INTRODUZIONE ALLA ANALISI DEI DATI SPERIMENTALI

Roma, Giugno 2005

DRAFT VERSION

Indice

1

I DATI

I dati sono informazioni elementari che descrivono alcuni aspetti particolari di un fenomeno. Ad esempio se consideriamo un individuo possiamo identificare alcuni dati che ne descrivono caratteristiche particolari come altezza, peso, colore pelle, concentrazione dei composti chimici nel sangue, composizione DNA, taglia abiti e calzature,…
Di per se un dato non ha significato, ad esempio, nessuno dei dati precedenti rappresenta una informazione significativa in quanto non permette do conoscere l'individuo al quale si riferisce. Affinchè un dato possa aumentare la "conoscenza" su di un fenomeno è necessaria una forma di analisi in grado di collegare il dato con qualche aspetto "significativo" del fenomeno stesso. Nel caso precedente per dare senso alla composizione chimica del sangue è necessario un modello del corpo umano e delle azioni delle patologie. Il processo mostrato in figura 1.1 è quindi il percorso necessario per trasformare un dato in un elemento di conoscenza

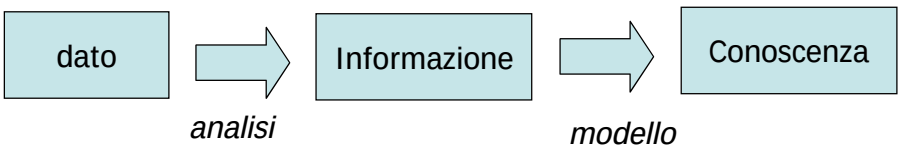


Figura 1.1

I due passaggi importanti di questo processo sono l'analisi che estrae l'informazione dal dato grezzo ed il modello che consente di includere l'informazione in un contesto interpretativo che ne definisce il significato e ne stabilisce la correlazione con le altre informazioni contribuendo, in questo modo, alla conoscenza del fenomeno.
Da un punto di vista generale i dati possono essere classificati in due categorie: *dati qualitativi* e *dati quantitativi*.
I dati quantitativi (anche detti "hard data") sono espressi da un valore numerico ed unità di misura. Ad esempio: la temperatura dell'acqua è 400.0 K.
E' importante sottolineare come i dati quantitativi sono la base della scienza galileiana e delle cosiddette "hard sciences" cioè quelle discipline basate su dati rigorosi connessi tra loro da modelli matematici.
Di estrema importanza sono anche i dati qualitativi ("soft data"). A questa categoria vanno ascritti le etichette ed i descrittori. I dati qualitativi sono generalmente sono espressi verbalmente, ad esempio: "l'acqua è calda". Questo tipo di dati è difficilmente standardizzabile e riproducibile, ma è importante mettere in evidenza che sono proprio i dati qualitativi ad essere elaborati dal nostro cervello. Infatti, i sensi umani producono dati di tipo qualitativo e tutte le elaborazione e le decisione che continuamente vengono prese, sia incosciamente sia consciamente, dagli esseri viventi sono basate sull'elaborazione di questo genere di dati.
Un'altra distinzione importanti sui dati riguarda la differenza tra dati discreti, quelli espressi in intervallo limitato e con valori pre-definiti, e i dati continui, espressi con valori continui. Le caratteristiche degli strumenti di acquisizione dati, soprattutto quelli basati su conversioni analogico-digitali tendono a rendere discreti anche quei dati che per loro natura sono continui.
Infine, una importante caratteristica dei dati riguarda la dimensionalità. Chiamiamo dati univariati quei dati che esprimono una sola grandezza. Per cui, un dato quantitativo univariato è formato da un solo scalare corredato dalla sua di unità di misura.
Ad esempio sono dati univariati i risultati delle seguenti misure: la misura di una resistenza elettrica è 100KΩ; Il peso di una mela è 80g; La concentrazione di K^+ in un acqua è 1.02 mg/l.

Invece, si ha un dato multivariato quando l'applicazione di una misura ad un campione produce una sequenza ordinata di grandezze univariate il cui l'ordine è relativo al significato fisico della misura stessa.
Sorgenti di dati multivariati sono ad esempio quegli strumenti che forniscono sequenze ordinate di valori (chiamati spettri) come gli spettrofotometri ed i gas-cromatografi. Allo stesso modo abbiamo un dato multivariato quando un fenomeno è descritto da un insieme di descrittori o attributi.

1.1 Rappresentazione analitica dei dati

La analisi dei dati è la applicazione delle tecniche e dei concetti della matematica e della statistica ai dati sperimentali, cioè a quelle particolari grandezze che derivano da misure strumentali. Il concetto fondamentale della analisi dati e la rappresentazione delle misure in spazi vettoriali, generalmente euclidei. In questa rappresentazione ad ogni osservabile (detto anche variabile, grandezza misurata o osservabile) viene fatta corrispondere una dimensione dello spazio ed è quindi associata ad un vettore di base. Il sistema di riferimento dello spazio di rappresentazione è perciò formato da una base di vettori ortonormali il cui numero è pari al numero degli osservabili. Date due grandezze univariate, il concetto è ovvio. Consideriamo ad esempio un sensore amperometrico di glucosio. Questo è un dispositivo che converte la concentrazione di glucosio in un liquido in una corrente elettrica. In questo caso si hanno due variabili univariate (concentrazione e corrente) che sono messe in relazione dall'azione del sensore per cui ad ogni valore di concentrazione corrisponde una corrente, e viceversa. Possiamo formare uno spazio vettoriale attraverso il prodotto cartesiano delle due grandezze univariate, questo conduce al piano cartesiano dove rappresentiamo la curva di risposta del sensore stesso. Il processo ora descritto è riassunto in figura 1.2.

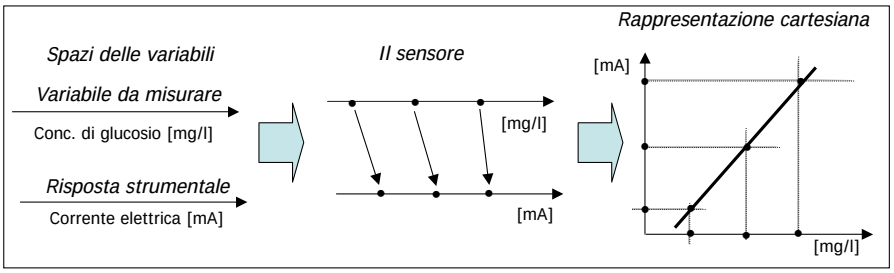


Figura 1.2

In questo caso attraverso l'analisi dei dati è possibile determinare la relazione funzionale che lega la corrente del sensore e la concentrazione di glucoso. Tale relazione si esprime nel luogo di punti (nell'esempio la retta) che definisce il comportamento del sensore.
Nel caso di dati multivariati, il numero delle variabili coinvolte è maggiore di due e lo spazio di rappresentazione diventa uno spazio multidimensionale.
Supponiamo ad esempio che nel caso precedente il sensore sia sensibile oltre che alla concentrazione di glucosio anche al pH della soluzione. In questo caso (figura 1.3) la corrente viene ad essere funzione di entrambe le grandezze, lo spazio cartesiano di rappresentazione ha tre dimensione e il luogo di punti che definisce l'effetto del sensore non è più una retta ma una superficie bidimensionale.

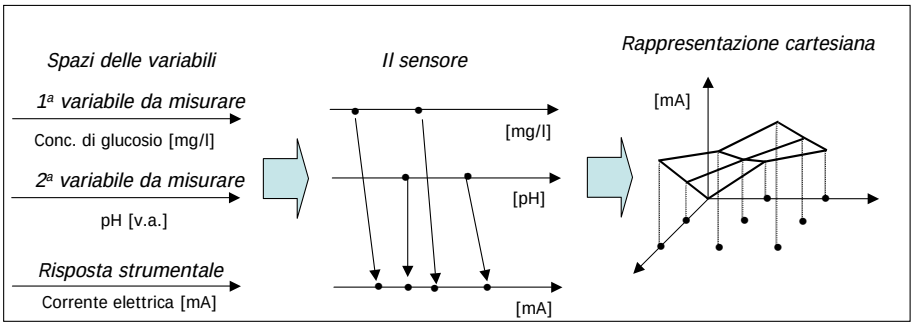


figura 1.3

1. 2 Calibrazione e regressione

Il problema generale dell'analisi dati consiste nella estrazione da una misura strumentale di informazioni riguardanti un campione (o fenomeno) misurato. Generalmente, questa informazione viene estratta dal risultato di uno strumento di misura (sensore). In pratica, il sensore fornisce una quantità osservabile dalla quale, attraverso la conoscenza della modalità di funzionamento del sensore, possiamo ricavare il valore di quella quantità che si vuole conoscere sebbene non sia direttamente osservabile.
Considerando un metodo di misura univariato in cui la risposta dello strumento dipende linearmente dalla grandezza misurata, si ha la seguente relazione

$$y = k \cdot x$$

dove y è la risposta dello strumento (osservabile), x è la grandezza da misurare e k è la caratteristica dello strumento (sensibilità per il caso lineare).
Lo scopo dello strumento è quello di permettere di conoscere la grandezza x. Quindi il problema generale è dato y come posso ricavare x? Ovviamente attraverso la conoscenza di k; Ma come viene ricavato k ? attraverso un processo chiamato *calibrazione*.
Calibrare lo strumento vuol dire esporlo a sollecitazioni (x) note, per cui misurando l'output y posso ricavare il valore di k e quindi rendere lo strumento utilizzabile. La calibrazione è un processo fondamentale per ogni strumentazione e coinvolge sia l'aspetto sperimentale sia l'analisi dati.
L'operazione precedentemente descritta è da un punto di vista matematico banale. Per calcolare k infatti basta disporre di una unica coppia (y,x) e dividere le due quantità.
Il problema in pratica è completamente differente perché ogni misura è affetta da errori. Infatti, ripetendo la stessa misura nelle stesse condizioni si ottengono risultati diversi. Questa proprietà non è relativa alla capacità o meno di eseguire la misura ma ad una proprietà intrinseca del metodo sperimentale: il risultato di ogni misura non è una grandezza deterministica ma una grandezza aleatoria che può essere soddisfacentemente descritta dalla statistica. Per questo il problema della calibrazione non può essere risolto semplicemente considerando una coppia x,y ma attraverso un processo più complesso detto regressione statistica.

Gli errori di misura sono in pratica la deviazione tra la forma funzionale teorica che lega la risposta y alla sollecitazione x ed i valori sperimentali. Consideriamo ad esempio una strain gauge, sappiamo che la relazione teorica che lega la tensione ai capi di un partitore contenente una strain gauge e la deformazione della gauge è del tipo:

$$V = k \cdot \varepsilon$$

Supponiamo di avere eseguito delle misure di calibrazione e di avere ottenuto il set di dati raffigurato in figura 4:

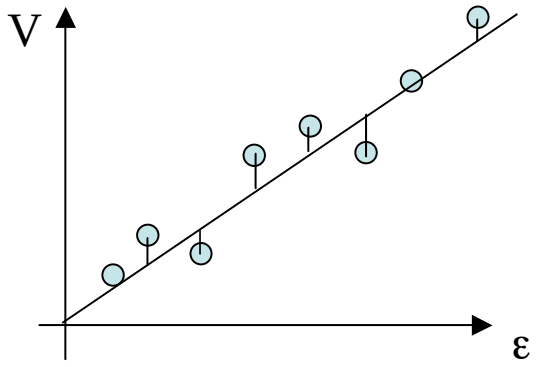


Figura 1.4

Osservandi i dati sperimentali graficati in figura 4 è evidente che per conservare la trattazione teorica della strain gauge è necessario considerare il dato sperimentale come la somma del modello del sensore e di una termine aleatorio, detto errore di misura.
In questo modo, la risposta del sensore (V) sarà composta da un termine deterministico (kε) e da un termine aleatorio (errore) ($V = k \cdot \varepsilon + E$). La parte deterministica della risposta è quella che contiene l'informazione sulla grandezza da misurare. Vedremo che la regressione statistica consente, ponendo determinate condizioni sull'errore, di calcolare dai dati sperimentali la parte deterministica della risposta strumentale.
L'assunzione precedente contiene un importante conseguenza, e cioè che il dato sperimentale, in quanto risultato di una misura, è una grandezza aleatoria. Questo è un concetto fondamentale delle scienze

sperimentali secondo il quale ripetendo N volte "la stessa misura" si ottengono N valori differenti. E' importante considerare che se non si ottengono valori differenti lo strumento di misura ha una risoluzione non adeguata. Si può anzi dire che la risoluzione adeguata di uno strumento di misura è quella che consente di apprezzare la natura statistica della grandezza stessa.
Come esempio consideriamo la sequenza di tabella 1.1 relative a misure di lunghezza, effettuate con differenti strumenti di diversa risoluzione. Al diminuire della risoluzione la misura inizia a fluttuare ed il dato diventa di tipo statistico. Questo consente di definire quando la risoluzione diventa adeguata per la misura.

Strumento	*Risoluzione*	*Misure [cm]*
Metro da sarta	1 cm	120, 120, 120, 120, 120
Metro da falegname	1 mm	119.8, 119.9, 120.1, 120.0, 120.2
Calibro	0.1 mm	119.84, 120.31, 119.83, 120.10, 120. 34
Micrometro	0.01 mm	119.712, 120.032, 119.879, 120.320, 119.982
Interferometro laser	0.5 μm	119.9389, 120.0032, 119.8643, 119.3492, 120.2010

Tabella 1.1

1.3 La legge di Propagazione degli errori

Dalla tabella 1.1 possiamo dedurre che l'errore che deve essere attribuito al risultato finale di una serie di misure della stessa grandezza è pari o all'errore di lettura dello strumento di misura (nel caso in cui le misure siano tutte uguali) o alla semidispersione massima nel caso in cui le misure presentino delle fluttuazioni attorno ad un valore medio.
Questi errori sono detti errori massimi poiché si ha la pratica certezza che ripetendo la misura si ottiene un valore compreso nell'intervallo dell'errore.
Determiniamo ora qual è l'errore da attribuire ad una grandezza G ottenuta per via indiretta, cioè come funzione di misure di altre grandezze $g_1, g_2,\ldots,g_n$.
Poiché il differenziale totale della funzione G rappresenta la variazione infintesima che la $G(g_1, g_2,\ldots,g_n)$ subisce al variare delle singole variabili, facendo l'ipotesi che il differenziale sia estensibile alle variazioni finite e considerando le variazioni dg_i come gli errori massimi da cui sono affette le grandezze $g_1, g_2,\ldots,g_n$, si ottiene l'espressione per l'errore di G. Tale legge è nota come legge di propagazione degli errori:

$$\Delta G = \left|\frac{\partial G}{\partial g_1}\right| \cdot \Delta g_1 + \left|\frac{\partial G}{\partial g_2}\right| \cdot \Delta g_2 + \ldots + \left|\frac{\partial G}{\partial g_n}\right| \cdot \Delta g_n$$

Il valore assoluto delle derivate definisce ΔG come errore massimo, in quanto si ottiene quando i singoli contributi non si compensano ma si somma tra loro.

Esempio: misura della densità media di un cubo attraverso la misura della massa e del lato l.

$$\rho = \frac{m}{V} = \frac{m}{l^3}$$

$$\Delta\rho = \left|\frac{\partial\rho}{\partial m}\right| \cdot \Delta m + \left|\frac{\partial\rho}{\partial l}\right| \cdot \Delta l \Rightarrow \Delta\rho = \frac{1}{l^3} \cdot \Delta m + 3\frac{m}{l^4} \cdot \Delta l$$

$$esempio: \quad se \quad m = \left(1 \pm 0.1\right)g \quad e \quad l = \left(1 \pm 0.1\right)cm$$

$$\rho = 1\frac{g}{cm^3} \quad \Delta\rho = 0.4\frac{g}{cm^3}$$

2

CENNI DI STATISTICA

La statistica è quella disciplina della matematica che studia le grandezze rappresentate da distribuzioni di probabilità. L'oggetto dello studio della statistica riguarda la descrizione e la relazione tra grandezze non deterministiche ovvero tra grandezze per cui il valore con cui si manifestano non può essere predetto in modo assoluto. Perciò le grandezze statistiche hanno la caratteristica di presentarsi sempre con valori differenti. Una delle applicazioni più importanti della statistica riguarda proprio l'analisi dei dati sperimentali in quanto come abbiamo descritto nel paragrafo precedente se misurata con uno strumento opportuno, una grandezza misurabile si presenta sempre con valori differenti.

2.1 Probabilità e Densità di Probabilità

Nonostante la variabilità delle grandezze statistiche, il valore che esse assumono è comunque descrivibile attraverso delle quantità opportune. La più importante di queste è la probabilità.
La probabilità è un concetto fondamentale della statistica. Essa è definita a priori come il rapporto tra i casi favorevoli e tutti i casi possibili. Nel caso classico del lancio di un dado, la probabilità di ottenere, ad esempio un sei, è a pari al rapporto tra uno (solo una faccia del dado contiene il valore 6) e sei (il numero delle facce del dado). Tale rapporto vale $1/6 \approx 0.16$. per cui si dice che la probabilità

di ottenere il valore di sei (o di qualunque altro valore) lanciando un dado con sei facce è all'incirca del 16%.
La probabilità, precedentemente definita, è calcolabile a-priori senza la necessità di lanciare il dado. Per poter calcolare a priori la probabilità di un fenomeno è però necessario conoscere le modalità con le quali il fenomeno stesso si presenta. In generale, è dai dati sperimentali che ricaviamo informazioni sui fenomeni studiati e possiamo definire dei modelli teorici che consentono ad esempio di calcolare la probabilità a priori. Per cui la definizione data sopra di probabilità è applicabile solo dopo aver studiato sperimentalmente un fenomeno. Uno dei criteri generali della analisi dati riguarda il fatto che la conoscenza di un fenomeno è tanto migliore quanto più ampia è la conoscenza sperimentale riguardo al fenomeno stesso. Le grandezze estratte dai dati sperimentali vengono dette stime delle grandezze reali. E' legge generale che all'aumentare dei dati sperimentali la stima approssima sempre più il valore reale.
Lo stimatore dai dati sperimentali della probabilità è la frequenza. Dato un insieme di N misure, la frequenza è definita come il rapporto tra gli N casi favorevoli osservati e le N misure totali. Per N che tende all'infinito la frequenza converge al valore vero della probabilità.

$$Frequenza = \frac{N_{favorevoli}}{N_{totali}}$$

$$\Pr obabilità = \lim_{N_{totali} \to \infty} Frequenza$$

Nel caso in cui la grandezza in oggetto possa assumere una molteplicità di valori, supponiamo continui, come nel caso della misura di una grandezza fisica. Esiste un numero infinito di valori di probabilità, ognuna relativa ad un differente valore della grandezza stessa. L'insieme delle probabilità per una grandezza continua è la funzione densità di probabilità (PDF: Probability Density Function).
La PDF è una funzione estesa a tutti i valori possibili della variabile in oggetto, in generale quindi la PDF è estesa da $-\infty$ a $+\infty$.
In pratica, ogni misura definisce un intervallo attorno al valore misurato, tale intervallo è espresso dall'errore massimo del metodo di misura definito nel precedente capitolo. Per cui il risultato della misura è espresso come $x \pm \Delta x$. La probabilità di tale evento sarà quindi la somma di tutte le probabilità dei valori compresi

nell'intervallo x±Δx. Tale quantità corrisponde all'integrale della PDF nell'intervallo x±Δx:

$$p = \int_{x-\Delta x}^{x+\Delta x} PDF(x) \cdot dx$$

La condizione di esistenza della misura stessa si esprime con il fatto che sicuramente il valore di x della misura si trova nel range dei valori possibili, quindi in generale tra -∞ e +∞. Poiché statisticamente il concetto di sicurezza significa probabilità unitaria si ottiene la seguente regola di normalizzazione della PDF.

$$\int_{-\infty}^{+\infty} PDF(x) \cdot dx = 1$$

La PDF è caratterizzata da descrittori che intervengono come parametri della funzione stessa. Tali parametri descrivono proprietà importanti della funzione di distribuzione e quindi della grandezza che questa rappresenta.
Il più importante di tali descrittore è il *valore medio*. Il valore medio (m) di una grandezza x è quel particolare valore di x per il quale si ottiene:

$$\int_{-\infty}^{m} PDF(x) \cdot dx = \int_{m}^{+\infty} PDF(x) \cdot dx$$

In pratica il valore medio è il valore centrale della probabilità, per cui la probabilità di avere una misura di x maggiore di x eguaglia la probabilità di osservare una misura minore.
Il valor medio coincide inoltre con il valore aspettato della grandezza x, cioè con il valore di x dato solo dalla natura deterministica della grandezza.
Un altro parametro importante è la varianza (σ^2). La varianza definisce l'ampiezza della PDF attorno al valore medio. In pratica definisce il range di x per il quale la probabilità assume un valore

significativo. Vedremo poi come questa "significatività" possa essere quantificata per i differenti tipi di PDF.
In pratica, nonostante la PDF sia definita sull'intero asse reale, il valore finito di varianza definisce un range attorno al valore medio dove è "ragionevole" aspettarsi il valore della grandezza x. In altre parole, dove si hanno i valori maggiori di probabilità.
La varianza è definita come il valore aspettato del quadrato dello scarto. Dove lo scarto esprime la differenza tra x ed il valore medio. Vedremo in seguito come questa quantità giochi un ruolo fondamentale nella analisi dati.
La varianza è definita attraverso la PDF come:

$$\sigma^2 = \int_{-\infty}^{+\infty} \left(x - m\right)^2 \cdot PDF\left(x\right) dx$$

Queste definizioni di media e varianza dipendono ovviamente dalla PDF, sono quindi calcolabili solo conoscendo la PDF quindi la probabilità e quindi il fenomeno da cui x discende. Come per la probabilità, media e varianza possono però essere stimate dai dati sperimentali.
Lo stimatore del valor medio è la media aritmetica. Al crescere del numero dei dati la media aritmetica converge verso il valore medio:

$$m = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i$$

Per quanto riguarda la varianza, data una sequenza di misure di x, la miglior stima della varianza della PDF che genera le misure è data dalla dispersione quadratica media, ovvero dalla medie del quadrato della differenza tra le singole misure ed il valore medio.

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2$$

Si noti che la media è effettuata dividendo per N-1. Questo perché il valor medio di x è stato calcolato dai dati stessi per cui il numero di dati indipendenti è diminuito di 1.
La quantità N-1 è detta gradi di libertà dell'insieme di dati. La radice quadrata della varianza è la deviazione standard (σ). La deviazione standard, che ha la stessa unità di misura della grandezza x, è usata per esprimere l'incertezza della grandezza x stessa.
Data una sequenza di misure, la PDF può essere stimata attraverso il grafico della frequenza in funzione della grandezza x. La frequenza viene generalmente calcolata per intervalli di x, per cui il grafico non è continuo ma è un grafico generalmente a barre. Un grafico di tale genere è detto istogramma. L'istogramma rappresenta la stima della funzione di densità di probabilità.

Consideriamo ad esempio di aver misurato in 31 pesche il valore del grado zuccherino (tale grandezza è detta grado brix). Supponiamo che i valori misurati siano i seguenti: 9.80; 8.00; 10.20; 9.10; 11.30; 12.80; 11.30; 11.20; 9.90; 7.70; 11.50; 11.90; 11.60; 9.20; 8.30; 11.90; 10.40; 9.50; 11.20; 13.20; 14.60; 13.70; 13.20; 13.70; 13.20; 13.70; 11.50; 12.50; 14.30; 12.50; 12.10.
Da questa sequenza di valori possiamo stimare i descrittori della distribuzione di probabilità:

$$media = \frac{1}{31}\sum_{i=1}^{31} x_i = 11.45$$

$$\sigma^2 = \frac{1}{30}\sum_{i=1}^{N}\left(x_i - 11.45\right)^2 = 3.51$$

considerando intervalli unitari della grandezza in esame possiamo costruire l'istogramma della frequenza in funzione del °brix (vedi figura 2.1).

