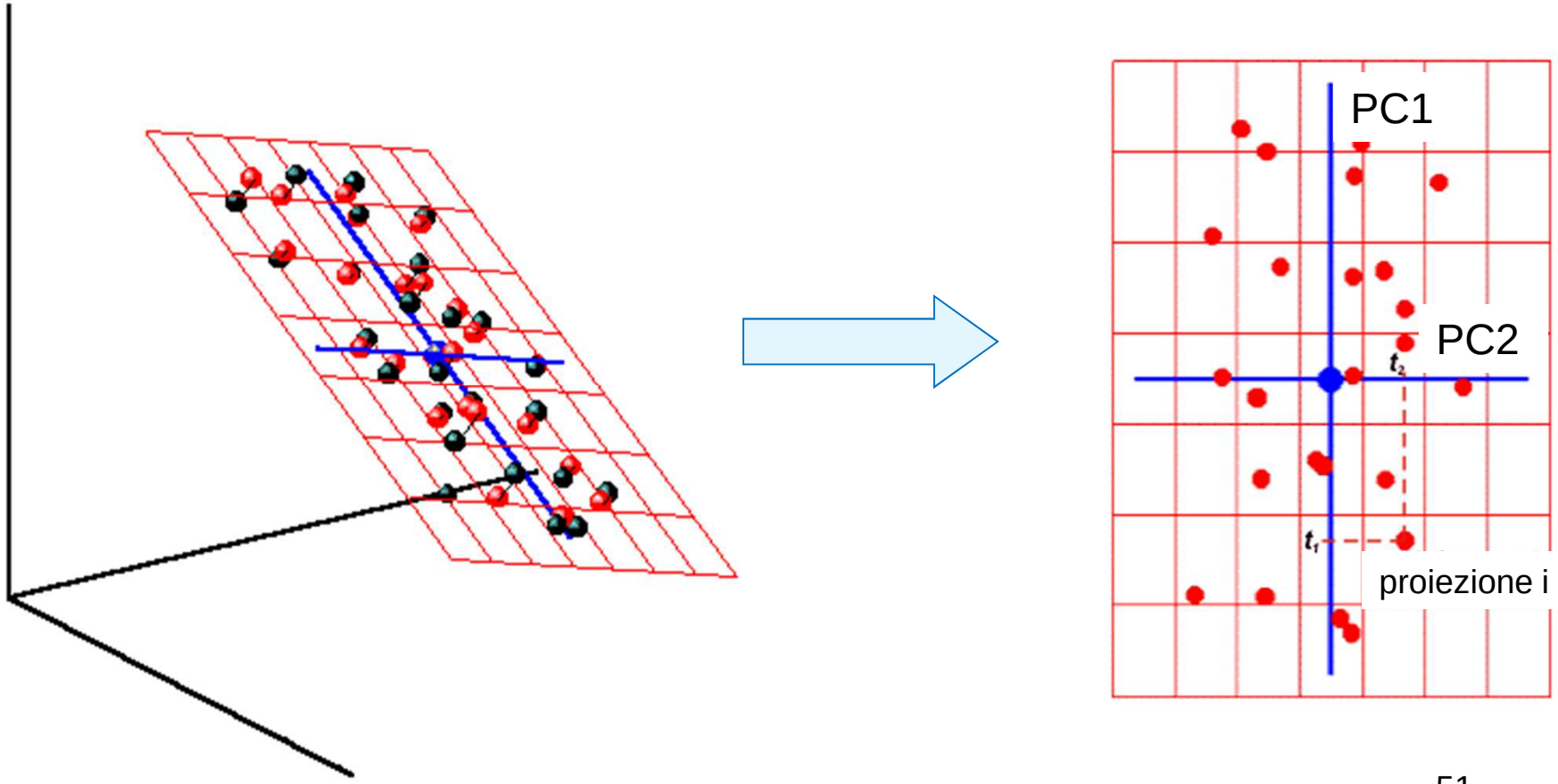


Analisi delle Componenti Principali



Punteggi (*scores*) nella PCA

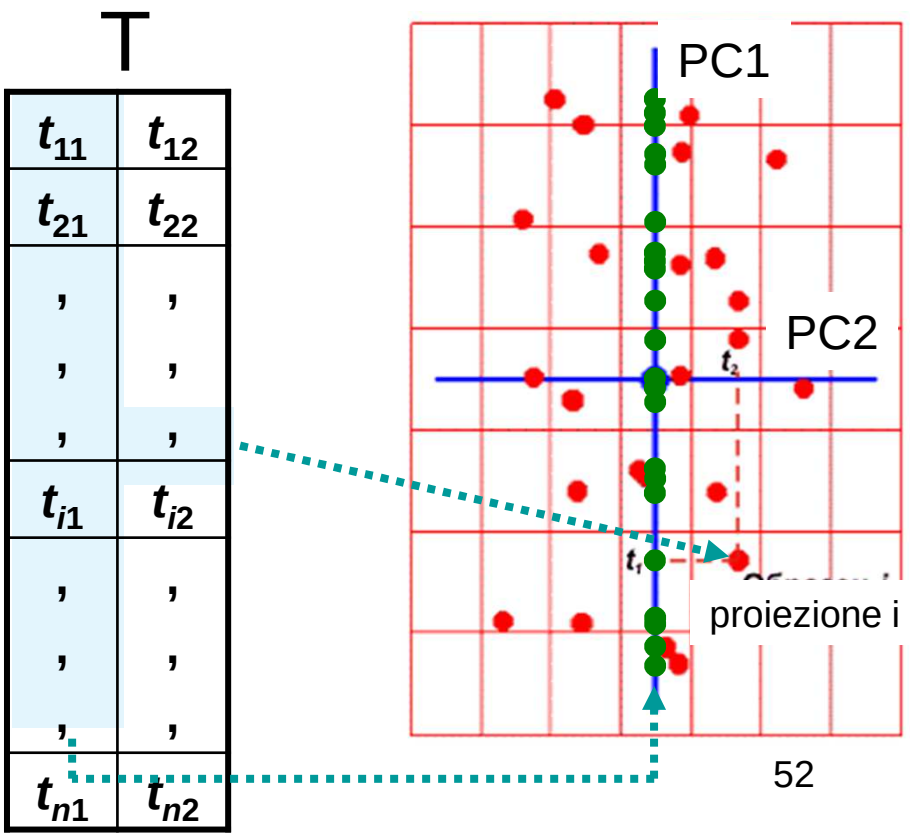