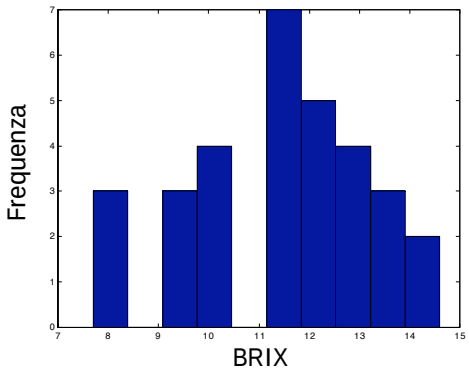


Figura 2.1

L'istogramma fornisce la stima della PDF, dall'istogramma precedente è difficile immaginare una funzione analitica che si adatti all'istogramma sperimentale. Vedremo in seguito che relazione c'è tra il numero di dati sperimentali ed il valore vero della PDF.

2.2 La distribuzione Normale (Gaussiana)

Nonostante esista una moltitudine di funzioni analitiche che rappresentano la PDF di grandezze fisiche, la distribuzione normale (detta anche Gaussiana) è la più importante. Vedremo più in avanti che la distribuzione Normale descrive, sotto opportune condizioni, i dati sperimentali.
Data una grandezza x continua, la distribuzione normale di x dipende dai due parametri precedentemente definiti: il valore medio e la varianza.
La forma funzionale della gaussiana è:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$$

A cui corrisponde la nota curva a campana:

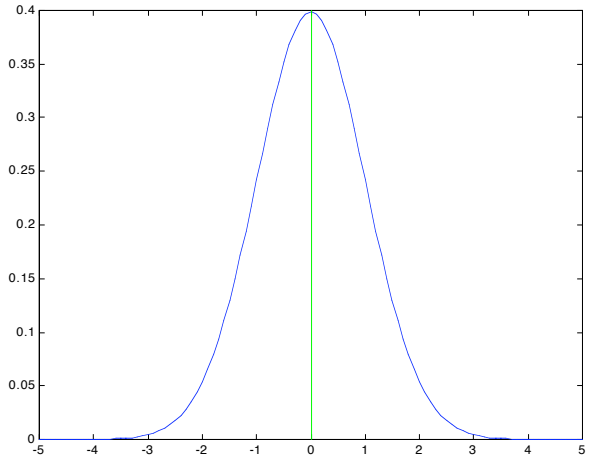


Figura 2.2

Nella distribuzione normale il valore medio coincide con il valore di massima densità di probabilità, mentre la deviazione standard definisce l'ampiezza della campana, cioè l'intervallo di valori di x dove la densità di probabilità è significativamente diversa da zero.
La funzione di Gauss obbedisce alla regola di normalizzazione definita precedentemente. Per cui essendo l'area totale sottesa dalla funzione pari uno, l'altezza della PDF è inversamente proporzionale alla varianza.
La funzione normale descrive in particolare quei fenomeni che derivano dalla somma di un numero molto alto di fenomeni discreti (per questo motivo è importante nella analisi dei dati sperimentali).
Supponiamo infatti di avere un fenomeno discreto che ammette solo 10 possibili risultati (0-9) di uguale probabilità. Dati N misure del fenomeno sia avrà un istogramma piatto. Supponiamo ora di considerare un fenomeno che è la somma di due fenomeni simili a quello appena descritto. Ci saranno 20 possibili valori che però non saranno più isoprobabili. L'istogramma della frequenza inizierà a piccarsi in corrispondenza di un valore centrale. Questo fenomeno spiega perché il gioco dei dadi viene giocato con due dadi per dare luogo ad una distribuzione di probabilità non costante.
Considerando fenomeni che siano somma di un numero crescente di fenomeni discreti elementari, si ottiene un aumento dei valori

possibili e un istogramma che diventa via via più piccato. Nel limite per cui il numero dei fenomeni elementari diventa infinito l'istogramma converge verso la distribuzione normale. Questa caratteristica della distribuzione normale è detta Teorema del Limite Centrale e verrà utilizzata più avanti come fondamento per la teoria degli errori di misura.

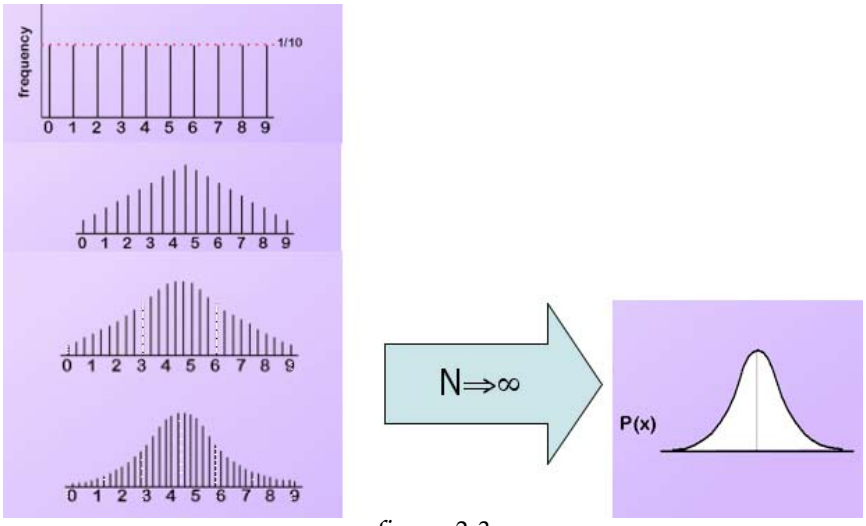


figura 2.3

Proviamo a rispondere ora alla seguente domanda: data una grandezza misurabile con distribuzione gaussiana quante repliche sono necessarie per ottenere una stima ottima della guassiana?
Prendiamo in esame una distribuzione gaussiana con media nulla e varianza unitaria ed "estraiamo" dati da questa distribuzione per ogni set di dati stimiamo la media e la varianza. Questa operazione può essere eseguita con una software di elaborazione matematica come Matlab™.
Calcolando per vari valori di N la media, la varianza e l'istogramma si ottegono i risultati della figura 2.4.

$N_{campioni}$	Media	varianza
3	0.54	0.22
6	0.25	1.58
10	-0.34	0.95
20	-0.39	0.81
30	-0.21	1.08
100	0.00	0.90
1000	0.03	1.04
10000	0.00	0.98

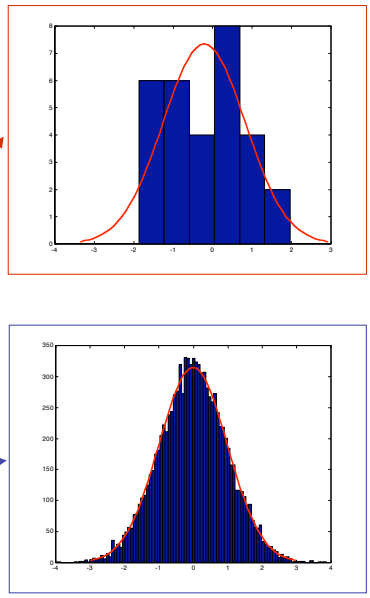


Figura 2.4

Si osservi che per ottenere una stima adeguata di media e varianza (con scarti di 0.00 e 0.02 rispetto ai valori veri) sono necessari 10000 dati! La differenza tra istogramma e PDF è ancora più grande. In figura sono riportati istogramma e PDF per 30 e 10000 dati. Si consideri che dato il carattere aleatorio dell'esperimento al ripetere dello stesso ci si devono attendere conclusioni numericamente differenti.
Possiamo comunque concludere che la stima dei valori veri di una distribuzione è una operazione che richiede un grande impegno sperimentale.
Per facilitare l'uso della distribuzione normale la variabile x viene di solito ridotta ad una variabile con media nulla e varianza unitaria.

$$u = \frac{x - \mu}{\sigma}$$

In questo modo il calcolo della probabilità diventa indipendente dalla particolare natura della variabile e possono essere usate delle tabelle generali. Vedremo più avanti che questo tipo di scalatura dei dei dati è di grande importanza nella analisi dei dati.

Una volta ridotta la variabile la probabilità è indipendente dal significato della variabile stessa. La tabella 2.1 fornisce il valore della probabilità (area) da -∞ a -u che corrisponde a quella dell'intervallo da u a +∞, per questo motivo il valore di u è indicato in tabella in valore assoluto.

$\lvert u \rvert$	*Area*	$\lvert u \rvert$	*area*
0	0.5000	2.0000	0.0228
0.2000	0.4207	2.2000	0.0139
0.4000	0.3446	2.4000	0.0082
0.6000	0.2743	2.6000	0.0047
0.8000	0.2119	2.8000	0.0026
1.0000	0.1587	3.0000	0.0013
1.2000	0.1151	4.0000	$3.2*10^{-5}$
1.4000	0.0808	6.0000	$39.9*10^{-10}$
1.6000	0.0548	8.0000	$6.2*10^{-16}$
1.8000	0.0359	10.000	$7.5*10^{-24}$

Tabella 2.1

Illustriamo ora un paio di esempi pratici sull'utilizzo della distribuzione normale.

Esempi dell'uso della distribuzione normale:

1 La statistica relativa al chilometraggio di una produzione di pneumatici è la seguente
m=58000 Km; s=10000 Km
Quale kilometraggio deve essere garantito per avere meno del 5% di pneumatici da rimpiazzare?
Nella tabella precedente un area di circa 0.05 la si ottiene per u=-1.6. Ovviamente il valore deve essere minore del valore medio. Utilizzando la equazione della variabile ridotta si ottiene:

$$u = \frac{x - m}{\sigma} \Rightarrow x = \sigma \cdot u + m = 10000 \cdot (-1.6) + 58000 \Rightarrow x = 42000\,Km$$

2 Un elettrodo per il monitoraggio del pH viene fornito dal produttore con la seguente statistica relativa al tempo di vita: m=8000 ore; s=200 ore
Se l'elettrodo deve essere sostituito dopo 7200 ore si può affermare che l'elettrodo era difettoso?

Si calcola la variabile ridotta

$$u = \frac{x - m}{\sigma} = \frac{7200 - 8000}{200} = -4.0$$

Nella tabella per tale valore di u si ottiene una probabilita $p=3.2*10^{-5}$. Quindi solo lo 0.0032% degli elettrodi si guasta dopo 7200 ore quindi l'elettrodo era sicuramente difettoso.

Considerazioni sui piccoli campionamenti

Spesso nella raccolta dei dati sperimentali non è possibile raccogliere un numero sufficientemente grande di dati. In questo caso bisogna considerare se il campione raccolto sia effettivamente rappresentativo del fenomeno in esame oppure se sia una sottopopolazione, cioè si siano considerati dei sottoinsiemi del fenomeno. Questo vuol dire che i valori raccolti devono veramente discendere dalla distribuzione statistica che rappresenta il fenomeno.

Una set di dati "mal-campionati" si dice biased.

Nel caso delle distribuzione normale ci sono alcuni strumenti per valutare se una serie di dati sia affetta da bias.

Un metodo consiste nel valutare i seguenti tre descrittori di una sequenza di campioni: la media, la moda e la mediana. La media è stata precedentemente definita, la moda è il valore più frequente nella sequenza di dati e la mediana è il valore centrale. La mediana è un valore della sequenza se il numero di misure è dispari altrimenti è il valore medio dei due valori centrali.

Nel caso della distribuzione gaussiana i tre valori coincidono, per cui questo test fornisce un criterio semplice per decidere se una sequenza di misure è veramente distribuita gaussianamente o se siamo in presenza di un bias.

Un'altra condizione di test per la gaussianità consiste nel misurare la deviazione dalla simmetria della distribuzione dei dati. Questa quantità è espressa da una grandezza detta skewness. Lo skewness è detto anche momento terzo dei dati ed è pari alla media dello scarto cubico:

$$s = \frac{1}{N \cdot \sigma^3} \sum_{i=1}^{N} \left(x - \mu \right)^3$$

Il valore di s indica la asimmetria attorno al valor medio della distribuzione, in particolare, per la distribuzione normale si ha s=0. Esempi sul significato dello skewness sono mostrati in figura 2.3

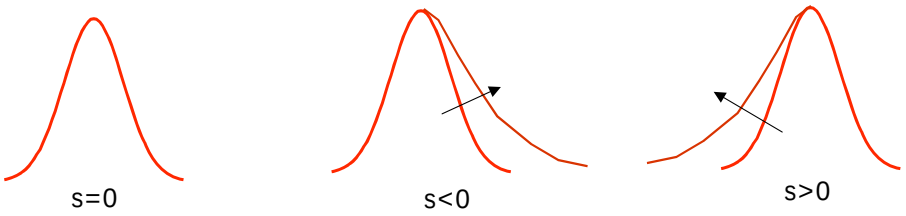


Figura 2.5

2.3 Teoria degli Errori di Misura

Gli errori casuali che affettano qualunque misura e qualunque strumento di misura, quindi qualunque sensore, sono trattati quantitativamente in modo da poter dare un significato non ambiguo sia al risultato di una misura che al suo errore. Il caso particolare che qui interessa è quello in cui ripetendo più volte la stessa misura, apparentemente nelle stesse condizioni, si ottengono risultati diversi.
Nel caso dei sensori è interessante specificare che la misura può solo apparentemente essere eseguita nelle stesse condizioni, cioè a condizioni in cui il misurando non varia, in realtà possono variare altre grandezze non controllate a cui il sensore è seppur in minima parte sensibile.
Si questo capitolo si vedrà che nel caso di fluttuazione della risposta, la media aritmetica delle fluttuazioni è la quantità che meglio rappresenta il risultato della misura; inoltre si vedrà che la bontà di tale rappresentazione è tanto migliore tanto più alto è il numero n delle misure.
Per questo si procederà nel modo seguente:

i tramite la *Legge dei Grandi Numeri* si mostrerà che la media aritmetica di n variabili casuali converge verso il valore aspettato della distribuzione di probabilità della variabile stessa, qualunque sia la distribuzione di probabilità;
ii per mezzo del *Teorema del Limite Centrale* si vedrà che, sotto ipotesi molto generali, la somma di n variabili casuali, di

distribuzione qualsiasi, ha una distribuzione che, al crescere di n, tende alla distribuzione normale;

iii infine facendo l'ipotesi che gli errori di misura siano dovuti a molte cause indipendenti, ciascuna delle quali produce un errore piccolo, applicando il teorema del limite centrale si potrà dedurre come la distribuzione di probabilità degli errori casuali è gaussiana e da questo dedurre che è gaussiana anche la distribuzione di probabilità della media delle misure.

2.4 Teorema del Limite Centrale e Teorema di Gauss

Esistono due teoremi, di non immediata dimostrazione, che rendono conto del perché tanto spesso ci si trovi a che fare con distribuzioni di probabilità che si approssimano molto a distribuzioni gaussiane e che giustificano l'assunto, che verrà fatto in seguito della distribuzione gaussiana degli errori di misura propri delle risposte di sensori.

Il primo di tali teoremi è il teorema del limite centrale che possiamo così enunciare:

Date n variabili casuali, ciascuna delle quali distribuita con legge arbitraria, ma aventi tutte valori aspettati dello stesso ordine di grandezza e varianze finite e dello stesso ordine di grandezza, si ha che la variabile casuale costruita con qualunque combinazione lineare di dette variabili ha una distribuzione che tende a quella normale al crescere di n.

Il secondo teorema detto teorema di Gauss, è una applicazione del teorema del limite centrale al caso degli errori di misura. Esso semplicemente fa l'ipotesi che un errore di misura è il risultato di molte cause indipendenti, ciascuna delle quali produce un errore piccolo, con segno a caso e dello stesso ordine di grandezza di tutti gli altri.

Se cioè l'errore di misura può essere considerato come dato da una combinazione lineare, in particolare della somma, di un numero molto grande di errori piccoli, si può applicare il teorema del limite centrale ed affermare che gli errori di misura seguono essi stessi una distribuzione di probabilità normale. Quindi se l'errore ε, definito come differenza tra valore misurato e valore aspettato (si ricordi ignoto) $\varepsilon = x - m$, è dato dalla somma degli ε_i prodotti dalle n cause indipendenti:

$\varepsilon = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n$

con

$E(\varepsilon_i) = 0 \;\; ; \;\; \sigma^2(\varepsilon_i) = E(\varepsilon_i)^2 \;\; ; \;\; E(\varepsilon) = 0$

da cui

$\sigma^2(\varepsilon) = \sum \sigma^2(\varepsilon_i)$

Allora per il teorema del limite centrale, la distribuzione di ε è:

$$f(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

che può essere riscritta come:

$$f(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

Quest'ultima espressione può essere considerata la distribuzione di x. In conclusione, nelle ipotesi fatte,, le misure seguono una distribuzione normale centrata intorno al valore aspettato m e con varianza uguale a quella dell'errore.

3

REGRESSIONE STATISTICA

In questo capitolo viene affrontato il problema della stima, cioè di ricavare informazioni estraibili da misure sperimentali. In particolare si vedrà come stimare i parametri caratterizzanti la funzione di distribuzione delle misure e i parametri che compaiono in relazioni funzionali che legano più misure tra loro.
Esistono molte soluzioni a questi problemi, qui verranno illustrate le più comunemente utilizzate.

3.1 Metodo della Massima Verosimiglianza (Maximum Likelihood)

Si supponga di aver eseguito n misure, x_i, di una variabile casuale x, e si supponga di voler utilizzare tali misure per stimare i parametri λ_j (j=1…m) che compaiono nella funzione di distribuzione, di forma nota, $f(x; \lambda_1, \lambda_2, \ldots, \lambda_m)$.
Il principio di massima verosimiglianza afferma che, tra i valori possibili per i parametri incogniti, si devono scegliere quelli che rendono massima la probabilità dell'evento osservato. Cioè, con i simboli adottati, si deve assumere per λ_j il valore che rende massima la probabilità di osservare gli n valori x_i effettivamente ottenuti.
Se la variabile x è continua si dovrà effettivamente parlare anziché della probabilità di osservare gli n valori x_i della probabilità che la variabile x assuma valori compresi in n intervalli di ampiezza dx_i intorno ai valori x_i ottenuti. Si avrà allora:

$$dP(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \lambda_1, \ldots, \lambda_m) dx_i$$

Per massimizzare l'evento osservato è allora sufficiente massimizzare il prodotto:

$$\prod_{i=1}^{n} f(x_i; \lambda_1, \ldots, \lambda_m)$$

A questo prodotto si dà il nome di funzione di verosimiglianza: $L(x_i; \lambda_1, \lambda_2, \ldots, \lambda_m)$.
Si osservi che L è da intendersi come funzione delle variabili $\lambda_1, \lambda_2, \ldots, \lambda_m$; i valori x_i ottenuti sperimentalmente sono invece parametri di tale funzione.
Per trovare il valore del parametro λ_j che rende massima L è sufficiente uguagliare a zero la derivata di L rispetto a λ_j e controllare che il valore trovato renda negativa la derivata seconda di L sempre rispetto a λ_j. Poiché quando L è massima anche la funzione Log(L) è massima, si può usare il logaritmo della funzione di verosimiglianza in quanto è più conveniente trattare somme che prodotti. Si ha perciò la seguente equazione:

$$log(L) = \sum_{i=1}^{n} log\left[f(x_i; \lambda_1, \ldots, \lambda_m)\right]$$

$$\frac{\partial\, log(L)}{\partial \lambda_j} = 0 = \sum_{i=1}^{n} \frac{\partial\, log\left[f(x_i; \lambda_1, \ldots, \lambda_m)\right]}{\partial \lambda_j}$$

Evidentemente in tal modo si ottiene una equazione da risolvere per ogni parametro da determinare. Il valore così trovato per ciascun parametro è detto "stima di massima verosimiglianza" di quel parametro.

Applicazione: Media Pesata
Una applicazione particolare e molto importante del principio della massima verosimiglianza si ha nel caso in cui le x_i provengono da

distribuzioni normali caratterizzate tutte dallo stesso valor medio m ma da diverse deviazioni standard σ_j. Questo è ad esempio il caso che si presenta in pratica quando si eseguono più misure di una stessa grandezza con strumenti o sensori di diversa precisione: in questo caso il valore aspettato è sempre la misura della grandezza incognita. Si tratta quindi di stimare m, cioè il valore più attendibile della grandezza in misura, sfruttando nel modo migliore tutte le misure eseguite, comprese le meno precise.
La funzione di verosimiglianza, in questo caso si scrive come:

$$L(x;m) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} exp\left(-\frac{(x_i - m)^2}{2\,\sigma_i}\right)$$

quindi

$$\frac{d\ log(L)}{dm} = \frac{d}{dm}\sum_{i=1}^{n}\frac{(x_i - m)^2}{2\sigma_i^2} = \sum_{i=1}^{n}\frac{x_i - m}{\sigma_i^2} = 0$$

da cui

$$m = \frac{\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}$$

La stima così eseguita del valor medio comune a tutte le distribuzioni considerate prende il nome di media pesata o media ponderale. Si noti che come è da aspettarsi alle misure meno precise viene attribuito un peso , cioè un'importanza, minore.

3.2 Metodo dei Minimi Quadrati (Least Squares)

Il metodo dei minimi quadrati è di grande importanza pratica in quanto consente di ricavare stime di parametri di relazioni funzionali che coinvolgono più grandezze misurabili. Come vedremo questo metodo è ampiamente utilizzato per l'analisi multicomponente di sensori oltre che per ricavare i parametri funzionali dei sensori stessi.
Per illustrate il significato di questo metodo si considerino due grandezze, x ed y, legate da una dipendenza funzionale g di forma nota, del tipo:

$$y = g(x; k_1, \ldots, k_m)$$

ove $k_1, \ldots, k_m$ sono parametri incogniti. Si supponga che siano state eseguite n coppie di misure di x e y e siano esse x_i e y_i. Se le incertezze con le quali sono note x_i e y_i fossero trascurabili, si potrebbero scrivere n equazioni del tipo della * sostituendo di volta in volta ad x ed y i valori sperimentali. Ottenendo così un sistema di n equazioni in m incognite (i parametri $k_1, \ldots, k_m$); l'aver supposto gli errori casuali trascurabili implica che le equazioni ottenute debbano essere compatibili.
Se n=m il sistema è risolubile e le soluzioni sono i valori degli m parametri. Se n>m un qualsiasi sistema di equazioni estratto dalle n disponibili fornisce le m soluzioni mentre le rimanenti m-n equazioni risultano essere delle identità qualora ad esse si sostituiscano i valori dei parametri determinati.
Se n<m, n equazioni forniscono i valori di n parametri incogniti mentre m-n di essi rimangono indeterminati.
Nel caso in cui però x ed y siano i risultati di misure sperimentali è evidente che non si può prescindere dal considerare gli errori ed è quindi necessario seguire un procedimento diverso da quello ora descritto.
Si supponga innanzitutto che gli errori sulle x_i siano percentualmente minori rispetto a quelli sulle y_i, di modo che si possa pensare che ciascuna x_i si riproduca, al ripetersi delle misure, mentre la y_i corrispondente fluttui seguendo una certa legge di distribuzione con valore aspettato $g(x_i; k_1, \ldots, k_m)$ e deviazione standard σ_i. Questa condizione è plausibile ad esempio quando consideriamo la calibrazione di un sensore sottoposto a condizioni ambientali note, è infatti quasi sempre rispettato in questo caso il fatto che l'incertezza sulle condizioni ambientali è molto minore di quella relativa alla risposta del sensore stesso.
Per stimare il valore dei parametri $k_1, \ldots, k_m$ si faccia l'ulteriore ipotesi che le misure y_i siano distribuite normalmente ed applichiamo il principio della massima verosimiglianza.
La funzione di verosimiglianza si scrive come:

$$L(x_i;k_1,\ldots,k_m) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} exp\left[-\frac{\left[y_i - g(x_i;k_1,\ldots,k_m)\right]^2}{2\sigma_i^2}\right]$$

per cui ci si riduce a minimizzare la somma:

$$\sum_{i=1}^{n} \frac{\left[y_i - g(x_i;k_1,\ldots,k_m)\right]^2}{\sigma_i^2}$$

cioè a minimizzare la somma dei quadrati degli scarti, ciascuno pesato con l'inverso della propria varianza.
Tale modo di procedere si può qualitativamente giustificare considerando che in primo luogo è accettabile che i valori più probabili dei parametri si ottengano minimizzando la somma di opportune funzioni dei moduli degli scarti perché questo è il modo algebricamente più semplice per imporre che le differenze tra i valori sperimentali ed i loro valori aspettati siano le più piccole possibili.

Applicazione: Approssimazione (fitting) di dati sperimentali in relazione lineare

Si supponga di avere eseguito un certo numero di misure disposte come in figura 3.1 che sono legate da una relazione lineare y=ax+b per cui si vogliono determinare, a partire dalle misure sperimentali, i valori dei parametri a e b.

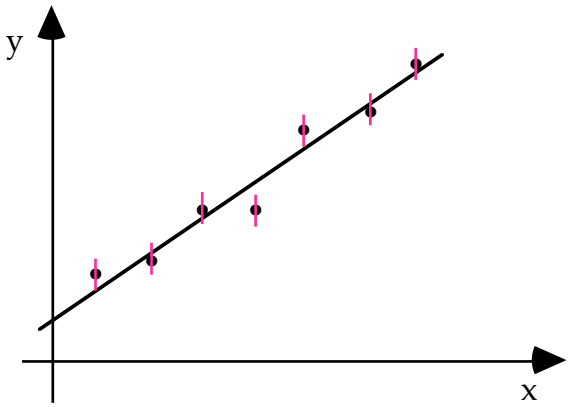


Figura 3.1

Si consideri inoltre che, come mostrato in figura, gli errori sperimentali presenti sulla variabile y siano molto più grandi di quelli presenti sulla variabile x che possiamo considerare assolutamente nota rispetto a y. Per semplicità si consideri inoltre che tutte le deviazioni standard di y siano uguali. I valori dei parametri che minimizzano la somma quadratica degli scarti sono allora dati dalle seguenti equazioni.

$$\frac{\partial \sum \left[y_i - (a x_i + b) \right]^2}{\partial a} = 0$$

$$\frac{\partial \sum \left[y_i - (a x_i + b) \right]^2}{\partial b} = 0$$

Che risolte forniscono le stime dei parametri della relazione lineare:

$$a = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$b = \frac{\overline{y} \sum x_i^2 - \overline{x} \sum x_i y_i}{\sum (x_i - \overline{x})^2}$$

Le deviazioni standard di a e b si ottengono dalla legge di propagazione degli errori, ricordando che per ipotesi l'unica variabile affetta da errore è y e che tutte le σ_i sono uguali.

$$\sigma_a = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial a}{\partial y_i} \right)^2 \sigma_i^2} = \sigma \cdot \sqrt{\sum_{i=1}^{n} \left(\frac{\partial a}{\partial y_i} \right)^2}$$

$$\sigma_b = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial b}{\partial y_i} \right)^2 \sigma_i^2} = \sigma \cdot \sqrt{\sum_{i=1}^{n} \left(\frac{\partial b}{\partial y_i} \right)^2}$$

3.3 Esempi di applicazione del metodo dei minimi quadrati

Calibrazione di un sensore chimico.

Consideriamo un sensore di gas sensibile a vapori di metanolo.
La concentrazione di metanolo viene stimata calcolando la parte volatile di una quantità di metanolo tenuta a temperatura nota e costante. La concentrazione satura è poi miscelata con azoto, da un sistema di erogatori controllati di flusso, per ottenere vari valori di concentrazione. La legge di regressione lineare non contiene l'intercetta.
Verifichiamo le ipotesi dei minimi quadrati

L'errore su y è molto maggiore dell'errore su x	SI	La stabilità della temperatura e le prestazioni dei flussimetri sono sufficienti
Y è distribuita normalmente	≈ SI	Essendo misure sperimentali obbediscono probabilmente al teorema di Gauss
Gli eventi osservati sono indipendenti	SI	Controllare l'assenza di effetto memoria
Le misure hanno varianza uguale	SI	Le misure sono svolte in condizioni simili e sempre con gli stessi sensori

Il risultato della calibrazione è mostrato in figura 3.2. Si osservi che il coefficiente del sensore è la sensibilità e che dagli errori di misura si può stimare il limite di rivelazione del dispositivo.

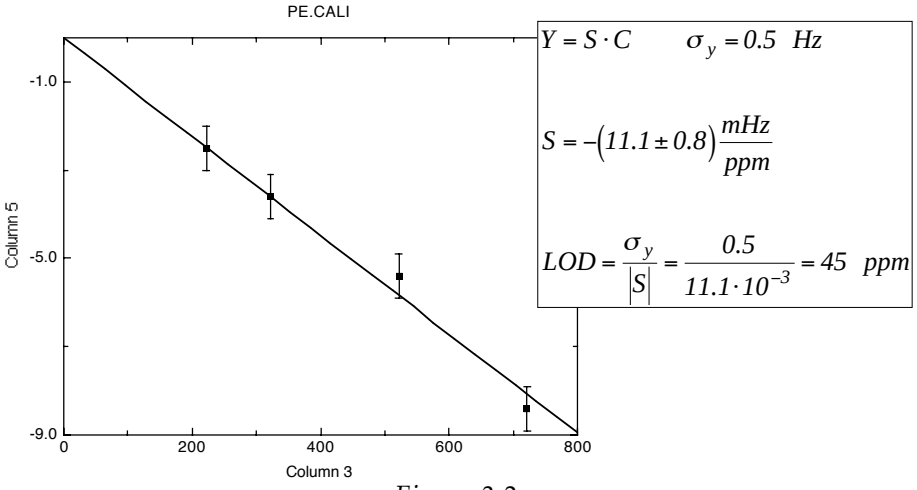


Figura 3.2

Studio della relazione tra glucosio e fruttosio in una popolazione di pesche

Natura dei dati: 21 pesche e nettarine. Le concentrazione di glucosio e fruttosio sono misurate con opportuni biosensori.
Le concentrazioni dei due zuccheri sono chiaramente "correlate". Le due quantità "co-variano": se l'una cresce anche l'altra aumenta e viceversa. Questi concetti saranno definiti più avanti. Ora ci chiediamo se esiste una legge lineare che rappresenta questa relazione.
Controlliamo prima le ipotesi dei minimi quadrati.

L'errore su y è molto maggiore dell'errore su x	NO	Entrambe le quantità discendono da misure con sensori simili
Y è distribuita normalmente	≈ SI	Essendo misure sperimentali obbediscono probabilmente al teorema di Gauss
Gli eventi osservati sono indipendenti	SI	Le misure sono indipendenti (attenzione all'effetto memoria dei sensori)
Le misure hanno varianza uguale	SI	Le misure sono svolte in condizioni simili e sempre con gli stessi sensori

La prima condizione non è verificata. Applichiamo ugualmente il metodo dei minimi quadrati, discuteremo in seguito questa ipotesi.
La figura 3.3 mostra l'esito del calcolo nel quale verifichiamo la esistenza ovviamente della relazione lineare. E' interessante notare il valore dell'intercetta (B) il quale valore è solo leggermente superiore all'errore di determinazione. Spetta chiaramente all'indagine fisiologica stabilire se B è 0 oppure se è possibile avere uno zucchero indipendentemente dall'altro.

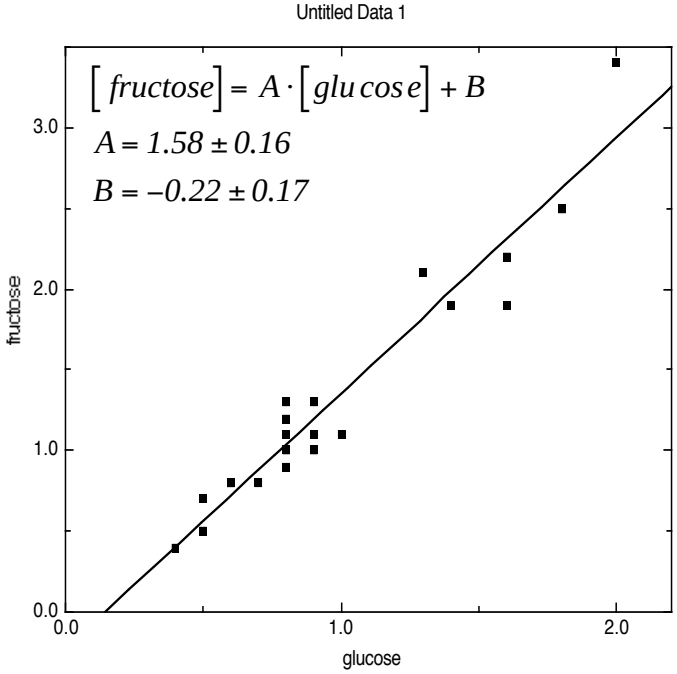


Figura 3.3

Discutiamo ora la prima ipotesi dei minimi quadrati.
La legge lineare è stata determinata ma le misure deviano dalla legge stessa. Tale deviazione va imputata agli errori di misura che esistono sempre tanto più in questo caso in cui si sono usati sensori e non strumenti di alta precisione.
La motivazione più importante però è che la legge lineare è di per se approssimata in quanto sia perché la relazione glucosio-fruttosio può non essere lineare e sia perché la quantità di fruttosio non dipende solo dal glucosio ma da tanti altri parametri chimici, fisici e biologici per cui la relazione esatta contiene molte più variabili
Tutte queste "inesattezze" del modello se sono distribuite casualmente possono essere trattate come errori di misura ed in questo contesto la prima ipotesi dei minimi quadrati può considerarsi soddisfatta. Infatti, la relazione tra i due zuccheri si può scrivere come:

$$[Fr] + \Delta[Fr] = K \cdot \left([Gl] + \Delta[Gl]\right) + \Delta Mod$$

Δ[Fr] e Δ[Gl] sono gli errori di misura dei rispettivi sensori, ΔMod è l'errore del modello lineare. Quindi l'errore globale di [Fr] è (Δ[Fr]+ΔMod) e poiché gli errori di misura sono verosimilmente simili la prima ipotesi dei minimi quadrati si può considerare soddisfatta.

3.4 Generalizzazione del Metodo dei Minimi Quadrati

Come noto dall'algebra lineare un sistema di equazioni lineari può essere espresso e risolto in termini di matrici e vettori. Si considerino due variabili x y poste in relazione tra loro da una legge lineare del tipo y=ax +b e si consideri di aver eseguito una serie di n misurazioni x_i, y_i con lo scopo di determinare da queste misure sperimentali il valore dei parametri a e b della legge lineare. A questo scopo si può costruire la seguente relazione matriciale:

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{pmatrix} a & b \end{pmatrix} \begin{bmatrix} x_1 & 1 \\ \dots & \dots \\ x_m & 1 \end{bmatrix}$$

che in forma compatta si può scrivere, usando la convenzione di adottare caratteri minuscoli in grassetto per indicare vettori e maiuscolo in grassetto per indicare le matrici, come:

$$\boldsymbol{y}^{\boldsymbol{T}} = \boldsymbol{k} \cdot \boldsymbol{X}$$

Se le misurazioni di entrambe le variabili x e y non sono affette da errori allora è sufficiente considerare solo due delle n misurazioni, così il vettore **y** prende dimensione 2 e la matrice **X** è di dimensioni 2x2. Il sistema diventa quindi determinato ed ha le seguente soluzione:

$$\boldsymbol{k} = \boldsymbol{y}^{\boldsymbol{T}} \cdot \boldsymbol{X}^{\boldsymbol{-1}}$$

dove la matrice $\mathbf{X^{-1}}$ rappresenta la matrice inversa di **X**, definita come quella matrice che, data **X**, soddisfa la seguente relazione:

$\mathbf{X^{-1} \bullet X = I}$ cioè la matrice identità. La condizione per l'esistenza della matrice $\mathbf{X^{-1}}$ è che il determinante della matrice **X** sia diverso da 0, cioè che tutte le righe della matrice **X** sia linearmente indipendenti una dalle altre.
Come però discusso nel paragrafo precedente il caso interessante si verifica quando le misure delle variabili x e y risultano affette da errori sperimentali. Si faccia inoltre la stessa assunzione, che si è già mostrato essere ragionevole, che gli errori sulla variabile x siano in percentuale trascurabili rispetto a quelli sulla variabile y.
In presenza di errori la relazione, supposta lineare, tra y e x si modifica nel modo seguente:

$$y = ax + b + e$$

dove la grandezza e rappresenta l'errore. Le considerazioni svolte di seguito hanno come ipotesi di partenza che la variabile e sia distribuita normalmente ed abbia media nulla, si è gia visto che praticamente il teorema del limite centrale assicura che queste ipostesi diventano sempre più vere tanto più alto è il numero di misure effettuate. La relazione precedente quando si siano eseguite n misure si scrive, in forma matriciale nel modo seguente:

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{pmatrix} a & b \end{pmatrix} \begin{bmatrix} x_1 & 1 \\ \dots & \dots \\ x_m & 1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \dots \\ e_n \end{bmatrix}$$

$$\boldsymbol{y^T = k \cdot X + e}$$

L'obiettivo del metodo dei minimi quadrati consiste quindi nel determinare il valore dei parametri che rende minima il modulo quadro (norma due) del vettore **e**, cioè che minimizza l'influenza degli errori di misura, giova sottolineare che nell'espressione precedente il vettore **y** rappresenta il vettore dei valori aspettati della variabile y, che è una grandezza ignota.
La presenza dell'errore fa si che tutte le equazioni del sistema sono indipendenti e quindi la matrice **X** non può più essere considerata quadrata ma ha dimensioni mx2.

La soluzione minimi quadrati del problema lineare è espressa da un teorema, detto di Gauss-Markov, che nella ipotesi di distribuzione normale e della media nulla del vettore **e** stabilisce che la soluzione è data dalla seguente relazione:

$$\boldsymbol{k} = \boldsymbol{y}^{\boldsymbol{T}} \cdot \boldsymbol{X}^{+}$$

La matrice $\mathbf{X^+}$ è detta matrice pseudoinversa di **X**. Questa rappresenta una generalizzazione del procedimento di inversione al caso di matrici non più quadrate ma rettangolari. La pseudoinversa, talvolta denominata "inversa generalizzata", è definita formalmente attraverso le relazioni di Moore-Penrose:

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

Per praticità di calcolo la pseudoinversa può essere calcolata come:

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

Se il determinante di $\mathbf{X^TX}$ è sufficientemente diverso da 0.

Regressione Polinomiale
La rappresentazione matriciale ora esposta consente di stimare i parametri anche per relazioni polinomiali di ordine qualsiasi. Si considerino n misure della coppia di variabili x ed y e che tra x ed y esista le seguente relazione polinomiale di ordine m:

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

Le n coppie di misure consentono quindi di scrivere la seguente relazione matriciale:

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & \dots & a_m \end{bmatrix} \cdot \begin{bmatrix} 1 & x_1 & \dots & x_1^m \\ & \dots & \dots & \\ 1 & x_n & \dots & x_n^m \end{bmatrix} + \begin{bmatrix} e_1 \\ \dots \\ e_n \end{bmatrix}$$

$$\boldsymbol{y^T = k \cdot X + e}$$

Una volta impostate correttamente le matrici il problema si risolve come nel cao lineare applicando la relazione del teorema di Gauss-Markov: $\boldsymbol{k = y^T \cdot X^+}$.

Regressione non-lineare
Il metodo dei minimi quadrati può essere esteso anche a problemi in cui la relazione che lega le variabili y e d x sia una funzione non lineare qualsiasi. Questo problema può essere affrontato come il caso lineare partendo da un sistema di equazioni nonlineari in cui il numero delle equazioni sia uguale al numero delle incognite, la soluzione in questo caso corrisponde al caso in cui si considerino le misure delle variabili x ed y non affette da errori. La soluzioni minimi quadrati del caso realistico in cui sono presenti errori sperimentali si ottiene generalizzando la soluzione del sistema quadrato.
La caratteristica principale dei sistemi di equazioni nonlineari è che la loro soluzione non può essere determinata univocamente ma è frutto di un algoritmo iterativo che presuppone un punto di partenza. Per comprendere il meccanismo di questi algoritmi è opportuno studiare la soluzione ad una equazione non-lineare del tipo f(x)=0. Questa relazione può essere risolta tramite un algoritmo iterativo detto algoritmo di Newton o metodo della tangente. Il metodo presuppone una soluzione di partenza arbitraria e poi procede iterativamente alla soluzione. la formula iterativa è la seguente:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

dove con f'(x) si indica, come consueto, la derivata di f(x). Questa formula ha una semplice spiegazione geometrica come illustrato nella figura 3.4.

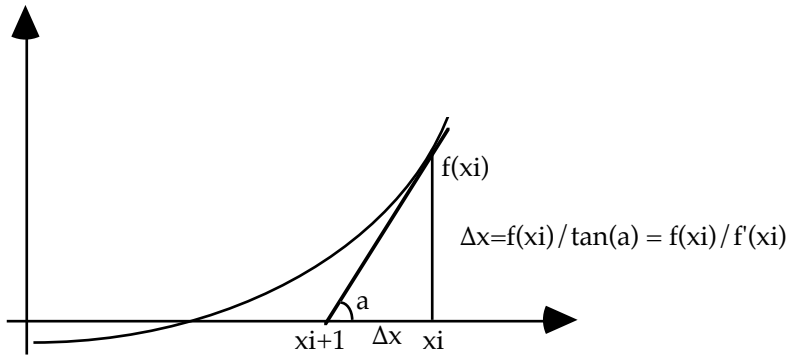


Figura 3.4

La grande sconvenienza di tale algoritmo, che si ripercuote anche nelle sue generalizzazioni ai sistemi di equazioni e ai minimi quadrati, è che la soluzione viene determinata se il punto di prova iniziale e la soluzione giacciono in un intervallo in cui la derivata della funzione non si annulla mai, cioè dove la funzione è monotona.
L'algoritmo di Newton si estende al caso dei sistemi di equazioni non lineari e prende il nome di algoritmo di Newton-Rhapson. La generalizzazione consiste nel considerare il seguente sistema di n equazioni in n incognite:

$$\begin{cases} y_1 = f(x_1, k_1, \ldots, k_n) \\ \quad\quad \ldots \\ y_n = f(x_n, k_1, \ldots, k_n) \end{cases}$$

per esprimere l'algoritmo iterativo di Newton-Rhapson è conveniente considerare il sistema scritto in formato vettoriale $\mathbf{y}=\{f_1,\ldots,f_n\}$ e considerando tutti i parametri incogniti elementi di un vettore $\mathbf{k}$, per cui la soluzione al passo i+1 e:

$$\boldsymbol{k}_{i+1} = \boldsymbol{k}_i - \boldsymbol{J}^{-1}(\boldsymbol{k}_i)\cdot\left(\boldsymbol{f}(\boldsymbol{k}_i) - y\right)$$

dove

$$\boldsymbol{J}_{ij} = \frac{\partial f_i}{\partial k_j}$$

la matrice **J** è detta matrice Jacobiana di **f**.

3.5 La validazione del modello

La regressione statistica è basata sull'esistenza degli errori di misura, cioè sul fatto che il risultato di una misura è descritto da una parte deterministica (la legge fisica) e da una parte casuale (statistica) che abbiamo chiamato errore. Il metodo dei minimi quadrati consente di determinare, una volta fissata la natura della funzione deterministica, i parametri della funzione che minimizzano la parte statistica.
E' chiaro come la minimizzazione dell'errore possa essere compiuta efficacemente aumentando il grado di complessità della funzione. In particolare, è ovvio come dati n dati sperimentali, questi sono perfettamente descritti da un polinomio di grado n.
La scelta della funzione però non può essere arbitraria, in quanto spetta alla descrizione fisica del fenomeno decidere il tipo di relazione che lega le grandezze oggetto del problema.
Come abbiamo accennato precedentemente, la scelta della funzione di regressione non può essere orientata alla minimizzazione dell'errore in quanto l'errore diminuisce aumentando il numero di parametri da stimare, e quindi il grado di complessità della funzione.
Il criterio con il quale si possono confrontare differenti funzioni è quello della stima (predizione) di dati non utilizzati durante la fase di calibrazione. In pratica, il data set di calibrazione viene diviso in due porzioni, una per la calibrazione, cioè la stima dei parametri, e l'altra per la validazione cioè per la misura della capacità di predizione del modello stimato.
Ad esempio consideriamo il set di dati rappresentato in figura 3.5 nel piano x-y

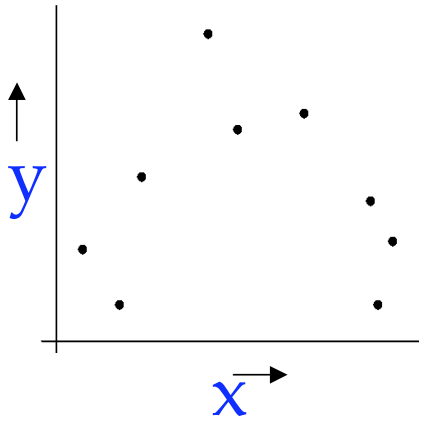


Figura 3.5

Supponiamo di voler trovare la relazione deterministica che lega le due sequenze di dati (x, y) supponendo valide le condizioni del metodo dei minimi quadrati. Quindi supponendo che la relazione tra y e x si a $y=f(x)+e$. Dove $f(x)$ è la parte deterministica ed e l'errore.
Nella figura 3.6 consideriamo tre tipi di funzioni possibili: una funzione lineare (la retta), una moderatamente non lineare (parabola), ed una altamente non lineare (un polinomio di grado n che passa perfettamente per i dati sperimentali).

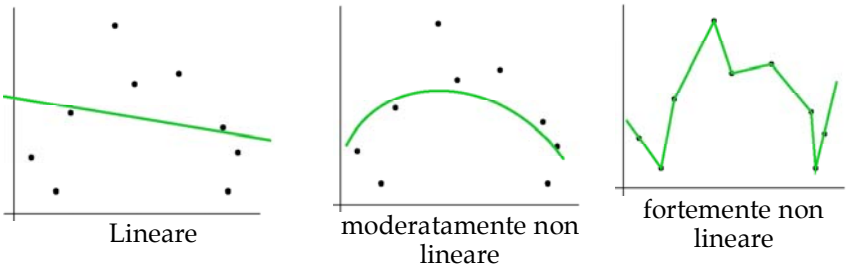


Figura 3.6

Ovviamente la terza scelta è quella che rende minimo l'errore di calibrazione, in pratica nella funzione che descrive i dati scompare il termine statistico e.