$$X = TP^T + E$$



Punteggi (*scores*) nella PCA

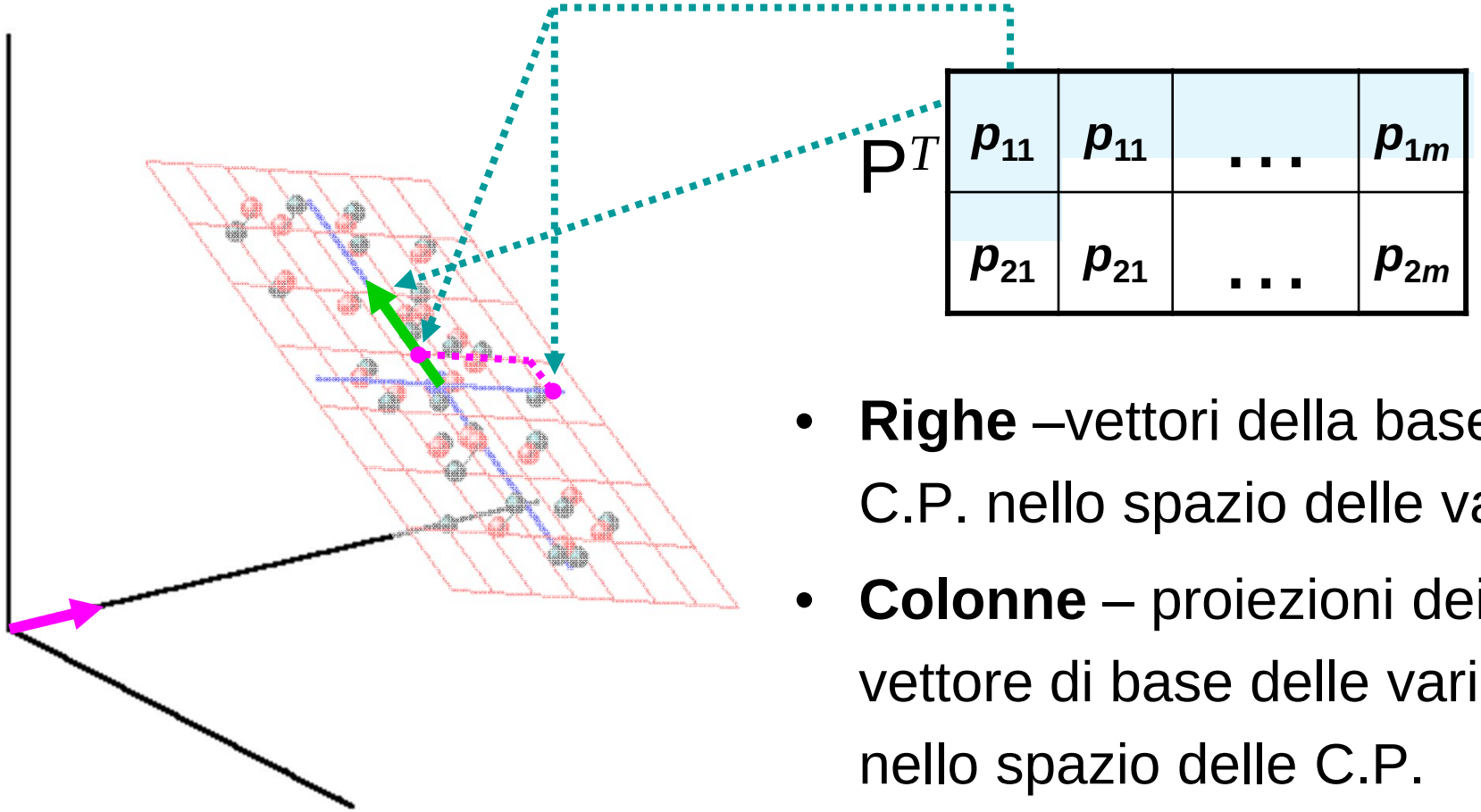
$$X = TP^T + E$$

- **Righe** – coordinate del campione sulle componenti principali
- **Colonne** – proiezioni dei campioni sulla componente principale