Nel metodo della validazione si tolgono dall'insieme di calibrazione alcuni dati che vengono utilizzati successivamente per *misurare* la

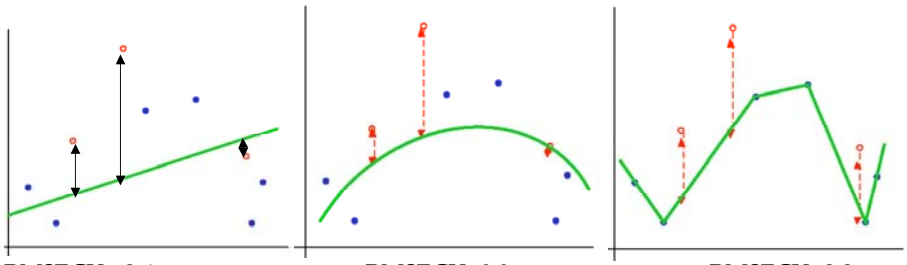
capacità del modello sviluppato di predirre il valore vero della variabile y conoscendo la variabile x.
È quindi necessario definire alcune quantità che rendono possibile la misura della capacità di predizione. L'errore compiuto in un insieme di dati si chiama PRESS (PREdicted Sum of Squares). Il PRESS è la somma dei quadrati delle differenze tra valore vero (y) e valore predetto (y^{LS}).

$$PRESS = \sum_i \left(y_i^{LS} - y_i \right)^2$$

Gli scarti vengono sommati quadraticamente perché gli errori non si compensano.
Le quantità necessarie per valutare il potere di regressione di un modello sono l'errore sul seti di calibrazione e quello sul set di validazione, i due errori si chiamano rispettivamente RMSEC (Root Mean Square of Calibration) e RMSECV (Root Mean Square of Cross-Validation).
Tali quantità sono relative al PRESS calcolato sui rispettivi data set nei modi seguenti:

$$RMSEC = \sqrt{\frac{PRESS}{N}} \;\; ; \;\; RMSECV_k = \sqrt{\frac{PRESS_k}{k}}$$

Nel secondo termine k indica il numero di dati sui quali si calcola la validazione.
Nella figura 3.7, vediamo la valutazione dell'RMSECV considerando 2 dati come set di validazione:



RMSECV= 2.4 RMSECV=0.9 RMSECV=2.2

Figura 3.7

L'esempio mostra come la funzione moderatamente non lineare sia quella che minimizzando l'errore di predizione ottiene una migliore

descrizione dei dati. Si osservi come la funzione lineare e quella altamente non lineare diano luogo alle stesse prestazioni nonostante la più elevata complessità di calcolo necessaria nel caso fortemente non lineare.
Il caso del modello fortemente non lineare, cioè il caso di errore di calibrazione nullo ed errore di validazione elevato si chiama overfitting. Il metodo di regressione cioè diventa sopra-specializzato nel descrivere i casi di calibrazione ma diventa non adeguato alla generalizzazione, cioè a descrivere i casi non utilizzati. Tale problema è tipico in tutti i metodi non lineare e non parametrici come ad esempio le reti neurali.
Il metodo della validazione, detto generalmente cross-validation, è semplice ma di non facile attuazione. In pratica, quanti e quali campioni considerare come calibrazione e test non è immediato. Una regola generale dovrebbe essere quella della uguaglianza statistica tra i due insiemi di dati (in termini di media e varianza). E' d'altronde ovvio che il risultato (RMSECV) può cambiare a seconda di quali campioni vengono considerati per la validazione.
Il metodo più generale per stimare l'errore di predizione di un modello è quello del leave-one-out, letteralmente uno-lasciato-fuori. Il metodo consiste nel ridurre a 1 il contenuto dell'insieme di test, e nel valutare l'errore di predizione rispetto al dato eliminato. Eliminando consecutivamente tutti i dati e facendo la media degli errori ottenuti si ottiene una stima robusta dell'errore di predizione del modello funzionale.
Ad esempio consideriamo nella figura 3.8 l'applicazione del metodo al data set precedente e nel caso della funzione lineare.

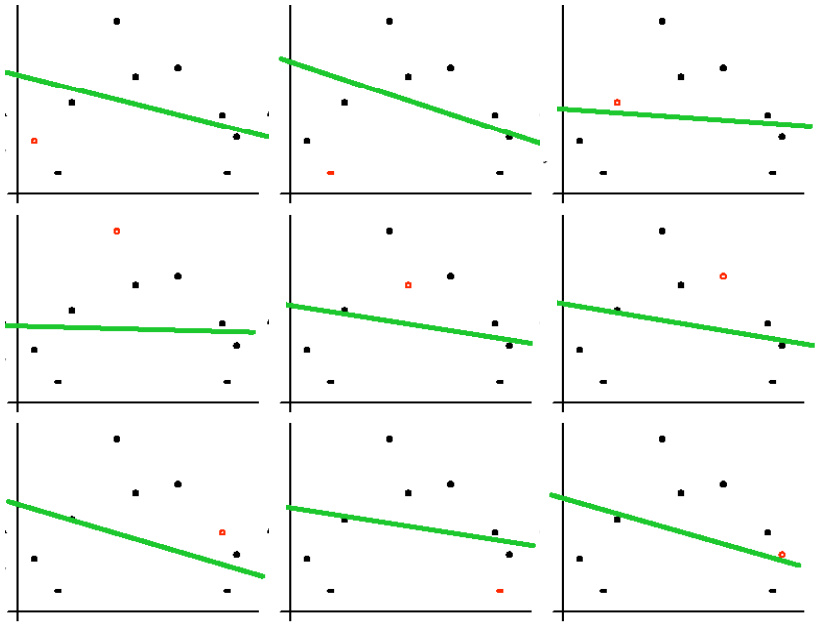


Figura 3.8

L'errore complessivo sarà la media degli RMSECV ottenuti in ogni passo, si noti che in ogni passo la retta di regressione è ovviamente diversa. Applicando questo metodo ai dati precedenti si ottiene:
$RMSECV_{lineare}$=2.12; $RMSECV_{parabola}$=0.96; $RMSECV_{non\text{-}ineare}$=3.33.
L'errore di generalizzazione del metodo non lineare è adesso decisamente più elevato rispetto agli altri due, e ancora la curva di secondo ordine è quella che dà il risultato migliore.
Il Leave-one-out è il test migliore disponibile per avere una stima della predizione su nuovi dati di un modello di regressione. Nel caso di insiemi molto numerosi, il metodo risulta dispendioso in termini di tempo di calcolo; si può ovviare a questo ammorbidendo il leave-one-out considerando blocchi di k dati invece che di 1.

3. 6 Il test del χ^2

Nel paragrafo precedente abbiamo visto come sia possibile stimare per un modello regressione l'errore aspettato su un set di dati non utilizzati nella regressione. D'altro canto, il metodo dei minimi quadrati richiede di conoscere a priori il tipo di relazione funzionale

tra le variabili. Tale relazione è in genere suggerita da un analisi teorica dei fenomeni rappresentati dalle variabili stesse, oppure in assenza di questa si procede utilizzando funzioni semplici.
In generale però ci si può chiedere se esista un metodo per stimare la probabilità che un insieme di dati possa essere rappresentato da un certo tipo di relazione funzionale.
A tal proposito esiste il test del χ^2. Questi è un procedimento statistico che consente di assegnare una certa probabilità alla ipotesi relativa alla forma funzionale tra due variabili statistiche.
La variabile χ^2 è la somma dello scarto quadratico tra la variabile misurato ed il suo valore teorico, ogni scarto è pesato per la deviazione standard della misura. Date N variabili casuali indipendenti (x_k) con valori medi (m_k) e varianza (σ^2) il χ^2 è definito come:

$$\chi^2 = \sum_{k=1}^{N} \frac{\left(x_k - m_k\right)^2}{\sigma_k}$$

Si osservi che tale grandezza è simile a quella minimizzata dal metodo dei minimi quadrati.
Si può dimostrare che la variabile χ^2 ha una distribuita secondo una PDF particolare, detta distribuzione χ^2, la cui forma funzionale è:

$$f\left(\chi^2\right) = \frac{1}{2^{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu}{2}\right)} e^{-\frac{\chi^2}{2}} \left(\chi^2\right)^{\frac{\nu}{2}-1} \qquad con \;\; \Gamma\left(\frac{\nu}{2}\right) = \left(\frac{\nu}{2} - 1\right)!$$

la distribuzione χ^2 dipende da unico parametro ν detto gradi di libertà.
La figura 3.9 mostra l'andamento della distribuzione del χ^2 per alcuni valori di ν.

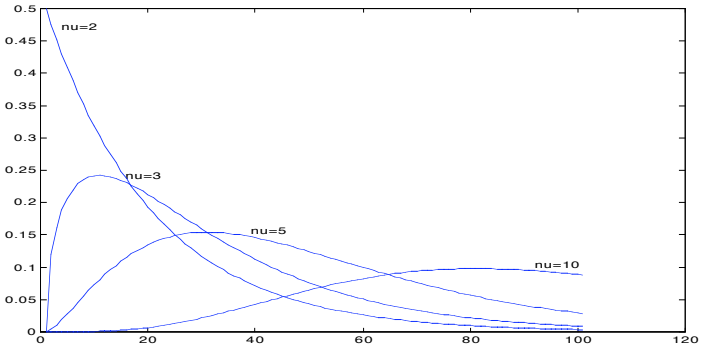


figura 3.9

Ad ogni valore di ν corrisponde□□□□ diversa distribuzione, Per ν=1 la distribuzione diverge per χ^2 =0, per ν⇒∞ la distribuzione converge alla distribuzione normale.
Per ogni valore di ν in corrispondenza di χ^2=ν/2 si ottiene il valore massimo della distribuzione.
Il test del χ^2 consiste nel calcolare il χ^2 per una data distribuzione di dati e per una determinata ipotesi di relazione funzionale e di valutare la probabilità di ottenere un valore di χ^2 superiore al valore calcolato.
In pratica, supponiamo di avere eseguito la regressione con il metodo dei minimi quadrati di un insieme di dati ogni volta usando differenti forme funzionali.
Per ogni forma funzionale si calcola il corrispondente valore del χ^2 con la seguente relazione:

$$\chi^2 = \sum_{k=1}^{N} \frac{\left(y_k - f\left(x_k ; a_1 , a_2 , \ldots , a_n\right)\right)^2}{\sigma_k^2}$$

Il numero di gradi di libertà, fondamentale per poter applicare la distribuzione del χ^2 è pari a $\nu = N - n_{param}$, dove N_{param} è il numero dei paramateri della funzione f stimati nella regressione. Tale definizione giustifica il nome di gradi di libertà per il parametro ν che indica il numero di dati indipendenti. Tale definizione è già

stata introdotta nel capitolo 2 per la stima della varianza e dello skewness.
Una volta calcolati i valori di χ^2 e ν si può calcolare il valore della probabilità di avere un χ^2 maggiore del χ^2 calcolato

$$\alpha = \int_{\chi_0^2}^{\infty} f\left(\chi^2\right) \cdot d\chi^2$$

ovviamente tanto più alta è tale probabilità tanto più attendibile è l'ipotesi.
E' interessante notare che il valore di tale probabilità aumenta al crescere di ν ed al diminuire di χ^2. Incrementare ν significa semplicemente, a parità di ipotesi (e quindi del numero di parametri), aumentare il numero dei dati sperimentali. Il valore di χ^2 invece diminuisce ovviamente col calare dello scarto tra valori misurati ed valori ipotetici e con l'aumentare dell'errore di misura. Questa ultima proprietà è interessante in quanto stabilisce che all'aumentare dell'errore di misura è plausibile che un gran numero di forme funzionali si accordino con i dati sperimentali.
Come esempio del test del χ^2 consideriamo in figura 3.10 il caso di 8 coppie (X,Y) di dati affette da un errore di misura del 10% sulla variabile Y, e confrontiamo due relazioni una lineare ed una logaritmica.

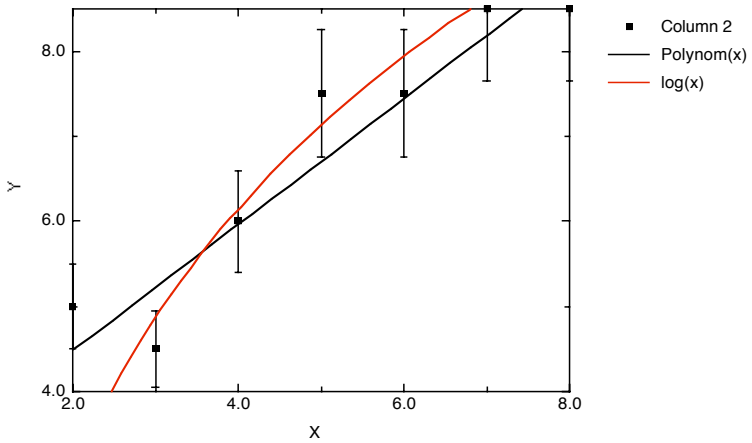


figura 3.10

Per le due ipotesi funzionali otteniamo i seguenti valori della regressione, del χ^2, dei gradi di libertà e della probabilità α.

Ipotesi	y=s+k x	y=k log x
Parametri stimati	s=2.99; k=0.74	k=4.43
χ^2	5.17	16.91
ν	8-2=6	8-1=7
α	50%	2.5%

Il test mostra come la probabilità dell'ipotesi lineare sia molto maggiore della logaritmica.
Il test del χ^2 è molto sensibile rispetto ai dati utilizzati, supponiamo ad esempio di dubitare, per qualche motivo, della attendibilità della misura relativa a x=2, oppure supponiamo di non avere eseguito affatto tale misura!
Eseguendo la regressione omettendo tale misura e ripetendo il test del χ^2 si ottengono i seguenti valori:

Ipotesi	y=s+k x	y=k log x
Parametri stimati	s=2.20; k=0.87	k=4.27
χ^2	3.01	4.27
ν	8-2=6	8-1=7
α	80%	99%

Si osservi come il risultato della regressione sia praticamente lo stesso del caso precedente, mentre la probabilità del test sia diventata dell'80% per l'ipotesi lineari e addirittura del 99% per l'ipotesi logaritmica che adesso diventa quella statisticamente più attendibile.
Questo esempio mostra l'importanza dell'accuratezza sperimentale per poter eseguire delle deduzioni corrette dagli esperimenti. D'altro canto però bisogna sempre tenere presente che il risultato statistico, che è sempre espresso in termini di probabilità, non si impone mai sulla teoria fisica. Proprio per questo motivo, tra l'altro, attribuiamo tutte le deviazioni dalla teoria agli errori di misura.

3. 7 La correlazione
Uno dei problemi generali nella interpretazione dei risultati sperimentali è quello di intepretare le relazioni esistenti tra grandezze. In particolare, è di estrema importanza poter dedurre da

un insieme di dati, se due grandezze sono legate da una relazione funzionale. Inoltre, è estremamente importante poter decidere se la relazione indica una legge fondamentale oppure se è legata ad aspetti particolari degli esempi misurati.
La quantità più semplice che serve ad indicare la esistenza di una relazione funzionale tra due grandezze misurate è la correlazione.
Date due grandezze Y e X queste sono dette correlate se dall'andamento di una è possibile dedurre l'andamento dell'altra. In pratica, se una grandezza aumenta e l'altra aumenta di conseguenza e viceversa. Si dice anche che le due grandezze "co-variano" cioè variano assieme.
La correlazione è espressa attraverso una quantità numerica detta coefficiente di correlazione lineare che ha valori compresi nell'intervallo –1 1. I valori negativi si riferiscono alla cosiddetta anticorrelazione quando all'aumentare di una grandezza l'altra "tende" a diminuire.
La correlazione perfetta è indicata dai valori del coefficiente di correlazione pari a ±1 mentre un valore vicino a 0 del coefficiente indica l'assenza della correlazione. Il coefficiente di correlazione lineare definisce in pratica se è possibile determinare un modello di regressione lineare tra due grandezze.
La correlazione parziale indica, in generale, che oltre a dipendere dalla variabile X, la variabile Y dipende da altre grandezze non determinate.
Nella figura 3.11 si mostrano alcuni esempi di correlazione relativi ad alcuni parametri misurati in una popolazione di vini.

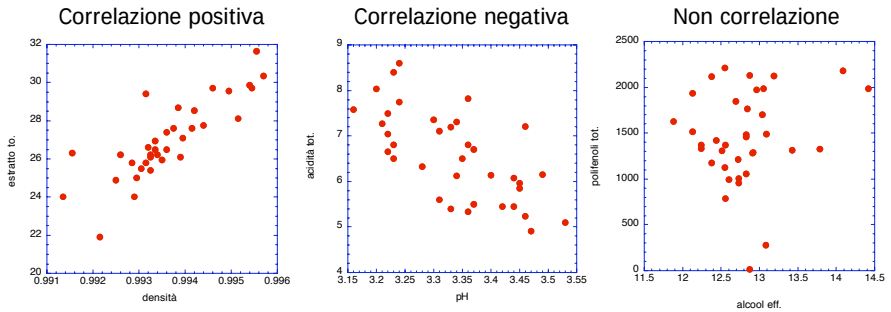


Figura 3.11

Nel primo caso a sinistra si vede la quantità di estratti graficata in funzione della densità, queste due grandezze sono sicuramente

proporzionali tra loro in quanto l'estratto totale è proprio la massa di sostanze che aggiunte alla acqua determinano la densità della soluzione. La figura centrale mostra la ovvia anticorrelazione tra acidità totale e pH; il pH infatti è inversamente proporzionale alla concentrazione degli ioni H^+ che determinano il grado di acidità di una soluzione. Infine vediamo l'esempio di due grandezze assolutamente non correlate come i polifenoli e l'alcool; i polifenoli infatti derivano dalle proprietà delle bucce degli acini mentre l'alcool dai processi fermentativi che trasformano il contenuto in zuccheri delle uve.
Il coefficiente di correlazione lineare può essere calcolato considerando il coefficiente angolare della retta di regressione tra le grandezze Y e X. Supponiamo di avere le coppie di dati sperimentali (X,Y) mostrate in figura 3.12:

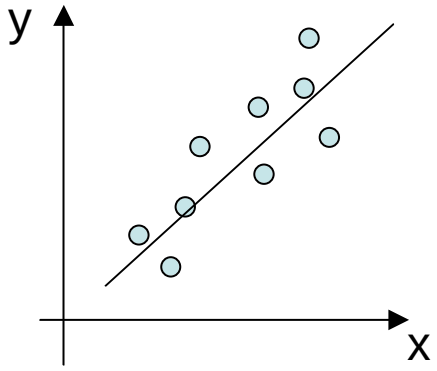


Figura 3.12

e di calcolare le rette di regressione sia diretta che inversa.

$$y = m \cdot x + b$$

$$x = m' \cdot y + b' \Rightarrow y = \frac{x}{m'} - \frac{b'}{m'}$$

Nel caso in cui si abbia correlazione si ha m=1/m′ quindi mm′=1. Nel caso in cui la correlazione sia assente è facile verificare che m′=0 e quindi mm′=0. Il prodotto mm′ può quindi essere assunto come coefficiente di correlazione.

Data una sequenza di misure (x_i,y_i), il coefficiente di correlazione può essere calcolato direttamente dalla espressione del metodo dei minimi quadrati del coefficiente angolare della retta di regressione:

$$\rho = m \cdot m' = \frac{N\sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right] \cdot \left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]}$$

Il coefficiente di correlazione è dotato di segno in modo da distinguere tra correlazione e anticorrelazione. Nonostante questo, due coefficienti di correlazione c=C e c=-C esprimono la stessa linearità a meno del segno del termine di proporzionalità.
Per questo motivo il valore assoluto del coefficiente di correlazione si usa per descrivere la bontà di un regressione lineare. Tale termine è indicato come R (regression coefficient). La figura 3.13 mostra il coefficiente R nel caso della relazione tra estratto totale e densità nel vino, discusso in precedenza.

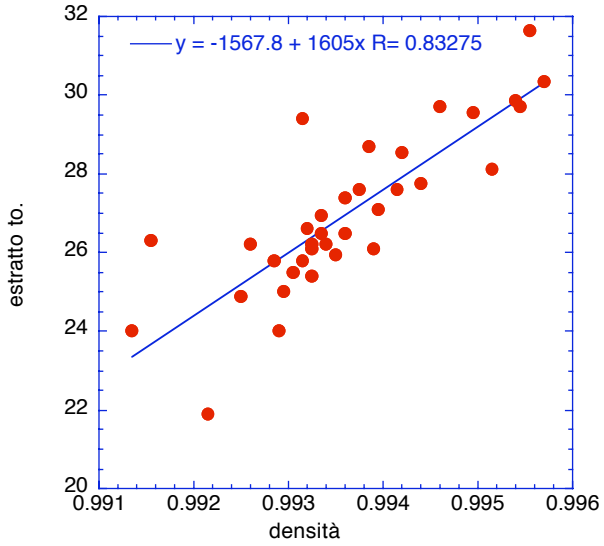


figura 3.13

4

MATRICI DI DATI E SENSOR ARRAYS

Supponiamo di voler misurare contemporaneamente due grandezze x_1, ed x_2. Ovviamente abbiamo bisogno di almeno due sistemi indipendenti di misura. Ci sono due possibilità estreme che riguardano la selettività dei metodi di misura utilizzati. In un caso possiamo avere a che fare con metodi di misura assolutamente selettivi e nell'altro con metodi non selettivi.
Due metodi assolutamente selettivi, nei confronti di x_1, ed x_2 si esprimono con le seguenti relazioni

$$y_1 = k_1 \cdot x_1$$
$$y_2 = k_2 \cdot x_2$$

In questo modo, le due grandezze vengono misurate indipendentemente l'una dall'altra.

Introduciamo a questo punto due importanti spazi di rappresentazione, lo spazio delle variabili e lo spazio degli osservabili (cioè delle grandezze misurate) nel caso appena descritto, lo spazio degli osservabili riproduce esattamente lo spazio delle variabili (a parte ovviamente le differenze di scala e di unità di misura).

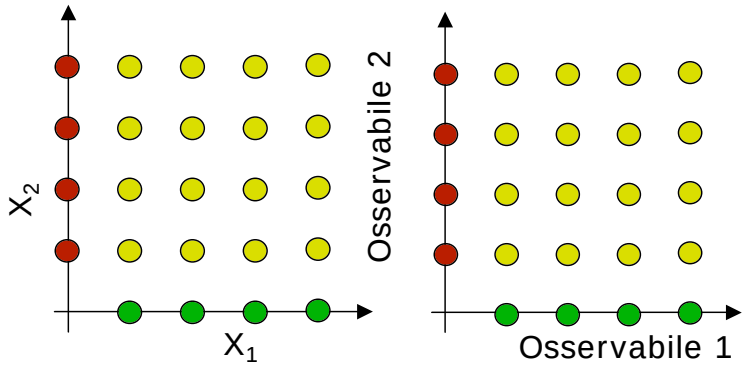


Figura 4.1

L'altro caso, più interessante per i nostri scopi, è quello che si osserva quando gli osservabili non sono indipendenti, cioè quando i metodi di misura sono non selettivi, cioè quando una misura contiene, in proporzione variabile, informazioni su entrambe le variabili.
Rimanendo nel caso lineare, possiamo scrivere le seguenti relazioni

$$y_1 = A \cdot x_1 + B \cdot x_2$$
$$y_2 = C \cdot x_1 + D \cdot x_2$$

Nello spazio degli osservabili questo corrisponde alla situazione mostrata in figura 4.2:

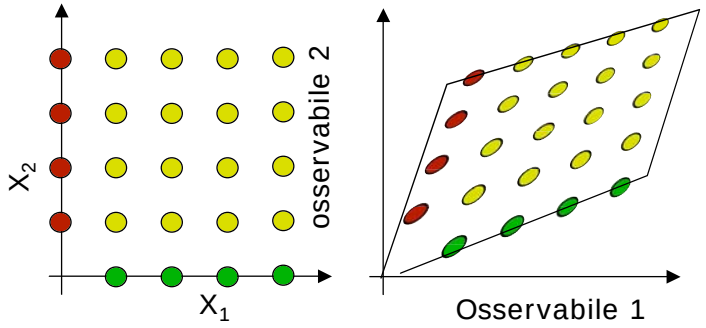


Figura 4.2

La differente situazione si esprime in maniera sintetica, scrivendo le relazioni tra misure e variabili in termini matriciali e considerando la correlazione espressa dalla matrice stessa.
Dato infatti un sistema di equazioni lineari $Y=\boldsymbol{K}X$ la correlazione tra gli elementi del vettore Y è data da $1\text{-}det(\boldsymbol{K})$. Nel caso di osservabili selettivi la matrice **K** è diagonale, il determinante è massimo e la correlazione è minima (0 nel caso omo-dimensionale), il caso di osservabili non selettivi, il determinate risulta dipendente dai valori di A, B, C, e D e la correlazione è compresa tra 0 e 1. Nell'altro caso limite in cui A=γB, e C=γD. Il determinante è nullo e la correlazione massima (1). Questo ultimo caso corrisponde a due misure che forniscono risultati solo proporzionali tra loro, ovviamente le due misure corrispondono ad una sola.
Geometricamente, la correlazione corrisponde alla ampiezza del parallelogramma descritto nel piano degli osservabili. Il parallelogramma diventa un quadrato nel caso di C=0 e si riduce ad una retta nel caso C=1.
Ovviamente, quanto descritto nel piano è generalizzabile a qualunque dimensione sia nel numero delle variabili che in quello delle misure.
In termini generali bisogna ricordare che le relazioni vere che legano osservabili e variabili sono le seguenti

$$y_1 = k_1 \cdot x_1 + k_2 \cdot x_2 + e_1$$
$$y_2 = w_1 \cdot x_1 + w_2 \cdot x_2 + e_2$$

Si deve infatti considerare che, in quanto misure sperimentali, sono sempre affette da un termine statistico aggiuntivo che chiamiamo errore.
La presenza di e fa si che ricavare le variabili x dalle grandezze y sia un problema di tipo statistico che risolvibile con il metodo dei minimi quadrati se le ipotesi del metodo sono confermate.
Misure non indipendenti, o meglio misuratori non selettivi, corrispondono al caso più comune e generale. Sono ad esempio misuratori non selettivi gli spettrofotometri (in un unico spettro sono contenute informazioni su molte specie chimiche) i gas-cromatografi (in quanto un gas-cromatogramma è tipicamente uno strumento per la misura di miscele complesse) o una matrice di sensori chimici (molti sensori chimici sono intrinsecamente non selettivi come ad

esempio le micro-bilance al quarzo o i sensori ad ossidi-metallici semiconduttori).