Pesi (*loadings*) nella PCA

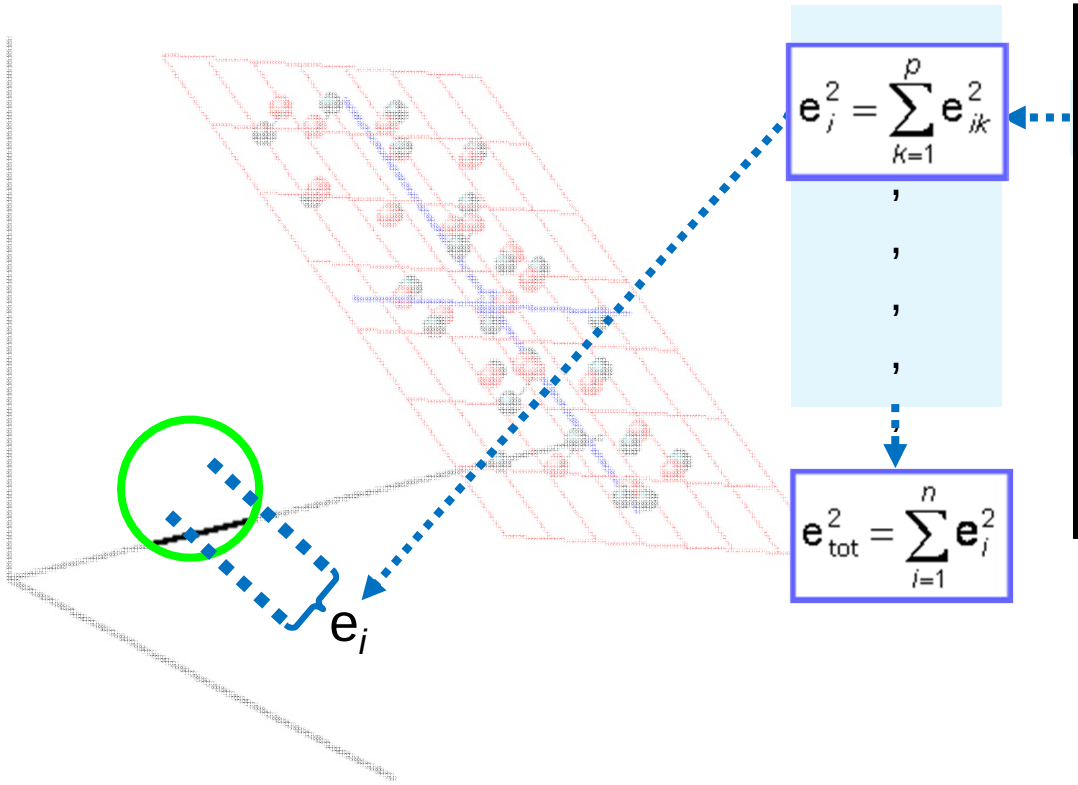
$$X = TP^T + E$$



- **Righe** –vettori della base delle C.P. nello spazio delle variabili
- **Colonne** – proiezioni dei vettore di base delle variabili nello spazio delle C.P.

Matrice E,
variabilità non spiegata dal modello

$$X = TP^T + E$$



e_{11}	e_{12}	…	e_{1p}
e_{21}	e_{22}	…	e_{2p}
, , ,	, , ,	**E**	, , ,
e_{n1}	e_{n2}	…	e_{np}

- e_i – distanza tra il campione e lo spazion delle C.P.
- e^2_{tot} –varianza residua

L’Analisi delle Componenti Principali
è una particolare tecnica di Analisi Fattoriale

Variabili e fattori

- L'analisi fattoriale ha come obiettivo principale l'individuazione di pochi costrutti fattoriali in grado di sostituire un insieme di numerose variabili.
- I costrutti fattoriali vengono considerati, a loro volta, nuove variabili suscettibili di una appropriata interpretazione

Caratteristiche comuni ai metodi di estrazione dei fattori

- Si estraggono fattori (comunque fino a un massimo pari al numero di variabili) finché si ritiene che la varianza spiegata sia sufficientemente grande rispetto alla varianza totale.
- Qualunque metodo di estrazione deve fornire, per ciascuna variabile, un valore numerico (chiamato "saturazione" o "peso") che misuri l'importanza del legame tra variabile e fattore.

Misura del grado di rappresentatività dei fattori rispetto a una variabile

- Per ciascuna variabile è possibile calcolare la "comunalità", cioè la somma dei quadrati dei "pesi" dei fattori.
- La "comunalità" assume valori compresi fra 0 e 1: se è uguale a 1, la variabile può essere esattamente determinata dalla combinazione lineare dei fattori.
- L'interpretazione dei risultati viene facilitata dalla "rotazione" della tabella dei "pesi", operazione per la quale sono disponibili diverse tecniche

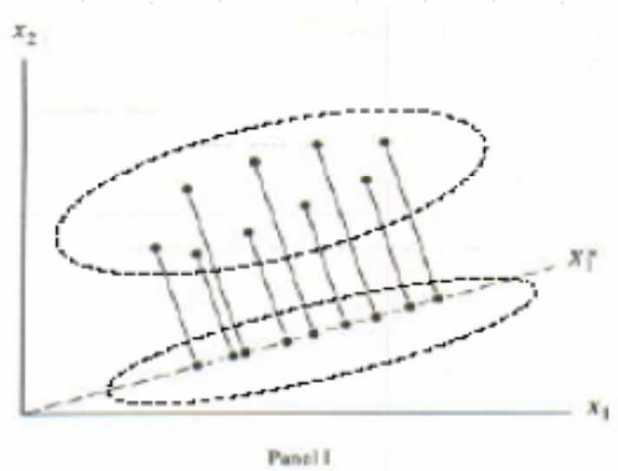
Analisi in componenti principali

L' Analisi in Componenti Principali (ACP) è una tecnica che a partire da un insieme di variabili quantitative (o al più binarie) osservate (originarie) $X_1, X_2, \ldots, X_j, \ldots, X_k$ produce un nuovo insieme di variabili artificiali $Y_1, Y_2, \ldots, Y_p$ ($p \leq k$) dove ciascuna Y_q (q=1,.. p) è una combinazione lineare di $X_1, X_2, \ldots, X_j, \ldots, X_k$

ACP da un punto di vista geometrico

- La matrice dei dati $\mathbf{X}_{n,k}$ è rappresentabile geometricamente come *n* punti in uno spazio R^k, cioè a *k* dimensioni
- Ciascuna riga della matrice $\mathbf{X}_{n,k}$ è chiamata vettore $\mathbf{x}_i=(\mathrm{x}_{i1}, \mathrm{x}_{i2}, \ldots, \mathrm{x}_{ik})$ costituito da *k* elementi numerici che rappresentano le coordinate cartesiane di un punto nello spazio di dimensione *k*
- L'ACP proietta gli *n* punti $\mathbf{x}_i$ rappresentabili nello spazio R^k in un sottospazio R^p di dimensione ridotta p<k individuato in modo tale che la "nuvola" degli *n* punti di R^k sia deformata il meno possibile

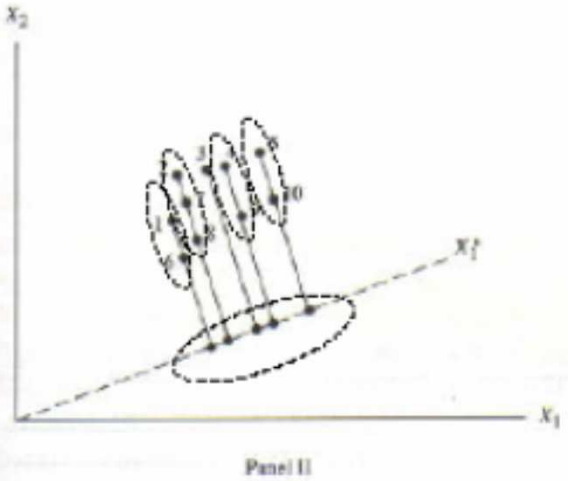
Qualità della riduzione



Panel I

Buona riduzione

Discriminazione tra le unità preservata



Panel II

Pessima riduzione

Formazione di cluster indesiderati

Procedura

Obiettivo: si vuole individuare il sottospazio di dimensione 1, e cioè una retta R^1, tale che la proiezione degli *n* punti $\mathbf{x}_i$ su di essa sia deformata il meno possibile

- Equivale a fare in modo che la varianza (l'inerzia) degli *n* punti $\mathbf{x}_i$ proiettati sulla retta sia la più grande possibile
- Le coordinate degli *n* punti-proiezione $\mathbf{y}_{i1}$ sulla retta costituiscono i valori della variabile artificiale Y_1 costruita dall'ACP: questa variabile è chiamata *prima componente principale* e tali valori sono le "osservazioni" di tale componente principale