4.1 Multiple Linear Regression

Il metodo più semplice per estrarre le variabili x da un insieme di misure Y è quello che considera una relazione lineare tra i due blocchi di variabili. Tale metodo si chiama Multiple Linear Regression (MLR).
Consideriamo n osservabili (sensori, canali spettrali, tempi di eluizione di un gas-cromatogramma,…) ed m variabili. Le m variabili possono essere stimate utilizzando il metodo dei minimi quadrati.
Per applicare correttamente il metodo è necessario verificare che siano rispettate le quattro ipotesi fondamentali:

1. l'errore su *Y* molto maggiore dell'errore su *X*
2. le *Y* sono distribuite normalmente
3. gli eventi osservati sono indipendenti
4. tutte le misure hanno uguale varianza.

Il problema lineare si può così scrivere nel modo seguente: $\boldsymbol{X}=\boldsymbol{YB}+\boldsymbol{E}$.
La scrittura matriciale può risultare più chiara considerando la espressione grafica rappresentata in figura 4.3:

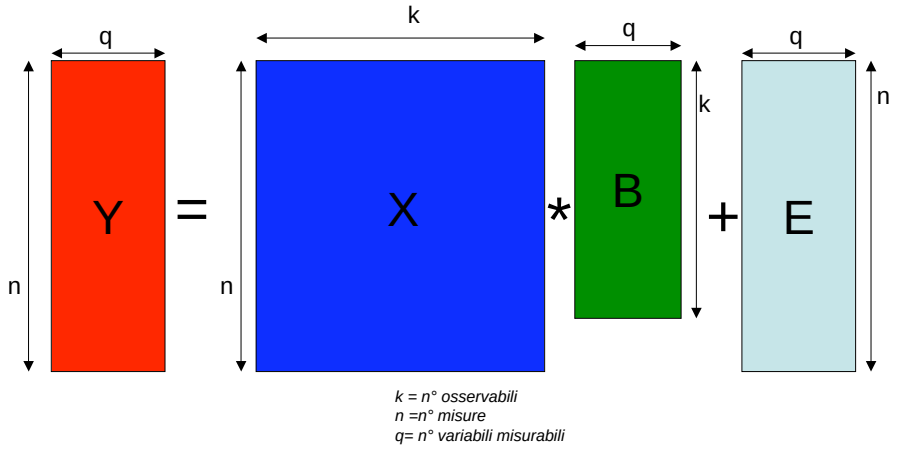


Figura 4.3

Formalmente questo problema è equivalente alla soluzione dei minimi quadrati per grandezze monovariate. La soluzione è fornita dal cosiddetto teorema di Gauss-Markov per cui la migliore stima per la matrice B è data da: $B_{MLR} = Y^{+} \cdot X$.

La soluzione del problema è agevole solo nel caso in cui la matrice Y sia di rango massimo cioè sia composta da grandezze poco correlate tra di loro. Questa condizione si verifica di rado, in pratica le correlazioni tra osservabili sono elevate.
Ad esempio per uno spettro ottico, la correlazione tra i canali che costituiscono le caratteristiche spettrali, righe e bande, sono elevate. Infatti, 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 eguale modo proporzionali alla caratteristica del campione, quindi le relative variabili (colonne nella matrice dei dati) risulteranno colineari. La figura 4.4 mostra un esempio relativo ad uno spettro nel vicino infrarosso (NIR).
E' importante notare che in genere sono colineari quelle variabili che dipendono in modo quantitativo con le caratteristiche del campione misurato.

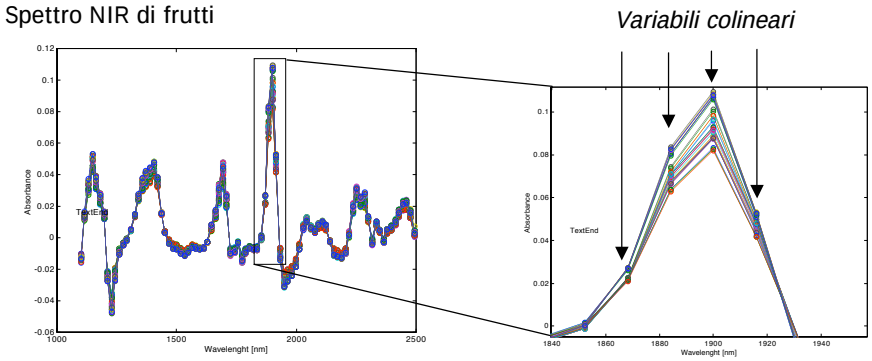


Figura 4.4

In questi casi la MLR non è applicabile, per superare questo scoglio è necessario eliminare la correlazione, cioè introdurre degli osservabili artificiali (ovviamente calcolati a partire da quelli reali) che risultino non correlati tra loro. Questo obiettivo sarà raggiunto nel paragrafo

precedente utilizzando del proprietà della cosiddetta analisi delle componenti principali.

4.2 Matrice di correlazione

Un insieme di dati multivariati costituisce una matrice con N righe (numero dei campioni misurati) ed M colonne (numero delle variabili multivariate). Un importante informazione sulle variabili che costituiscono l'insieme dei dati riguarda il loro grado di correlazione. Abbiamo visto nel paragrafo precedente come proprio l'esistenza di queste correlazioni complica i calcoli nel caso delle regressioni multivariate. I coefficienti di correlazione calcolati per tutte le coppie di variabili sono raggruppati in una matrice detta matrice di correlazione.

Supponiamo ad esempio di avere misurato per una popolazione di pesche i seguenti cinque parametri: Acidità totale, antociani, gradi brix, carotene e clorofilla. L'insieme delle correlazioni forma la seguente tabella 5x5

	ACID	ANTOC	BRIX	CAROT	CLOROF
ACID	1.00	-0.48	0.34	0.15	-0.32
ANTOC	-0.48	1.00	-0.70	-0.56	0.88
BRIX	0.34	-0.70	1.00	0.30	-0.76
CAROT	0.15	-0.56	0.30	1.00	-0.25
CLOROF	-0.32	0.88	-0.76	-0.25	1.00

La matrice è simmetrica, nel senso che Y rispetto a X ha la stessa correlazione di X rispetto a Y. Inoltre, i valori della diagonale valgono 1 in quanto ogni grandezza è perfettamente correlata con se stessa.

Possiamo osservare come non esistano praticamente mai grandezze assolutamente ne correlate (c=1) ne scorrelate (c=0) ma che la correlazione è quasi sempre parziale.

Quindi stabilire se due grandezze sono o no correlate dipende dal contesto e dall'applicazione.

La correlazione tra le variabili di un insieme di dati multivariati viene a volte visualizzata attraverso i grafici delle variabili una rispetto all'altra. In questo modo la dipendenza tra variabili è visivamente evidente. Poiché la matrice di correlazione è diagonale solo una parte dei grafici è riportata nella figura 4.5.

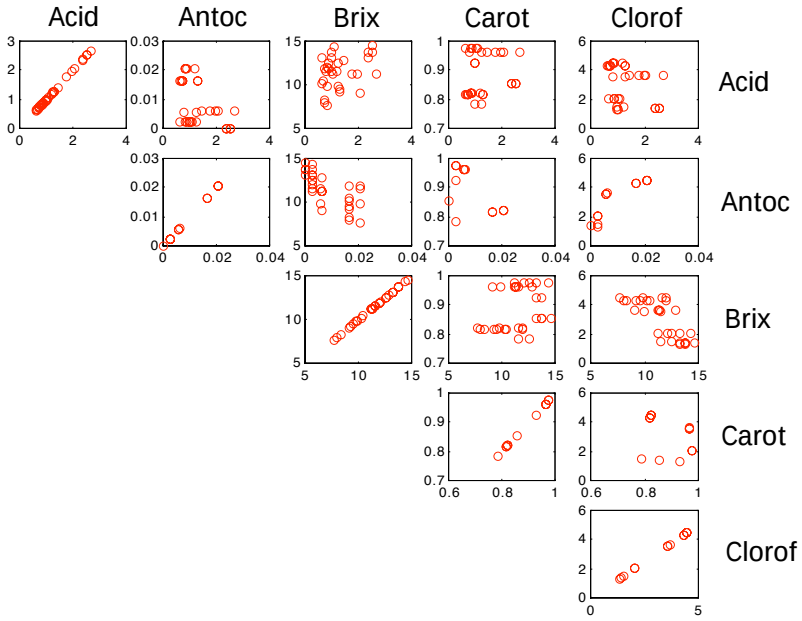


Figura 4.5

Confrontando i grafici precedenti con i valori della matrice di correlazione possiamo interpretare il significato dei coefficienti stessi. Osserviamo ad esempio che il grafico che mostra la relazione funzionale più evidente è quello tra antociani e clorofilla ed il coefficiente di correlazione per questa coppia di variabili è 0.88. La coppia con il vaore immediatamente minore è clorofilla-brix per la quale si ha un coefficiente pari a –0.76. La dipendenza anche in questo caso è evidente ma con una quantità rilevante di errore (nel senso dei minimi quadrati). Per tutte le altre coppie non c'è nessuna evidenza grafica di correlazione ed il coefficiente è minore di 0.56 (in valore assoluto) .

4.3 Analisi delle componenti principali

L'analisi delle componenti principali è un metodo per la scomposizione di dati multivariati in componenti scorrelate. Il metodo fornisce quindi la possibilità di risolvere la regressione lineare multipla eliminando le difficoltà di calcolo che provengono dalla correlazione degli osservabili.

Per definire le componenti principali è necessario considerare la statistica dello spazio degli osservabili. Dato un insieme di misure multivariate, rappresentate dalle righe della matrice *Y*, questi possono essere rappresentati come vettori in uno spazio vettoriale la cui dimensione è pari al numero delle colonne di *Y*. La analisi delle componenti principali si basa sulla ipotesi che l'insieme di dati multivariati sia distribuito normalmente. Proprio dalle proprietà della distribuzione normale multivariata discende il metodo di calcolo delle componenti principali.
Per definire una funzione di densità di probabilità (PDF) multivariata è necessario generalizzare i descrittori statistici definiti per le variabili monodimensionali al caso multivariato. Per la PDF gaussiana solo due descrittori sono utilizzati: la media e la varianza. La media multivariata è il vettore medio dell'insieme di dati, quindi da uno scalare, nel caso multivariato, si passa ad un vettore nel caso multivariato. Nel caso della varianza la generalizzazione è più complessa. La varianza infatti è sostituita dalla matrice di covarianza. La matrice di covarianza è definita come il valore aspettato della forma quadratica della differenza x-m, dove x è l'insieme dei dati ed m il loro valore medio. Formalmente questa definizione coincide con quella della varianza monovariata (valore aspettato del quadrato dello scarto, cioè della differenza tra i dati ed il loro valore medio).
La natura vettoriale di x ed m trasforma il quadrato nella forma quadratica.
La covarianza di un insieme di dati X si esprime come:

$$cov(X) = \Sigma = E\left[(x-m)^T \cdot (x-m)\right]$$

la matrice di covarianza è una matrice quadrata simmetrica. La matrice stabilisce il grado di covarianza delle singole variabili, gli elementi della matrice misurano come le singole variabili variano insieme (co-variano).
Gli elementi diagonali di Σ sono le varianze delle singole variabili, mentre gli elementi diagonali sono proporzionali al coefficiente di correlazione.

$$\Sigma_{ii} = \sigma_i^2 \quad ; \quad \Sigma_{ik} = \rho_{ik}\sigma_i\sigma_k$$

L'espressione della PDF multivariata è quindi:

$$PDF(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\Sigma}} exp\left[-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)\right]$$

La figura 4.6 mostra la PDF nel caso bivariato, l'unico graficamente rappresentabile)

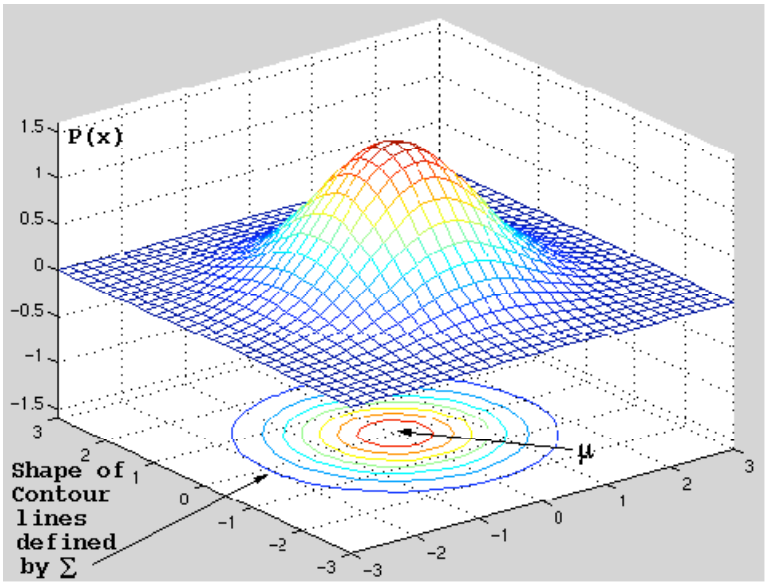
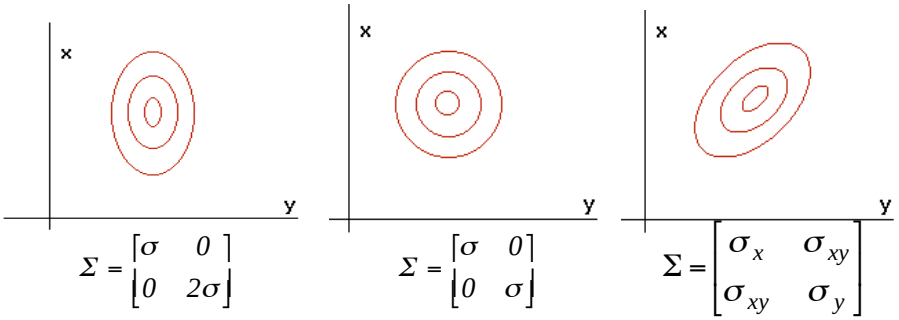


figura 4.6

E' importante osservare le ellissi nel piano della variabile X, queste curve costituiscono il luogo dei punti di iso-probabilità. Tali ellissi sono definite dalla forma quadratica costruita dalla matrice di covarianza.
La forma quadratica è costruita come:

$$\begin{bmatrix} x-\bar{x} & y-\bar{y} \end{bmatrix} \cdot \begin{bmatrix} a & b \\ b & c \end{bmatrix} \cdot \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix} = k$$

sviluppando la forma quadratica, cioè i vari prodotti righe per colonne, si ottiene proprio la forma di una quadrica. Poiché la matrice di covarianza è simmetrica per definizione, la quadrica che si ottiene dalla matrice di covarianza è sempre un'ellisse. Per i vali valori di k si ottengono i differenti livelli, cioè le ellissi corrispondenti alle varie probabilità.
Per inciso, si consideri che la PDF multivariata ha senso solo se le variabili di Y sono correlate tra loro, infatti la PDF di variabili non correlate ed indipendenti è data dal prodotto delle distribuzioni monovariate. Come noto infatti la probabilità congiunta di grandezze indipendenti è data dal prodotto delle probabilità delle singole variabili.
Abbiamo visto introducendo la Multiple Linear Regression come la soluzione, cioè la stima della variabili X, si ottiene invertendo (o meglio pseudoinvertendo) la matrice degli osservabili Y. Tale operazione risulta possibile solo se il rango della matrice Y è massimo cioè se il numero di colonne linearmente indipendenti coincide con il numero di colonne della matrice. Questo equivale a dire 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" in quanto le colonne di Y tendono a co-variare. E' da notare che i termini correlato, colineare e covariante sono sostanzialmente equivalenti in quanto descrivono l'esistenza di una relazione tra le variabili.
La colinearità si esprime quantitativamente attraverso la matrice di covarianza. Infatti, in caso di colinearità i termini non diagonali della matrice di covarianza sono diversi da zero. Per risolvere il problema MLR è quindi necessario rimuovere la colinearità, cioè ridurre la matrice di covarianza in forma diagonale introducendo delle nuove variabili latenti.
Prima di calcolare queste nuove variabili latenti è opportuno offrire degli esempi a riguardo della matrice di covarianza. Per poterli rappresentare graficamente consideriamo sempre una distribuzione bivariata. La figura 4.7 mostra la forma della ellissi di isoprobabilità associata alla matrice di covarianza e la forma della matrice di covarianza stessa.



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

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

figura 4.7

La matrice di covarianza può essere scritta in forma diagonale con un adeguato cambiamento del sistema di riferimento. Tale operazione coincide con la scrittura in forma canonica della ellissi associata alla matrice di covarianza.
Il sistema di riferimento che rende canonica la ellisse è formato dagli autovettori della matrice di covarianza, cioè dagli assi principali dell'ellisse costruita come forma quadratica dalla matrice di covarianza stessa. Le nuove variabili corrispondono al primo caso della figura precedente con variabili scorrelate e la PDF ottenuta come il prodotto di PDF monovariate.
Le variabili del nuovo sistema di riferimento ovviamente non sono più degli osservabili fisici (oggetto di misurazioni) ma si ottengono come combinazioni lineari di queste.
Le variabili sono chiamate Componenti Principali e l'insieme di procedure di calcolo e interpretazione delle componenti principali è generalmente chiamata analisi delle componenti principali (PCA).
Il cambio di coordinate è così descritto:

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

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

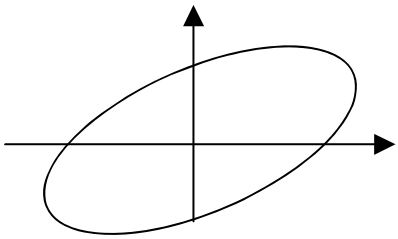
E' opportuno richiamare sia le definizioni che le proprietà più importanti del calcolo degli autovalore ed autovettori.
Data una matrice A quadrata e di dimensione n definiamo v autovettore di A se esiste uno scalare λ (□□□□□□□□□□□□) tale che:

$$A \cdot v = \lambda \cdot v$$

Il calcolo degli autovettori ed autovalori proviene dallo sviluppo della relazione precedente, attraverso la cosiddetta equazione secolare

$$A \cdot v = \lambda \cdot v \Rightarrow \left(A - \lambda I\right)v = 0 \Rightarrow \left|A - \lambda I\right| = 0$$

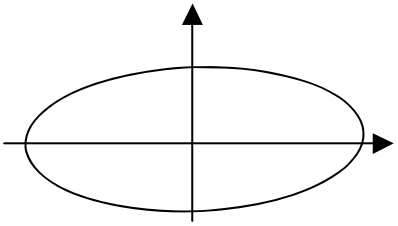
Se A è non singolare allora tutti gli autovettori sono diversi da zero. Inoltre, se A è simmetrica tutti gli autovalori sono distinti ed ortogonali tra loro. Questa ultima proprietà consente di considerare il sistema di riferimento formato dagli autovettori di una matrice di covarianza (simmetrica) come ortogonale. Questa proprietà è fondamentale per la PCA.
Per comprendere in maniera semplice il significato degli autovettori consideriamo la definizione precedente e consideriamo A come una trasformazione (operatore) che applicata ad un vettore x ne produce uno y=Ax. Tra tutti i vettori x a cui A si applica gli autovettori sono quei vettori per i quali la trasformazione agisce solo in termini di scala (prodotto vettore scalare) senza cambiarne l'orientamento nello spazio vettoriale. In pratica, gli autovettori sono quindi degli in varianti per la trasformazione A, nel senso che A non ne modifica la direzione.
Ad esempio, data una matrice di rotazione nel piano xy, la direzione z essendo non variata dalla rotazione risulta essere un autovettore della matrice.
Il calcolo degli autovettori in geometria è utilizzato ad esempio per determinare il sistema di riferimento nel quale una quadrica si scrive in forma canonica. Consideriamo una ellisse generica descritta da una equazione implicita.



$$a \cdot x^2 + 2b \cdot xy + c \cdot y^2 = k$$

Figura 4.7

L'equazione diventa scritta in forma canonica cambiando il sistema di coordinate considerando come base gli assi principali dell'ellisse.



$$\lambda_1 \cdot x^2 + \lambda_2 \cdot y^2 = k$$

Figura 4.8

Gli assi principali dell'ellisse coincidono con gli autovettori della matrice associata e gli autovalori sono proprio i termini λ che compaiono nella equazione canonica.

Mostriamo ora un esempio di calcolo di diagonalizzazione di una matrice simmetrica associata ad una ellisse.

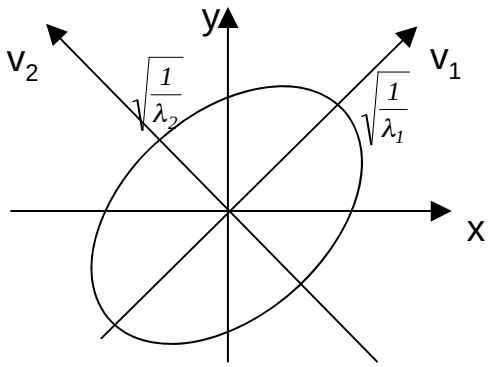


Figura 4.9

equazione in forma implicita: $13 \cdot x^2 - 10 \cdot x \cdot y + 13 \cdot y^2 = 1$

forma quadratica associata $\Rightarrow \begin{bmatrix} x & y \end{bmatrix} \cdot \begin{bmatrix} 13 & -5 \\ -5 & 13 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 1$

equazione secolare: $A \cdot v = \lambda \cdot v \Rightarrow (A - \lambda \cdot I) \cdot v = 0 \Rightarrow det(A - \lambda \cdot I) = 0$

calcolo autovalori: $det \begin{bmatrix} 13-\lambda & -5 \\ -5 & 13-\lambda \end{bmatrix} = 0 \Rightarrow (13-\lambda)^2 - 25 = 0$

$\lambda^2 - 26 \cdot \lambda + 144 = 0 \Rightarrow \lambda_1 = 8;\ \lambda_2 = 18$

calcolo autovettori

$$(A - \lambda_1 \cdot I) \cdot v_1 = 0 \Rightarrow \begin{bmatrix} 13-8 & -5 \\ -5 & 13-8 \end{bmatrix} \cdot v_1 = 0 \Rightarrow \begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix} \cdot v_1 = 0 \Rightarrow v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$(A - \lambda_2 \cdot I) \cdot v_2 = 0 \Rightarrow \begin{bmatrix} 13-18 & -5 \\ -5 & 13-18 \end{bmatrix} \cdot v_2 = 0 \Rightarrow \begin{bmatrix} -5 & -5 \\ -5 & -5 \end{bmatrix} \cdot v_2 = 0 \Rightarrow v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Come noto la quadrica descrive sia l'ellisse che l'iperbole il tipo della curva dipende dai coefficienti della forma canonica. Per l'ellisse entrambe i coefficienti devono essere maggiori di zero. Questa condizione è sempre rispettata per una forma quadratica generata da una matrice simmetrica.