Calcolo della prima componente principale

- Prima componente principale

$$y_{i1} = a_{11}x_{i1} + a_{21}x_{i2} + \ldots + a_{k1}x_{ik} = \sum_{j=1}^{k} a_{j1}x_{ij} \qquad i = 1,\ldots, n$$

- Problema: determinare i valori dei coefficienti a_{11}, $a_{21},\ldots,a_{k1}$ in modo tale che la varianza della variabile artificiale (componente principale) Y_1 sia massima e che la somma dei quadrati dei coefficienti sia =1

$$Var\,(Y_1) = \max$$

$$\sum_{j=1}^{k} a_{j1}^{2} = 1$$

Calcolo delle componenti principali successive

- Seconda componente principale

$$y_{i2} = a_{12} x_{i1} + a_{22} x_{i2} + \ldots + a_{k2} x_{ik} = \sum_{j=1}^{k} a_{j2} x_{ij} \qquad i = 1,\ldots, n$$

- Problema: determinare i valori dei coefficienti a_{12}, $a_{22},\ldots,a_{k2}$ in modo tale che la varianza della variabile artificiale (componente principale) Y_2 sia massima, che la somma dei quadrati dei coefficienti sia =1 e che la comp.princ. Y_2 sia incorrelata con la comp.princ. Y_1

$$Var\,(Y_2) = \max$$

$$\sum_{j=1}^{k} a_{j2}^2 = 1$$

$$r(Y_1, Y_2) = 0 \Rightarrow a_{11} a_{12} + a_{21} a_{22} + \ldots + a_{k1} a_{k2} = 0$$

Proprietà delle componenti principali

- Ciascuna componente principale è una combinazione lineare delle variabili originarie
- La 1ª cp spiega il massimo della varianza (inerzia) spiegabile attraverso una riduzione ad una dimensione
- La 1ª componente e la 2ª componente principale spiegano il massimo della varianza spiegabile attraverso una riduzione a due dimensioni

.....

- La 1ª componente, la 2ª componente, ... la *p*-esima componente principale spiegano il massimo della varianza spiegabile attraverso una riduzione a *p* dimensioni

$$Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p)$$

$$r(Y_t, Y_q) = 0 \qquad \forall t, q \text{ tale che } t \neq q$$

Risultati dell'ACP

- La risoluzione del problema di massimo vincolato ad ogni passo porta a trovare che
  - ✓Var(Y_1)= λ_1 primo autovalore della matrice S di var. e covar.
    Coefficienti $a_{11}, a_{21},\ldots,a_{k1}$ sono l'autovettore associato a λ_1
  - ✓Var(Y_2)= λ_2 secondo autovalore della matrice S
    Coefficienti $a_{12}, a_{22},\ldots,a_{k2}$ sono l'autovettore associato a λ_2
  - ✓Var(Y_3)= λ_3 terzo autovalore della matrice S
    Coefficienti $a_{13}, a_{23},\ldots,a_{k3}$ sono l'autovettore associato a λ_3
    .....
  - ✓Var(Y_p)= λ_p p-esimo autovalore della matrice S
    Coefficienti $a_{1p}, a_{2p},\ldots,a_{kp}$ sono l'autovettore associato a λ_p

$$Var(Y_1) \geq Var(Y_2) \geq \ldots \geq Var(Y_p) \Rightarrow \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

La matrice di correlazione R

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}$$

Le variabili vengono numerate da 1 a k, in modo che si potrà indicare con il simbolo r_{ij} il coefficiente di correlazione tra la variabile *i* e la variabile *j*. Data la simmetria del coefficiente di correlazione, si avrà $r_{ij} = r_{ji}$ (i, j = 1, …, k)