Definito il contesto matematico della PCA ricordiamo che lo scopo generale della PCA è quello di rappresentare 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 abbiamo visto si ottiene con una rotazione delle coordinate nella base degli autovettori che costituiscono le componenti principali. Ad ogni autovettore è associato un autovalore. L'autovalore descrive la "elongazione" della ellisse lungo la componente principale, tale elongazione è associata alla varianza della componente principale stessa.
E' importante notare che se le variabili originarie sono parzialmente correlate tra loro alcuni autovalori dovranno avere un valore trascurabile. In questo caso gli autovettori corrispondenti a gli autovalori minori possono essere trascurati e si può limitare la rappresentazione dei dati solo a quegli autovettori i cui autovalori hanno il valore più grande.
La PCA è un metodo del secondo ordine poiché sia le nuove coordinate che il criterio per la riduzione delle dimensioni sono basate unicamente sulle proprietà della matrice di covarianza. Oltre alla media, infatti la covarianza è l'unica grandezza che descrive una variabile distribuita normalmente.
La media è in genere resa nulla prima del calcolo delle componenti principali e quindi tutta l'informazione è contenuta nella matrice di covarianza. Quindi la PCA si basa sulla ipotesi che la variabile x sia distribuita normalmente
Solo nel caso di distribuzione normale le componenti principali risultano indipendenti e la probabilità multivariata diventa il prodotto delle probabilità univariate Nel caso in cui le variabili non sono distribuite normalmente le componenti principali risultano solo non correlate.
Ricordiamo la definizione dei concetti statistici di correlazione ed indipendenza. Due variabili X e Y sono dette non correlate se E[X Y]=E[X] E[Y], cioè il valore aspettato del prodotto delle due è il prodotto dei valori aspettati delle singole variabili. L'indipendenza coinvolge invece la distribuzione di probabilità, P(X Y)=P(X) P(Y), cioè la probabilità del prodotto è pari al prodotto delle singole probabilità.

Data una matrice X le componenti principali sono gli autovettori (**P**) della matrice di covarianza, la matrice di covarianza è data da: X^TX. I

coefficienti degli autovettori nella base originale degli osservabili sono detti loadings.
I dati della matrice X, espressi nel sistema di riferimento degli osservabili, si trasformano nella base degli autovettori P attraverso la proiezione degli stessi. Queste coordinate si chiamano scorse e si calcolano come T=XP.
La matrice **X** risulta quindi decomposta nel prodotto dei **P** (loadings) e delle coordinate dei patterns originali nella base degli **T** (scores): $X=TP^T$
Gli autovalori sono proporzionali alla varianza dei dati lungo la direzione principale identificata dall'autovettore corrispondente.

La figura seguente illustra il metodo della analisi delle componenti principali. Si inizia con la rappresentazione dei dati nello spazio degli osservabili. La matrice di covarianza è non diagonale e di conseguenza la ellisse, il luogo di punti di isoprobabilità, è in forma non canonica. Cioè gli assi principali non coincidono con gli assi del sistema di riferimento. Il calcolo della PCA consiste nel trovare gli autovettori e gli autovalori corrispondenti. Data X e determinati gli autovettori P, attraverso l'operazione XP si determinano le coordinate dei singoli dati nella base delle componenti principali. Il plot dei dati nella base delle componenti principali si chiama score plot. In questa base, la matrice di covarianza dei dati diventa diagonale. Cioè i dati sono correlati. L'analisi degli autovalori consente di studiare la distribuzione della varianza globale dei dati nelle varie componenti principali. Il plot degli autovalori prende il nome di scree plot. L'analisi degli autovalori consente di individuare gli autovettori maggiormente significativi, cioè quelli che portano la varianza maggiore. Ridurre la rappresentazione dei dati a solo quelle componenti principali che portanto una quantità significativa di varianza consente di ridurre la dimensione dello spazio di rappresentazione.

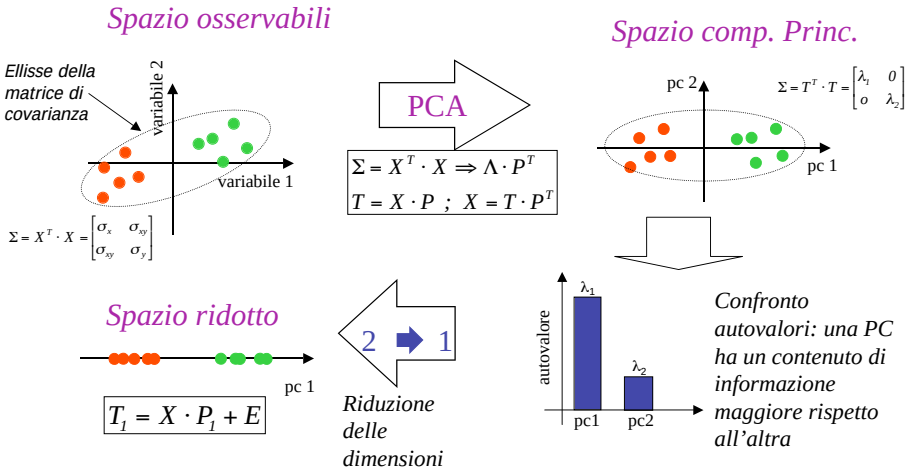


Figura 4.10

Scorse e Loadings sono il risultato maggiore della PCA. Ricordiamo che gli score sono le coordinate dei dati nella base delle componenti principali mentre i loadings sono le coordinate degli autovettori. In particolare, studiando i loadings è possibile determinare l'importanza di ogni variabile nell'insieme multivariato.
Come detto precedentemente, l'analisi degli autovalori consente di troncare lo sviluppo di una matrice di dati ad un numero di componenti minore del valore massimo che coincide con il numero N delle colonne di X cioè delle variabili multivariate. Tale sviluppo è scritto come: $X=TP^T+res$. Dove res è il residuo.

E' importante riflettere sul significato delle componenti principali. Dato un set di dati le cui variabili sono parzialmente correlate, le direzioni principali di massima varianza sono quelle che colgono le direzioni di massima correlazione. In pratica le componenti principali descrivono le direzioni di correlazione dei dati, procedendo ind direzioni ortogonali tra loro. Le componenti i cui autovalori hanno il valore più piccolo descrivono quindi le direzioni di correlazione minore, o nulla. Eliminare dalla matrice originale tali direzioni significa quindi eliminare le parti di scarsa o nulla correlazione. In questa parte dei dati rientra sicuramente il rumore intrinseco di ogni variabile. Infatti, per sua natura il rumore di ogni variabile è non correlato con le altre variabile. Per qui lo sviluppo di una matrice di

dati con la PCA è anche un metodo per filtrare il rumore non correlato. Ovviamente, il metodo non distingue il rumore dall'informazione significativa. Per cui se una delle variabili porta un informazione non correlata con il resto delle variabili, tale informazione sarà necessariamente reiettata nelle componenti minori. Per questo motivo la PCA risulta efficace solo se le variabili del problema hanno una correlazione significativa.
Tornando al filtraggio del rumore, questo fu uno dei primi utilizzi pratici del metodo. Come esempio consideriamo la sequenza di spettri delle figure 4.11, 4.12 e 4.13 relativi alla anisotropia dello spettro di riflessione in film organici.

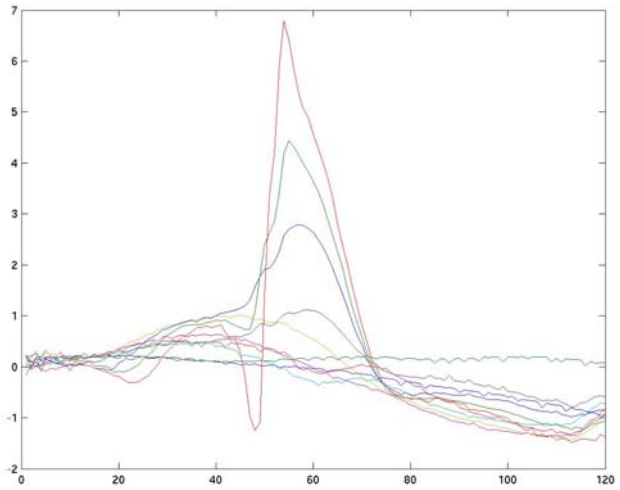


Figura 4.11: *spettri originali*

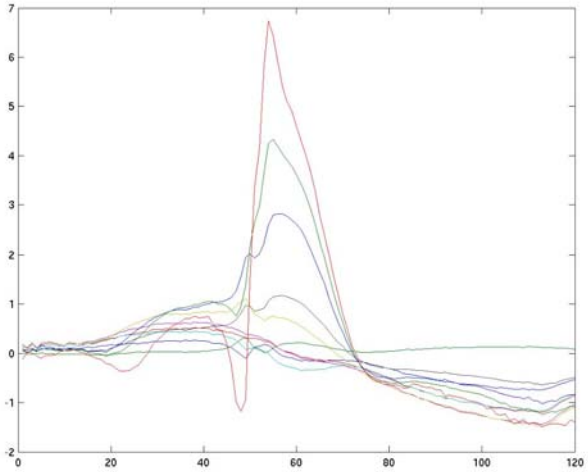


Figura 4.12: spettri sviluppati alle prime 3 componenti principali

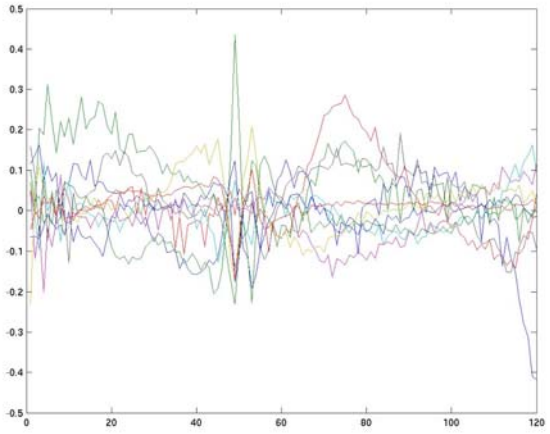


Figura 4.13: *residui degli spettri.*

4.4 Principal Component Regression (PCR e Partial Least Squares (PLS)

La PCA ed in particolare gli scores sono utili per la soluzione del problema della MLR. Lo sviluppo in componenti principali elimina la

colinearità dei dati e permette un calcolo sicuro della matrice di regressione.
La figura 4.13 mostra la procedura del metodo di regressione detto PCR (Principal Component Regression).

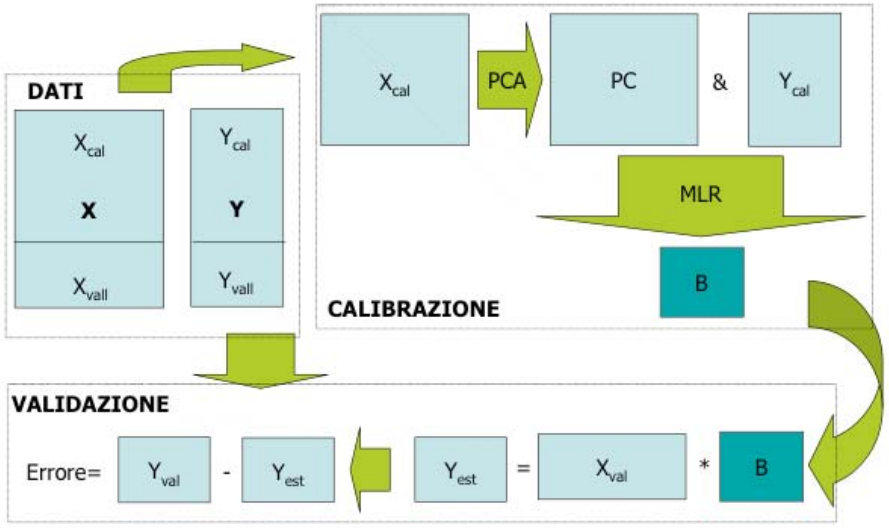


figura 4.13

La PCR coincide con la MLR per il calcolo della matrice B tranne per il fatto che B è calcolata non con i dati originali ma con le componenti principali, quindi con un set di dati non correlati e quindi sicuramente invertibili.
Si tenga comunque presente che la PCA non riporta il problema a quello degli osservabili selettivi perché le componenti principali non sono osservabili fisici.

La PCR non è comunque la soluzione ottimale del problema. Infatti le componenti principali sono calcolate indipendentemente dal tipo di variabili da determinare. Un passo ulteriore consiste nel considerare la PCA non solo della matrice X ma anche della matrice Y. In questo modo le componenti principali di Y vengono ruotate per massimizzarne la correlazione con le componenti principali di X questo metodo è detto **Partial Least Squares** (PLS) e rappresenta il

modo più avanzato per risolvere il problema della regressione lineare multipla.
Le componenti ruotate non coincidono più con le componenti principali e vengono dette variabili latenti.
PLS "funziona meglio" di PCR perchè le variabili latenti sono scelte massimizzando la correlazione con le PC della matrice Y. Le componenti principali infatti sono le direzioni di massima varianza, queste direzioni non dipendono dallo "scopo" (matrice Y) ma sono determinate univocamente dalla matrice X Le componenti principali quindi non sono di per se significative per la costruzione di un modello che stima Y, agevola solo la MLR eliminando la correlazione tra le variabili.

L'algoritmo della PLS è una tecnica di analisi che combina le caratteristiche dell'Analisi alle componenti principali con la regressione lineare. Tale tecnica diviene molto utile quando bisogna predire un insieme di variabili dipendenti a partire da un numero elevato di variabili indipendenti (predittori).
Come abbiamo visto quando nei paragrafi precedenti quando si parlava di Regressione multipla e di PCR, il problema può essere descritto prendendo in esame le seguenti condizioni generali.
Consideriamo un esperimento caratterizzato da M osservazioni. Per ognuna di queste osservazioni vengono registrate i valori delle variabili dipendenti (Y) e delle variabili indipendenti (X) .
L'obbiettivo della PLS è quello di predire le variabili Y a partire dalla sola conoscenza delle variabili indipendenti X, in formula
dove B è la matrice di regressione ed E è quella dei residui.
Quando i dati contenuti nella matrice X sono colineari, la matrice sarà singolare e la soluzione del problema non potrà essere risolta con la regressione lineare. Diversi approcci sono stati sviluppati per risolvere questo tipo di problema. Un approccio può essere quello della Principal Component Regression che riduce il set di predittori , utilizzando al posto della X le componenti principali ottenute dalla stessa matrice applicando PCA. La scelta della componenti principali elimina il problema della colinerarità, ma non risolve il problema della scelta del sottoinsieme ottimo di predittori della matrice X. Infatti le PCs di X possono non essere sufficienti per trovare la massima correlazione con Y.
La PLS , invece, trova quell'insieme di componenti di X, dette variabili latenti, che rappresentano una decomposizione della stessa matrice e che allo stesso tempo massimizzano la loro correlazione

con le variabili dipendenti Y. In formule X è decomposta nel seguente modo:

$$X = T \cdot P^T + E \quad con \quad P \cdot P^T = I$$

dove I è la matrice identità , T è la matrice degli scores , P quella dei Loadings ed E quella dei residui. La matrice T contiene le nuove coordinate di X nel nuovo spazio descritto dalle variabili latenti.
Nello stesso modo possiamo decomporre Y nello stesso modo utilizzando la stessa matrice degli scores T di X:

$$Y = T \cdot C^T + F \quad con \quad C \cdot C^T = I$$

dove C è la matrice dei loadings ed F quella dei residui.
Tenendo presente le equazioni precedenti possiamo scrivere la matrice delle variabili dipendenti come:

$$Y = X \cdot W \cdot C^T + F = X \cdot B + F \quad con \quad B = W \cdot C^T$$

dove la matrice W viene determinata attraverso la seguente relazione:

$$T = X \cdot W .$$

Lo scopo della PLS è quello di trovare le matrici T e C in modo da massimizzare la covarianza tra X ed Y.

Interpretazione geometrica della PLS
La Partial Least Square, cosi' come la regressione multipla, rappresenta un metodo di proiezione. Infatti questa tecnica puo' essere considerata come una proiezione dei vettori della matrice X in un sottospazio la cui dimensione è definita dalle k variabili latenti che massimizzano la covarianza tra X ed Y. Le coordinate di tale proiezione rappresentano dei buoni predittori di Y come indicato nella figura 4.14.

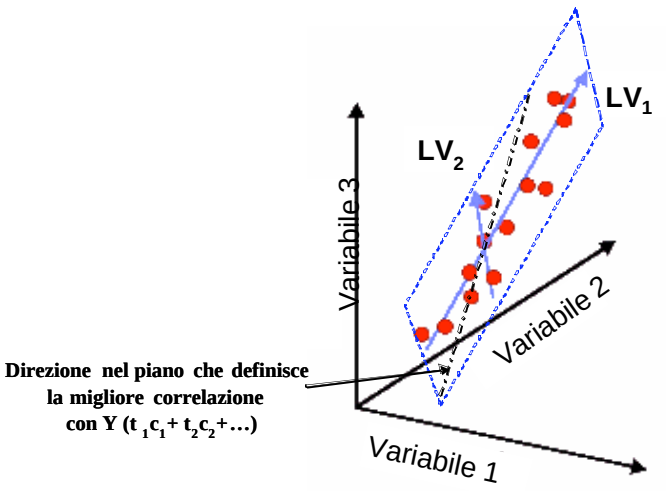


Figura 4.14

L'errore di predizione in PLS

Contrariamente a quanto avviene per la PCR, il metodo della PLS è soggetto ad overfitting, ovvero l'errore sul set di calibrazione è inferiore rispetto all'errore relativo ai dati di validazione.
L'overfitting definito nel capitolo 3 era dato dall'utilizzo di funzioni non lineari che riducono l'errore tra modello e dati in fase di calibrazione. Nel caso della PLS l'overfitting riguarda il numero delle variabili latenti utilizzate nel modello. In particolare, l'overfitting deriva dal processo di decomposizione nel quale procedendo per proiezioni su sottospazi ortogonali si adatta il modello ai dati di calibrazione riducendo l'errore.
Per evitare l'overfitting è quindi necessario ottimizzare il numero della variabili latenti attraverso un processo di cross-validation.
La bontà del modello ottenuto con la PLS puo' essere valutato attraverso il calcolo dell'errore quadratico medio ottenuto durante la fase di calibrazione (Root Mean Square Error of Calibration: RMSEC) e durante la fase di validazione (Root Mean Square Error of Cross Validation: RMSECV).

Nella figura 4.15 è rappresentato un possibile andamento in funzione del numero di variabili latenti del modello sia dell'errore di calibrazione (RMSEC) sia l'errore di validazione (RMSECV).

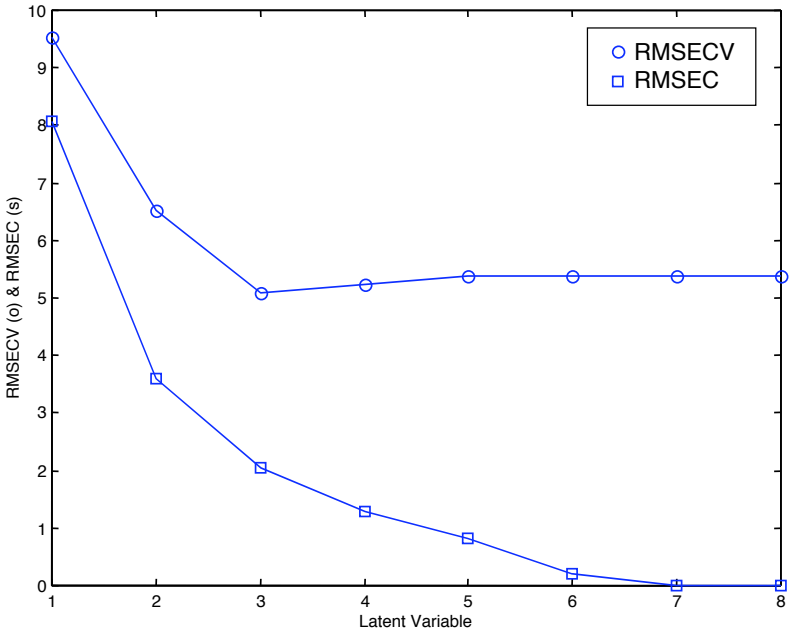


Figura 4.15

Si osservi come l'errore di calibrazione diminuisce monotonicamente all'aumentare del numero di variabili latenti, mentre l'errore di validazione presenta un valore minimo.
L'andamento monotono di RMSEC indica come incrementando il numero di variabili latenti il modello tende a inglobare anche il rumore dei dati. RMSECV d'altro canto fornisce l'andamento dell'errore di predizione, ovviamente in fase di predizione la quantità di rumore dei dati sarà sicuramente differente quindi la parte di modello che si adatta sul rumore dei dati di calibrazione provoca sicuramente un errore quando viene utilizzata con dati con rumore differente.
Il valore minimo di RMSECV indica proprio il numero di variabili latenti per il quale è massima la descrizione della parte deterministica dei dati ed oltre il quale il modello inizia a rappresentare il rumore dei dati di calibrazione.

Il numero di variabili latenti in corrispondenza dei quali si ottiene il minimo di RMSECV indica le variabili latenti ottimali da considerare per il modello.

La analisi dei loadings

Cosi' come per la Principal Component Analysis anche per la Partial Least Squares i loadings forniscono informazioni su quali variabili (colonne della matrice X) contribuiscono maggiormente ad ogni singola variabile latente e di conseguenza all'intero modello. Sulla base di queste informazioni si può ridurre il numero delle variabili riducendo la complessità del modello ma mantenendone invariate le prestazioni.
Illustriamo la analisi dei loadings in PLS attraverso un esempio.

Misura della concentrazione di propanolo ed etanolo con una matrice di quattro sensori di gas.

Consideriamo in questo esempio una matrice di quattro sensori, tipo microbilancia al quarzo, funzionalizzati con polisilossani modificati. La matrice di sensori è esposta a miscele di vapori di propanolo ed etanolo a concentrazione variabile e nota. Si vogliono utilizzare i dati di calibrazione per costruire un modello PLS in grado di predirre, dai segnali dei sensori, i valori della concentrazione dei due alcoli.
La tabella seguente riporta i valori dei segnali sensoriali. Le prime due colonne della tabella costituiscono la matrice Y e le restanti colonne formano la matrice X.

Concentr. Propanolo	[ppm[Etanolo	Segnali Sensore 1	[Hz] Sensore 2	Sensore 3	Sensore 4	Sensore 5
600.0	400.0	-5.5	-41.5	-25.9	-19.3	-15.6
800.0	300.0	-5.8	-49.1	-31.8	-23.6	-19.1
400.0	400.0	-3.6	-26.9	-18.0	-13.1	-11.1
800.0	0	-5.8	-46.9	-31.4	-21.1	-15.5
200.0	600.0	-2.9	-18.3	-12.1	-8.9	-8.1
800.0	800.0	-7.6	-61.5	-38.2	-27.5	-22.5
0	400.0	-1.1	-5.4	-3.6	-2.0	-2.1
600.0	600.0	-5.9	-49.2	-27.2	-20.1	-16.6
800.0	400.0	-6.5	-54.3	-32.4	-24.3	-19.5
400.0	0	-2.9	-21.5	-14.5	-10.1	-7.1
200.0	300.0	-1.7	-14.9	-9.0	-6.7	-6.0
600.0	800.0	-6.1	-49.6	-29.2	-21.9	-17.3
0	600.0	-1.5	-9.5	-5.4	-3.8	-4.0
400.0	800.0	-4.7	-41.4	-22.4	-17.1	-13.5
200.0	0	-1.9	-9.4	-6.6	-4.8	-4.9
400.0	600.0	-4.9	-36.3	-19.5	-15.1	-11.4

800.0	600.0	-6.7	-57.4	-33.7	-25.8	-19.9
0	300.0	0	-2.9	-2.5	-1.8	-1.5
200.0	400.0	-2.6	-19.7	-10.0	-8.4	-6.1
600.0	0	-4.1	-38.1	-21.5	-16.0	-13.3
0	800.0	-2.8	-13.6	-8.4	-5.8	-5.0
400.0	300.0	-3.2	-30.6	-16.9	-12.5	-10.4
600.0	300.0	-4.6	-41.0	-23.8	-18.0	-13.4
200.0	800.0	-4.1	-24.7	-14.8	-10.1	-9.1

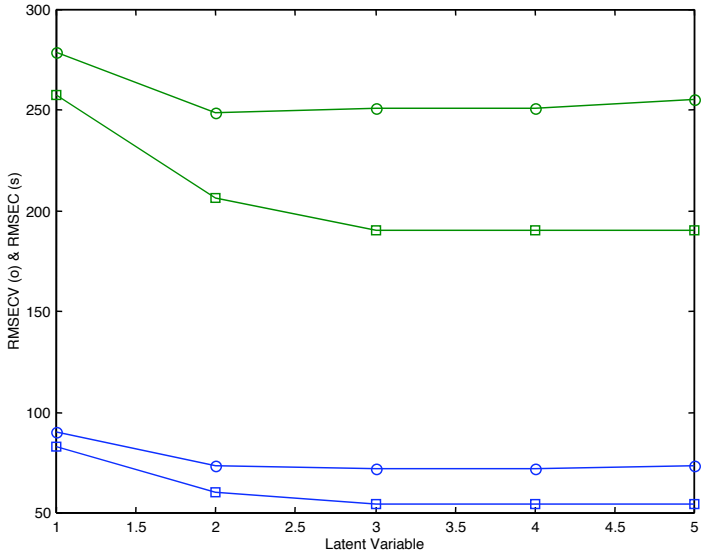


Figura 4.16

In figura 4.16 sono graficati i valori di RMSEC ed RMSECV per i due vapori ed in funzione del numero di variabili latenti considerate. Il modello mostra una maggiore accuratezza nella determinazione del propanolo. Il numero di variabili latenti che minimizza l'errore di predizione è uguale a tre.
È interessante confrontare i valori numerici degli errori di calibrazione e di predizione per osservare che i risultati ottenuti in calibrazione sono sempre peggiori delle prestazioni che la combinazione modello più sensori darà nel suo utilizzo pratico.
RMSEC è minimo considerando tutte e cinque le variabili, i valori sono 54 ppm e 190 ppm per propanolo ed etanolo rispettivamente. L'errore di predizione (RMSECV) invece risulta minimo con tre variabili con valori pari a 72 ppm per il propanolo e 250 ppm per l'etanolo.