Risultati dell'ACP applicata a variabili standardizzate

- Se si vuole lavorare su variabili standardizzate allora la procedura descritta deve essere applicata alla matrice **R** di correlazione
- La risoluzione del problema di massimo vincolato applicato alla matrice **R** ad ogni passo porta a trovare che
  - ✓ $\mathrm{Var}(Y_1) = \lambda_1$ primo autovalore della matrice **R** di correlazione
    Coefficienti $a_{11}, a_{21}, \ldots, a_{k1}$ sono l'autovettore associato a λ_1
  - ✓ $\mathrm{Var}(Y_2) = \lambda_2$ secondo autovalore della matrice **R**
    Coefficienti $a_{12}, a_{22}, \ldots, a_{k2}$ sono l'autovettore associato a λ_2
    .....
  - ✓ $\mathrm{Var}(Y_p) = \lambda_p$ p-esimo autovalore della matrice **R**
    Coefficienti $a_{1p}, a_{2p}, \ldots, a_{kp}$ sono l'autovettore associato a λ_p

$$Var(Y_1) \geq Var(Y_2) \geq \ldots \geq Var(Y_p) \Rightarrow \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Le componenti principali sono fra loro incorrelate

Il modello ACP garantisce che le nuove variabili (le componenti principali) $Y_1, Y_2, \ldots, Y_p$ siano tra loro tutte non correlate

$$r(Y_t, Y_q) = 0 \qquad \forall t, q \text{ tale che } t \neq q$$

Obiettivi dell'analisi in componenti principali

- Riduzione delle k variabili osservate in un numero inferiore $p<k$ di nuove variabili sintetiche dette componenti principali tra loro incorrelate tali che spieghino il massimo della varianza totale della nuvola di punti originaria
- Eliminazione della correlazione esistente tra le k variabili originarie osservate sostituendo ad esse le componenti principali che sono incorrelate
- Costruire indici sintetici e variabili sintetiche

Criteri per la scelta del numero di componenti principali

Quante componenti principali scegliere?

- La scelta deve essere fatta in base a
  - ✓Criterio di parsimonia: numero minimo possibile di componenti principali
  - ✓Minima perdita di informazione
  - ✓Minima deformazione nella qualità della rappresentazione

Scelta del numero di componenti principali

- Criteri di scelta
  - ✓Percentuale di varianza (inerzia) spiegata dalle componenti principali almeno superiore al 70%

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_p}{\text{var}\,ianza\ \ totale} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_p}{\sum_{j=1}^{k} \lambda_j}$$

- ✓Analisi della rappresentazione grafica degli autovalori delle componenti in ordine decrescente: si può disegnare una spezzata unendo i punti corrispondenti agli autovalori per individuare più facilmente le componenti davvero importanti
- ✓Componenti principali corrispondenti ad autovalori λ_q il cui valore
  - ✓ è maggiore dell'"inerzia" media $(\lambda_1 + \lambda_2 + \ldots + \lambda_k)/k$, OPPURE
  - ✓ è maggiore di 1 se nell'ACP si considerano le variabili standardizzate (analisi condotta sulla matrice di correlazione)

La matrice dei pesi (factor loading) delle componenti principali

- La *q*-esima componente principale Y_q è definita come la combinazione lineare delle variabili originarie $X_1, X_2, \ldots, X_j, \ldots, X_k$ con coefficienti $a_{1q}, a_{2q}, \ldots a_{jq}, \ldots, a_{kq} \Rightarrow$ il generico coefficiente a_{jq} rappresenta il peso che la variabile X_j ha nella determinazione della c.p. Y_q
  - ✓Valore assoluto di a_{jq} individua l'importanza della variabile X_j nella determinazione e spiegazione della cp Y_q
  - ✓Il segno (positivo o negativo) di a_{jq} fornisce indicazione della relazione esistente tra X_j e Y_q

Calcolo della comunalità di una variabile

Assumendo p componenti principali, la comunalità della variabile X_j è:

$$h_j^2 = (a_{j1}\sqrt{\lambda_1})^2 + (a_{j2}\sqrt{\lambda_2})^2 + \ldots (a_{jp}\sqrt{\lambda_p})^2$$

$$j = 1, 2, \ldots, k$$

La comunalità indica in quale misura le p componenti principali prescelte sono in grado di rappresentare ciascuna delle variabili originali.

Principal component analysis