Il modello PLS ottimale risulta quindi costituito da tre variabili latenti. Nella figura 4.17 vengono riportati i valori dei loadings per ciascuna delle tre variabili.

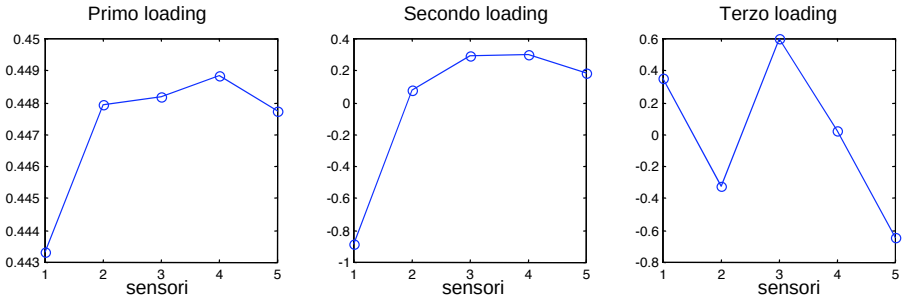


figura 4.17

è importante sottolineare che le tre variabili latenti non hanno la stessa importanza, infatti osservando la figura 4.16 possiamo notare come il grosso della quantificazione è ottenuto con le prime due variabili latenti, e che la terza riduce l'errore di una piccola quantità. Per quantificare l'effetto della terza variabile possiamo dire che l'aggiunta della terza variabile riduce l'RMSECV di circa 1 ppm.
In figura 4.17 osserviamo come i loadings relativi ai sensori 2, 3, 4, e 5 siano sostanzialmente simili. Possiamo ipotizzare di ridurre l'intera matrice a solo tre sensori: il sensore 1 e due sensori che rappresentino gli altri quattro, ad esempio la coppia 2 e 4. Se si costruisce un modello di regressione con il trio di sensori 1-2-4 si ottiene il PRESS graficato in figura 4.18 in cui il minimo dell'errore viene ottenuto con due variabili.

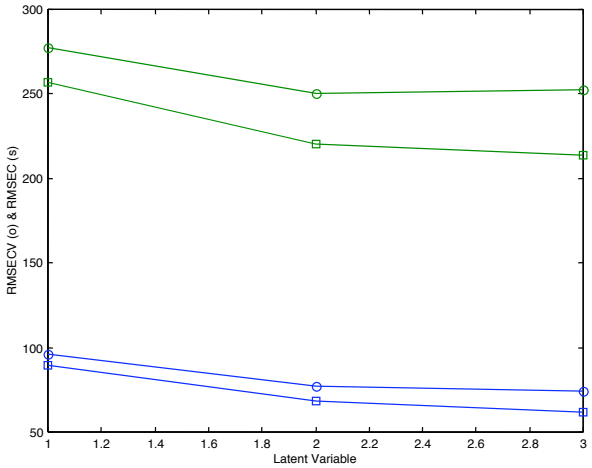


Figura 4.18

Con il modello PLS costruito con tre sensori si ottiene un errore di validazione pari a 77 ppm e 250 ppm per propanolo ed etanolo, cioè di prestazioni sostanzialmente uguali a quelle ottenute usando cinque sensori. La selezione delle variabili della matrice X è una fase importante per quanto riguarda l'ottimizzazione di un sistema di misura multivariato. Possiamo dire che esso consente di semplificare sia l'hardware che il software del sistema di misura mantenedo inalterata l'accuratezza della stima.

4.5 Esempio di PLS: misura di acidità ed umidità nelle nocciole attraverso la spettroscopia nel vicino infrarosso (NIR)

La spettroscopia NIR è un metodo estremamente potente e semplice per misurare alcuni componenti importanti nei vegetali. In questo esempio consideriamo gli spettri NIR di 36 campioni di farina di nocciole di differenti specie. Lo scopo della analisi è la misura della acidità totale e dell'umidità contenute nei frutti.
La figura 4.19 mostra l'insieme degli spettri.

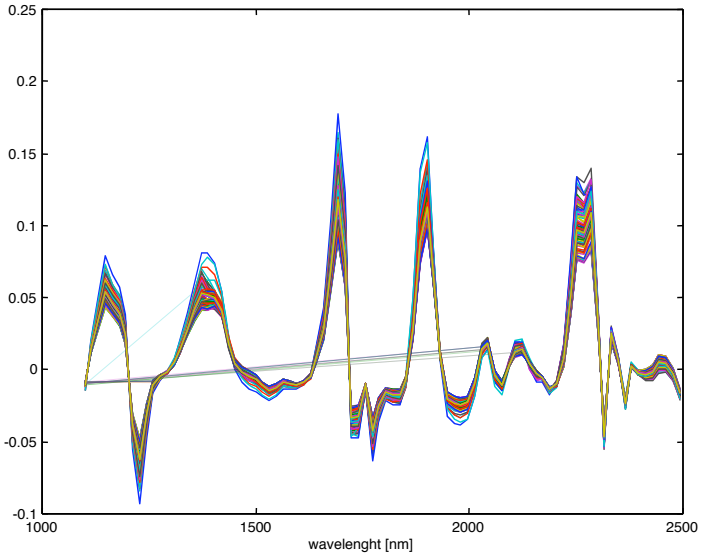


Figura 4.19

Per calibrare un modello è necessario come abbiamo già discusso misurare i parametri obiettivo del modello con un metodo di maggior accuratezza che funga da metodo di riferimento. In questo caso acidità totale ed umidità sono state misurate con i metodi analitici standard. I dati sono stati centrati, cioè resi a media nulla.
Il modello PLS calcola l'effetto delle variabili latenti sulla accuratezza del modello. Il problema è quello di scegliere un numero di variabili latenti che eviti l'overfitting dei dati. Attraverso il metodo di leave-one-out si calcola per ogni variabile latente l'errore statisticamente atteso su un set di validazione indipendente, cioè l'errore che ci si può aspettare utilizzando il modello per stimare i valori di umidità e di acidità in farine di nocciole diverse da quelle usate per la calibrazione. Il PRESS calcolato sull'insieme di calibrazione quello derivato dal leave-one-out viene graficato in figura 4.20 per valutare l'accuratezza del modello.

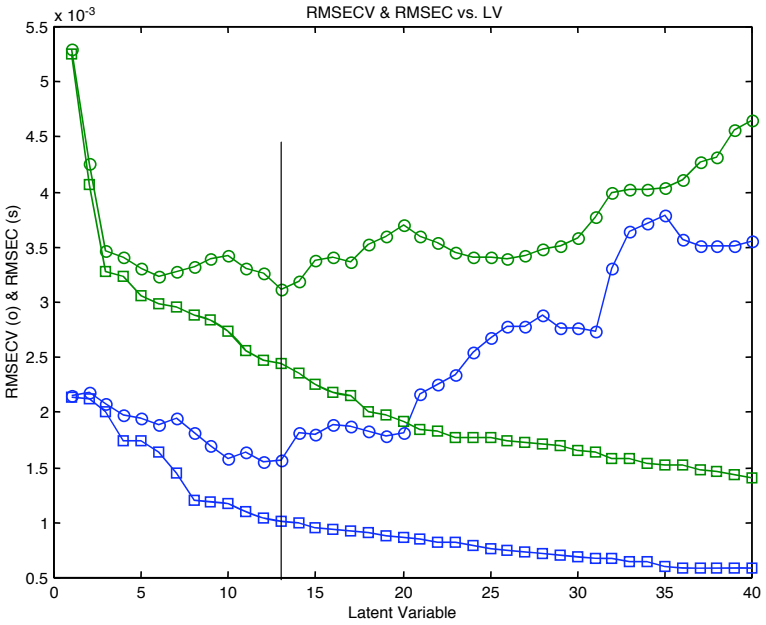


Figura 4.20

Nella figura precedente il PRESS relativo al set di calibrazione è indicato con un quadrato (RMSEC) mentre il cerchio indica il PRESS calcolato dal leave-one-out (RMSECV).
Ovviamente l'errore di cross-validazione è maggiore di quello compiuto sul set di calibrazione. Nel grafico seguente si osserva che con 13 variabili latenti si ottiene il minimo di errore di validazione. In questo esempio si ottiene:

	RMSEC	RMSECV
Acidità totale	0.0010	0.0016
Umidità	0.0024	0.0031

Le prestazioni del modello sul set di calibrazione possono essere raffigurate con lo scatter plot. Il grafico in cui si rappresenta la variabile stimata in funzione della variabile vera è rappresentato in figura 4.21. Nel caso di perfetta regressione i dati sono disposti lungo la bisettrice a 45° del piano cartesiano.

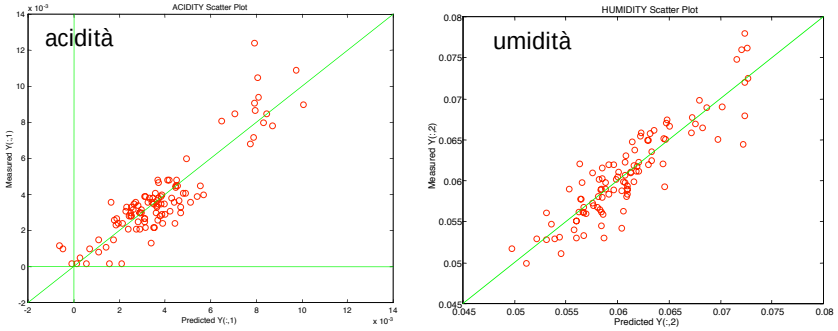

ACIDITY Scatter Plot
acidità
Measured Y(:,1)
Predicted Y(:,1)
x 10^-3
HUMIDITY Scatter Plot
umidità
Measured Y(:,2)
Predicted Y(:,2)

Figura 4.21

5

ELEMENTI DI PATTERN RECOGNITION

La analisi dei pattern ha come obiettivo l'assegnazione di una sequenza di dati ad una classe. Il linguaggio più opportuno per inquadrare la analisi dei pattern è quello degli insiemi. Un insieme è definito come una classe alla quale appartengono degli elementi che hanno in comune alcune caratteristiche. Uno degli aspetti importanti della analisi dei pattern è proprio il fatto che elementi (patterns) diversi possono essere assegnati a classi differenti a seconda della caratteristica che viene messa, di volta in volta, in evidenza.
Gli elementi fondamentali della analisi dei pattern sono i seguenti:

- Pattern: serie di descrittori (features) che definiscono le proprietà di un soggetto complesso.
- Classe: insieme a cui un soggetto appartiene.
- Classificazione: operazione matematica per cui un campione, descritto da una serie di features, viene assegnato ad una particolare classe.
- L'insieme delle features è detto Pattern, l'operazione di classificazione si dice Pattern Recognition.

È chiaro che le features utilizzate devono essere importanti per il tipo di classificazione che si vuole operare. In particolare, le features rappresentano il tipo di informazioni che possiamo estrarre da un campione per dettagliarne le caratteristiche.
La pattern recognition è un procedimento mentale che a livello sia conscio sia inconscio regola la vita degli essere viventi, ed in

particolare degli umani. E' infatti esperienza comune come gli uomini tendono a classificare le esperienze in categorie.
Esempio di pattern sono ad esempio i descrittori chimico-fisici dei frutti, come ad esempio peso, forma, colore, concentrazioni di zuccheri e di acidi, …. Questi indicatori ci consentono di classificare i frutti a seconda della specie, della cultivar (ovvero del tipo), del grado di maturità, fino al valore merceologico.
Come esempio di pattern, osserviamo un caso attinente alla vita quotidiana, le acque minerali. Sulla etichetta delle acque minerali, per legge, viene riportata la composizione ionica dell'acqua stessa. Tale composizione forma un pattern (che possiamo chiamare profilo ionico). L'intero pattern definisce l'appartenenza dell'acqua minerale ad una classe che ne definisce le caratteristiche generali come vediamo nella figura 5.1.

figura 5.1

Il modo più semplice per visualizzare un pattern è l'istogramma. L'istogramma (o grafico a barre) consente una immediata visualizzazione delle similitudini e differenze tra i vari patterns. La figura 5.2 mostra i pattern ionici di varie marche di acque minerali.
Ogni grafico mostra un'acqua, le specie a cui corrispondono le singole barre sono mostrati in fondo alla figura. Una prima analisi quantitativa, fatta semplicemente guardando alle scale dei grafici, individua subito le acque a basso contenuto ionico (le acque oligominerali) inoltre, la composizione delle acque, cioè la forma del

pattern rivela un certo numero di peculiarità. Ovviamente il contenuto dell'acqua non è standard in quanto dipendente dal tipo di rocce presenti nella regione della sorgente.
Tuttavia, è evidente come la analisi degli istogrammi non consente di definire con certezza la appartenenza o la non appartenenza di due acque ad un insieme comune. Per far questo serve un approccio che traduca il pattern in una entità matematica e che quindi consenta di stabilire la differenza tra pattern.
La pattern recognition è l'insieme delle procedure volte ad operare una classificazione matematica dei pattern. Il punto di partenza di ogni tecnica di pattern recognition è la assegnazione del pattern ad una entità matematica per la quale sia definibile un algebra ovvero una serie di operazioni che consentono di calcolare, ad esempio, la differenza tra pattern.

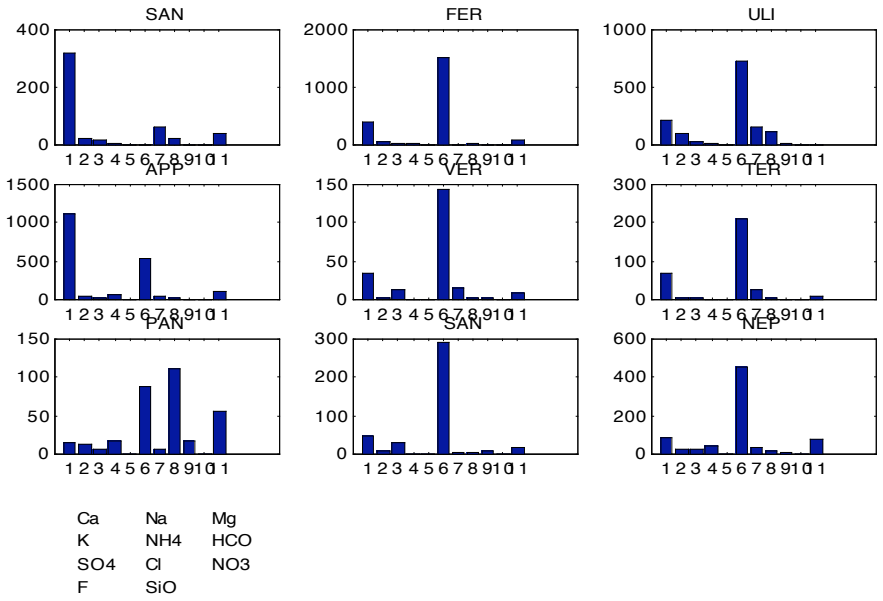


Figura 5.2

L'approccio più semplice consiste nel tradurre ogni pattern in un vettore, cioè in una sequenza ordinata di grandezze (le feature). Vedremo più avanti che tali vettori saranno considerati come elementi di uno spazio: lo spazio dei pattern, che per le tecniche di

pattern recognition più utilizzate coincide con uno spazio vettoriale euclideo.
Le classi di un problema di pattern recognition sono chiamate classi di appartenenza (membership classes). Le classi di appartenenza sono classi teoriche i cui elementi condividono una o più caratteristiche globali. Pattern uguali possono essere classificati in differenti classi a seconda del tipo di classificazione di interesse. Ad esempio consideriamo la classificazione dei vini mostrata in figura 5.3. Questi possono essere distinti in differenti classi relative ad esempio al colore del vino (bianco o rosso) alla regione di provenienza, alla denominazione, o ancora ad alcune caratteristiche come il fatto se un vino sia secco oppure frizzante.

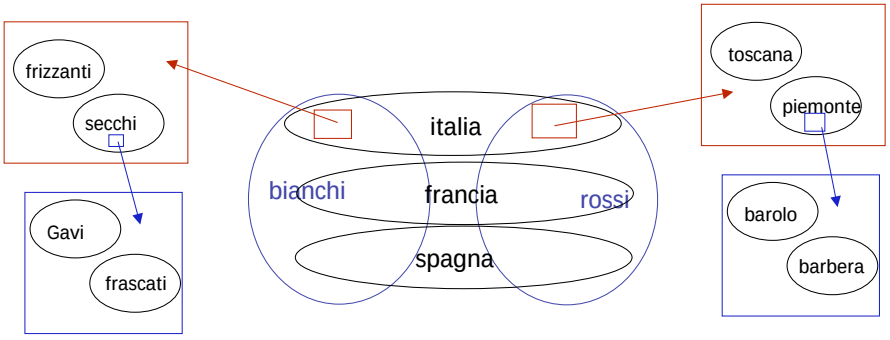


Figura 5.3

Ovviamente, per ogni classificazione ci saranno delle features che saranno più appropriate, nel senso che forniscono informazioni pertinenti allo scopo e features ridondanti che non aiutano alla classificazione.
In generale, possiamo dire che allo scopo di classificare correttamente dei patterns, dobbiamo essere una certa similitudine tra i pattern, o perlomeno tra quelle features che compongono il pattern e che sono rilevanti per la classificazione voluta.

5.1 Istogrammi e radar plots
La similitudine, abbiamo visto prima può essere studiata in termini qualitativi dall'osservazione della rappresentazione grafica dei pattern. Gli istogrammi di figura 5.4, ad esempio, ci consentono di distinguere facilmente le acque oligominerali dalle acque minerali.

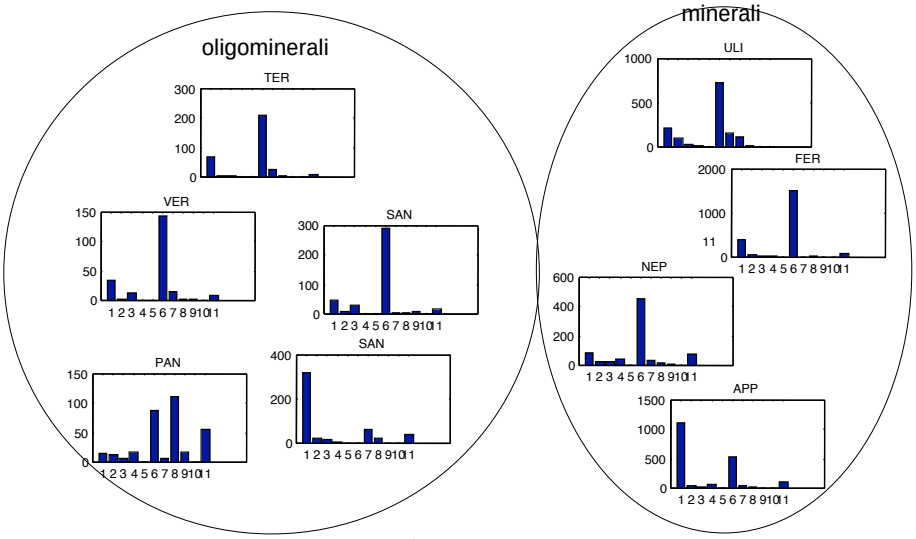


Figura 5.4

Tale metodo però non consente di stabilire in maniera certa i confini tra le classi, cioè qual è il pattern limite che separa una classe da un'altra.
Un altro modo semplice per visualizzare patterns multidimensionali è il cosiddetto radar-plot. In questa rappresentazione si definiscono N assi quanti sono le variabili del pattern, gli assi vengono disposti a raggiera a formare un cerchio e lungo ogni asse viene graficato il valore della variabile corrispondente. Unendo i punti sugli assi si ottiene una figura che forma il "profilo" del pattern. Le figure geometriche così ottenute possono essere considerate in termini di area (cioè della intensità del pattern) e in termini di forma (cioè della composizione del pattern stesso).
Tale metodo ottimizza la visualizzazione grafica e per questo motivo è molto utilizzato in contesti dove tradizionalmente la cultura matematica non è diffusa, ad esempio viene utilizzato frequentemente in analisi sensoriale per definire i profili sensoriali degli alimenti.
Nella figura 5.5 vengono mostrati i radar plot delle acque minerali dell'esempio precedente.

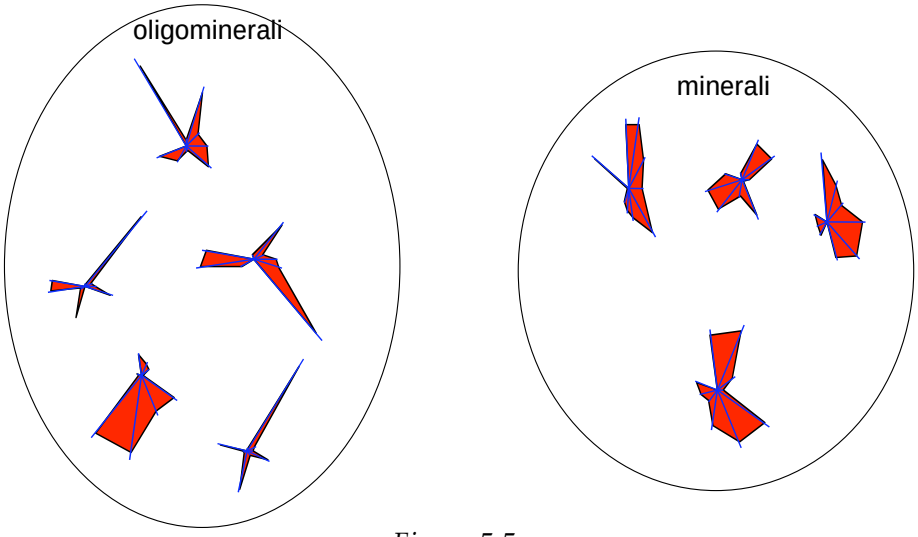


Figura 5.5

Come accennato precedentemente, per la pattern recognition abbiamo la necessità di definire un criterio univoco di distanza. Per questo motivo definito un insieme di pattern ciascuno con N features, definiamo uno spazio vettoriale euclideo di dimensione N dotato di una base orto-normale in cui ogni asse della base è relativo ad una delle features. Un insieme pattern, viene così a formare un insieme di punti nello spazio. Matematicamente un insieme di M pattern, ognuno descritto da N feature forma essere una matrice MxN.
La similitudine tra pattern è associata al concetto di distanza, per cui punto appartenenti alla stessa classe si troveranno più vicini rispetto a punti appartenenti a classi differenti.
I metodi di classificazione possono in termini generali essere suddivisi in due grandi categorie: i metodi *unsupervised* e *supervised*. I metodi unsupervised sono anche detti metodi esplorativi. In questi metodi non viene fornita alcuna ipotesi sulla appartenenza dei pattern alle classi, ma la classificazione viene determinata solamente osservando come i pattern si relazionano, in termini di distanza, tra loro, ed in particolare se i pattern tendono a formare degli ammassi (clusters) tra loro isolati. Il secondo metodo invece è simile alla regressione che abbiamo precedentemente studiato, in questo caso ogni pattern è associato ad una classe e scopo della pattern recognition è quello di trovare una trasformazione (funzione) che

associ il pattern alla classe. Vedremo come questo problema si può risolvere usando gli stessi metodi di regressione che abbiamo visto nel capitolo precedente.

5.2 Metodi unsupervised

I metodi unsupervised cercano di determinare se i pattern sono disposti nello spazio a formare degli ammassi isolati ai quali possono corrispondere classi significative. E' da dire innanzitutto che questo tipo di classificazione spontanea dei pattern può non essere interessante nel caso di scopi precisi. Nonostante ciò la conoscenza del tipo di classificazione spontanea che i pattern assumono è un punto di partenza importante per ogni problema di classificazione.
Il metodo più importante per la classificazione unsupervised è la cluster analysis. La cluster analysis considera il raggruppamento gerarchico dei dati. In funzione della distanza, i pattern vengono raggruppati a formare insiemi via via più grandi.
Lo strumento di base della cluster analysis è la matrice delle distanze. Dato un insieme di x_i pattern la matrice della distanza d_{ij} è così definita:

$$d_{ij} = \left\| X_i - X_j \right\|$$

dove l'operazione di norma è generalmente la norma euclidea. La matrice delle distanze ovviamente è una matrice simmetrica con diagonale nulla.
Consideriamo la figura 5.6 dove sono rappresentati dei pattern formati da due features, quindi lo spazio dei pattern è un piano.

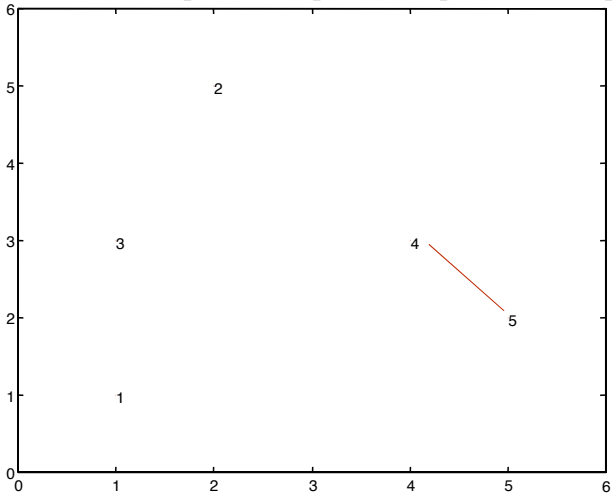


Figura 5.6

La matrice delle distanze dei cinque pattern è la seguente:

	1	2	3	4	5
1	0	4.1231	2.0000	3.6056	4.1231
2	4.1231	0	2.2361	2.8284	4.2426
3	2.0000	2.2361	0	3.0000	4.1231
4	3.6056	2.8284	3.0000	0	**1.4142**
5	4.1231	4.2426	4.1231	1.4142	0

Ovviamente i pattern 4 e 5 sono i più vicini e verosimilmente possiamo identificare due gruppi di dati 1 2 3 e 4 5.
La cluster analysis rende questi ragionamenti oggettivi e consente di operare con spazi N-dimensionali cioè con pattern di dimensioni qualsiasi. Il metodo è basato so di un processo gerarchico iterativo ed ha come risultato una struttura ad albero (dendrogramma) che definisce l'appartenenza alle classi in funzione della distanza.
Il passo i-esimo della analisi dei cluster consiste nelle seguenti operazioni:

- Calcolo matrice di distanza
- Individuazione della coppia di punti con distanza minore
- Formazione del cluster unendo i punti
- Sostituzione del cluster con il punto medio (o con uno dei punti della coppia)
- Reitarazione del procedimento finchè rimane un solo punto

Si definiscono differenti tipi di cluster analysis a seconda del criterio di raggruppamento e del conseguente calcolo della distanza tra cluster e tra pattern e cluster.
Senza entrare nel merito delle differenze tra vari tipi di cluster analysis mostriamo di seguito un esempio di applicazione al caso appena mostrato. La figura 5.7 mostra la sequenza delle operazioni.

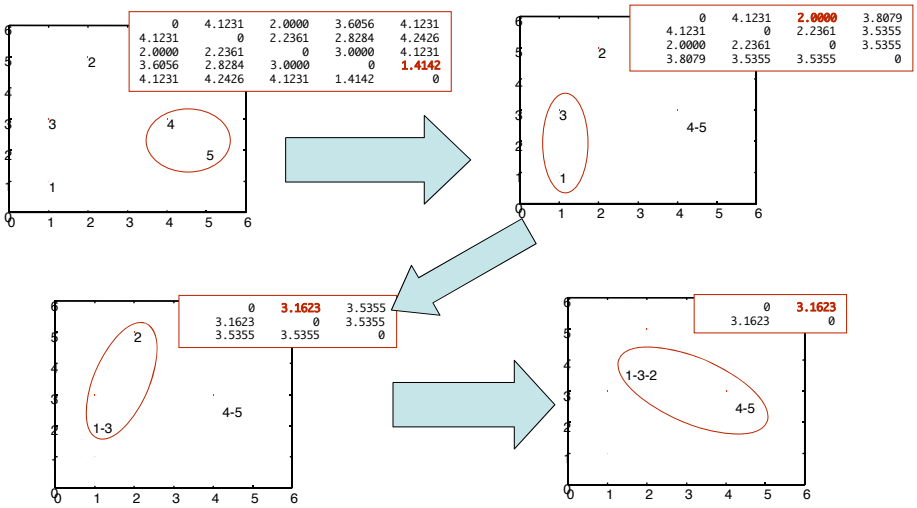


Figura 5.7

Il primo passo, in alto a sinistra, mostra i pattern originali e la matrice di distanza prima definita. La coppia 4-5 viene sostituita da un solo pattern, calcolato in questo caso come la media tra i due, ed una nuova matrice di distanze viene calcolata. Il valore minimo si ottiene per la coppia 1-3 che al secondo passo viene unita in un unico pattern. Nel terzo passo, il pattern 2 viene aggiunto al cluster 1-3 mentre nell'ultimo passo i due clusters 1-3-2 e 4-5 vengono uniti a formare il cluster di tutti i pattern. L'ultimo passo della analisi dei cluster coincide sempre con un unico cluster.
Il risultato della cluster analysis viene espresso attraverso un diagramma ad albero (dendrogramma) dove tutti i punti sono uniti, in funzione della distanza a formare dei cluster.
Il dendrogramma relativo all'esempio appena discusso è mostrato in figura 5.7.

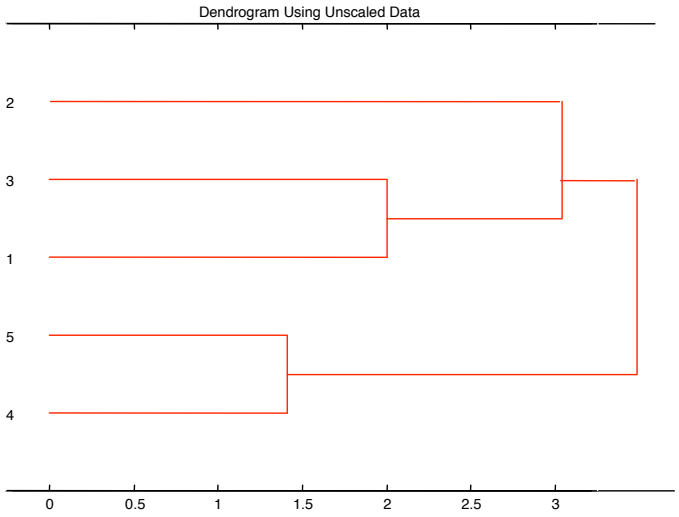


Figura 5.7

Il dendrogramma di figura 5.7 indica l'esistenza di due gruppi di dati: il primo che si forma alla distanza 1.5 ed il secondo alla distanza 3.0 Le differenti distanze di clustering indicano che un gruppo è più compatto dell'altro.
Applicata al. caso delle acque minerali, la analisi dei cluster fornisce il dendrogramma di figura 5.8.

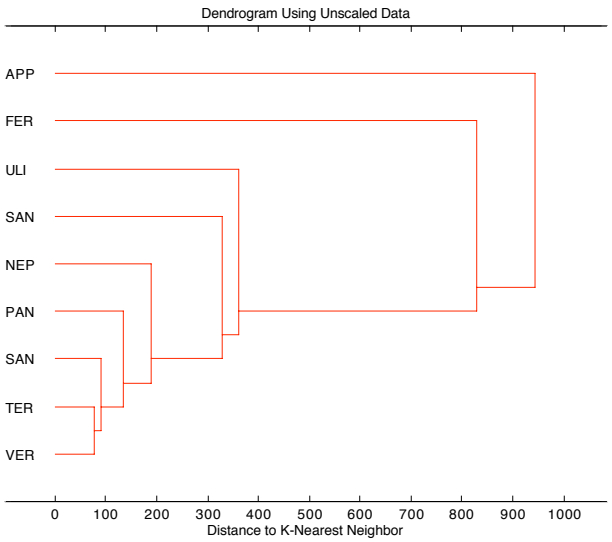


Figura 5.8

L'analisi dei cluster non fornisce una classificazione semplice come quella ipotizzata precentenemente delle acque oligominerali e minerali. Piuttosto, ad un gruppo compatto di acque (ver, ter, san) si aggiungono via via tutte le altre. Solo le marche app e fer si raggruppano ad una distanza maggiore lasciando supporre una sostanziale differenza dal resto del gruppo.

Metodo del Potenziale
Un interessante metodo "esotico" per la classificazione "unsupervised" è quel del potenziale basato sulla analogia con i campi di forze. Ad esempio consideriamo il campo di forze gravitazionali e supponiamo che ogni punto possegga una massa M (uguale per tutti). Tale massa genererà un potenziale V in ogni punto dello spazio. Poiché il potenziale è additivo, in ogni punto dello spazio si avrà un valore del potenziale pari alla somma dei potenziali generati da tutti i punti.
Dove i punti sono raggruppati (dove cioè esiste una classe) si genererà un potenziale maggiore. Studiando l'andamento del potenziale, ed in particolare i suoi massimi è possibile individuare delle regioni spaziali di massimo addensamento (classi).

L'analogia con le masse è solo un esempio si può utilizzare qualunque funzione potenziale.
Nel caso dell'analogia gravitazionale, dati N punti x_i il potenziale in un punto r è dato da:

$$V(\vec{r}) = \sum_i \frac{1}{\|\vec{x}_i - \vec{r}\|}$$

Supponiamo di avere un certo numero di pattern descritti da due features e rappresentati nel piano di figura 5.9.

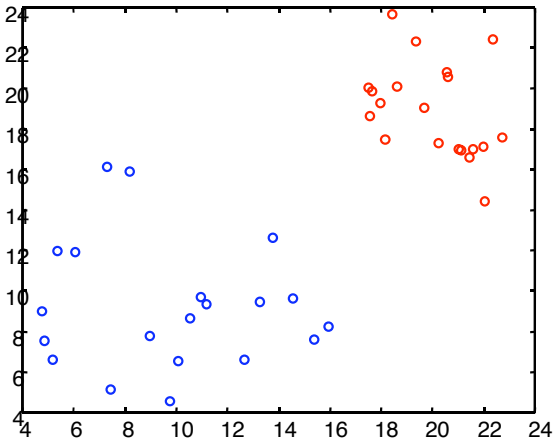


Figura 5.9

Il potenziale di tale distribuzione è rappresentabile sia come rappresentazione tridimensionale (figura 5.10) che come curve di livello (figura 5.11).

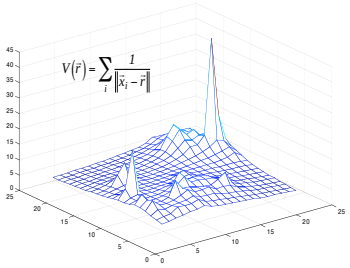


Figura 5.10

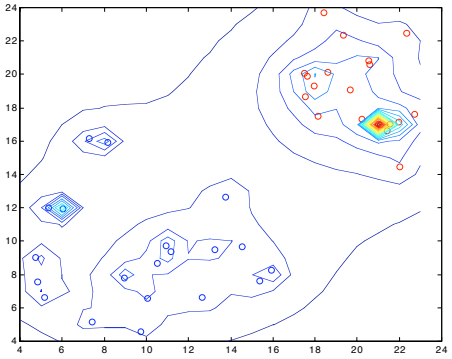


Figura 5.11

A seconda del livello considerato si possono quindi definire delle classi di appartenenza per i pattern.
E' importante notare che in entrambe i metodi, cluster analysis e potenziale, la assegnazione del pattern alla classe non è univoca ma dipende da un parametro, la distanza o il livello di potenziale, lasciando all'analizzatore la facoltà di stabilire qual è il livello del parametro affidabile per la classificazione stessa.

5.3 Il problema della normalizzazione dei dati

Uno degli aspetti più importanti della pattern recognition è quello della normalizzazione dei dati. La normalizzazione è un procedimento indispensabile per eliminare le differenze numeriche tra variabili e le differenze tra le unità di misura.
La prima motivazione è quella fondamentale, infatti se ad ogni feature si associa un asse è chiaro che se una feature è caratterizzata da valori numerici molto più grandi rispetto l'informazione portata da questa feature sarà dominante. Mentre uno dei criteri importanti per la scelta delle features è l'uguaglianza della informazione. Oltre al valore numerico assoluto è importante notare come anche il range, cioè l'escursione numerica della feature, sia importante.
Le normalizzazioni più importanti sono due: la media nulla e l'autoscaling.
Nella media nulla ogni feature viene scalata attorno al valore di 0, in modo che la media della feature in un insieme di pattern sia nulla
L'autoscaling consiste nel rendere pari a 0 la media e pari a uno la varianza di ogni feature.
Le espressioni della matrice dei pattern e delle corrispondenti matrici a media nulla ed autoscalate sono:

$$X$$

$$A_{ij} = X_{ij} - m_j$$

$$Z_{ij} = \frac{X_{ij} - m_j}{\sigma_j}$$

Da un punto di vista grafico l'applicazione della normalizzazione produce gli effetti mostrati in figura 5.12.

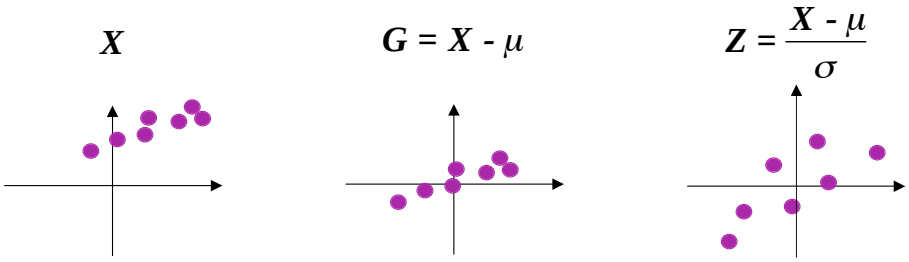


figura 5.12

L'autoscaling in particolare consiste nel dare lo stesso peso ad ogni features, la cosa è positiva se si è certi che ogni features, indipendentemente dal suo valore numerico, ha la stessa importanza nel problema.
L'autoscaling diventa pericoloso quando una o più features sono rumorose oppure quando le relazioni numeriche tra features sono importanti. Caso tipico è quello degli spettri dove l'autoscaling distrugge completamente l'informazione.
La normalizzazione modifica il risultato della pattern recognition e quindi è estremamente importante valutare se e quando applicarla. Consideriamo ancora il caso delle acque minerali, la cluster analysis valutata sui dati autoscalati produce il dendrogramma di figura 5.13

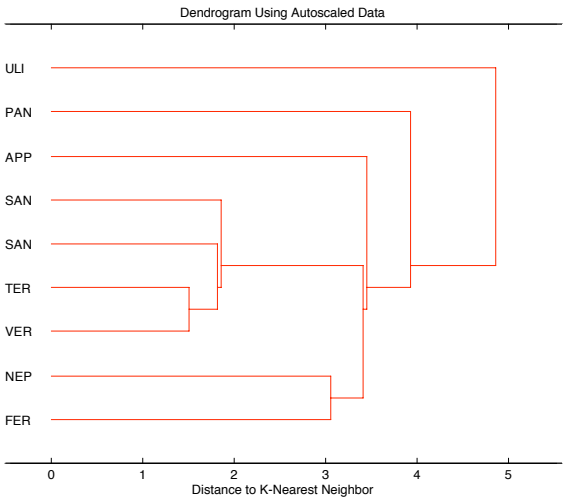


figura 5.13

In questo caso appaiono evidenti due classi di acque (san-san-ter-ver e nep-fer) mentre altre tre tipi (uli, pan e app) appaiono differenti sia tra loro sia rispetto ai due gia citati gruppi. Confrontando questo risultato con il precedente osserviamo situazioni e comportamenti completamente differenti e che il risultato ottenuto con i dati normalizzati da una descrizione delle acque sicuramente più prossima a quanto atteso.

5.4 Gli Spazi di Rappresentazione: Analisi delle Componenti Principali (PCA)

Negli esempi fatti finora abbiamo sempre mostrato pattern con due features, ovviamente il caso più frequente è quello relativo a pattern multidimensionali rappresentati da un numero qualsiasi di features. In questi casi si pone sia il problema della rappresentazione dei pattern sia il problema della individuazione del numero effettivo di dimensioni del problema.
Abbiamo già sfrontato questo aspetto parlando della regressione, ovvero quando abbiamo un problema multidimensionale spesso accade che la dimensione dello spazio realmente occupato dai pattern sia minore del numero delle features. Questo è dovuto a ciò che abbiamo chiamato colinearità; ovvero al fatto che Le features utilizzate per descrivere un fenomeno non sono assolutamente scorrelate tra loro.
Ad esempio nella classificazione degli individui il peso e l'altezza hanno sicuramente un certo grado di correlazione tra loro.
Come abbiamo accennato, la correlazione tra elementi del vettore delle features fa si che lo spazio delle features non sia "completamente" occupato ma che i pattern tendono a giacere in sottospazi.
E' quindi opportuno studiare dei metodi di riduzione delle variabili che consentono di individuare le variabili più significative, di ridurre le dimensioni del problema e di rappresentare pattern multidimensionali in spazi visualizzabili (2 o 3 D).
Tra i vari metodi che consentono di ottenere questi risultati uno dei più utilizzati è la analisi delle componenti principali (PCA). Abbiamo studiato i benefici della PCA per quanto riguarda la regressione, nel caso della pattern recognition la PCA 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.

Studiamo la applicazione della PCA alla pattern recognition unsupervised attraverso quattro tipi di dati.

Classificazione di frutti descritti da parametri chimici
Consideriamo un insieme di pesche, per ogni pesca di diverso cultivar, abbiamo definito un pattern costituito dalle seguenti features: pH, saccarosio, glucosio, fruttosio, acido malico ed acido citrico. In pratica si vuole studiare la distribuzione della popolazine di pesche in funzione della loro composizione acida e zuccherina.
La prossima figura mostra i valori dei parametri per tutte le pesche attraverso una rappresentazione di colore. La matrice dei dati viene graficata assegnato un colore da un massimo ad un minimo ad ogni elemento della matrice. Questo modo di rappresentare i dati consente di visualizzare le variabilità delle features e le loro intensità. In figura 5.14 sono mostrati sia i dati reali che i dati autoscalati.

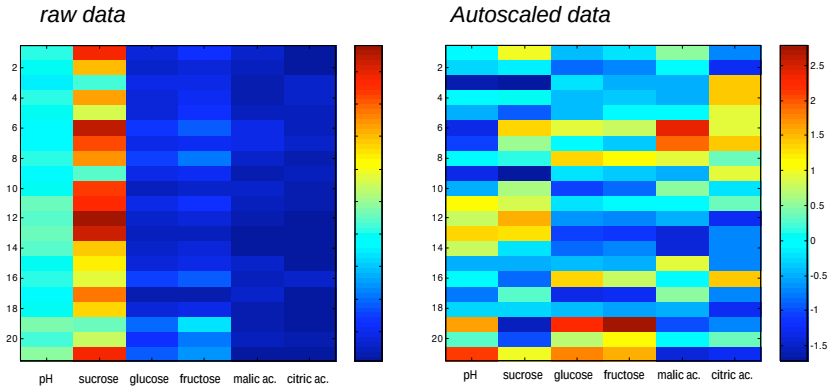


figura 5.14

I dati grezzi (graficati a sinistra) sono dominati dal saccarosio, mentre il pH ha una scarsa variabilità. I dati autoscalati (a destra) mostrano invece una omogeneità di informazione di tutte le feature. Usando i dati autoscalati siamo quindi in condizione di utilizzare l'informazione di tutte le feature.

La figura 5.15 mostra gli autovalori associati con le componenti principali. Come si ricorderà l'autovalore è proporzionale alla varianza portata dalla componente principale. In pratica possiamo ridurre la rappresentazione alle componenti principali i cui autovalori siano significativamente diversi da 0.

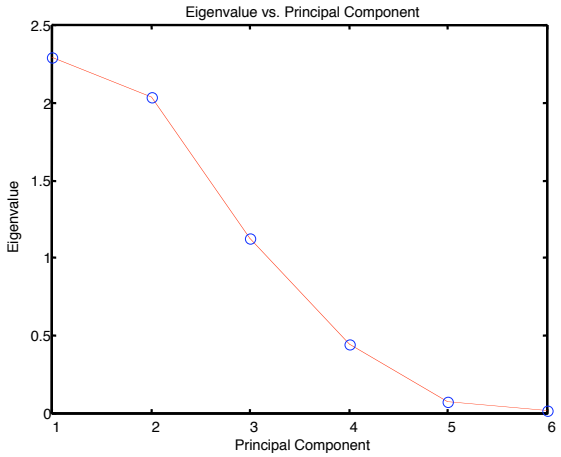


Figura 5.15

In questo caso solo l'ultima componente principale può essere veramente trascurata, si può quindi dire che per questo set di dati la colinearità delle feature è contenuta.
La proiezione dei dati autoscalati sul piano individuato dalle prime due componenti principali fornisce lo score plotdi figura 5.16 .

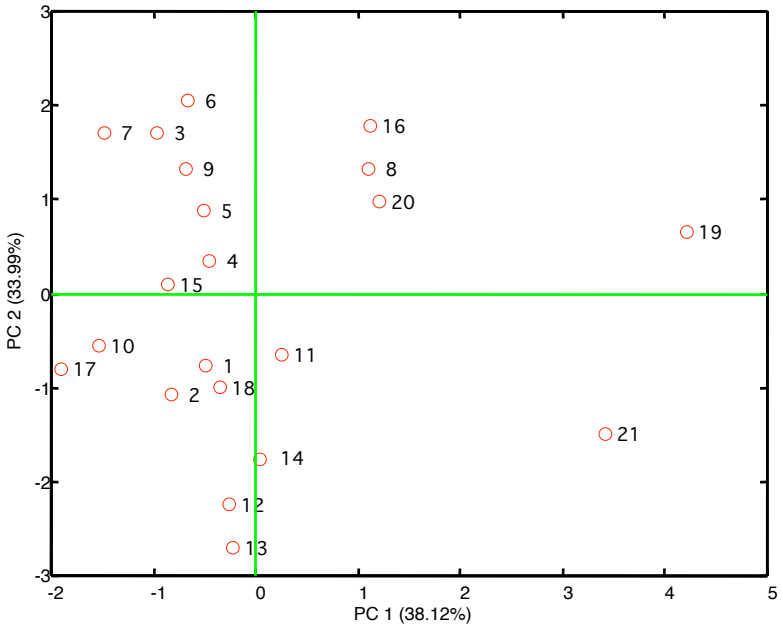


Figura 5.16

La quantità di informazione rappresentata nelle prime due componenti principali è pari alla somma della percentuale degli autovalori corrispondenti. In questo caso 38.12+33.39=71.51%.
La rappresentazione mostra alcune caratteristiche interessanti dei dati. Ad esempio, l'esistenza di un gruppo di 3 specie (16, 8 , 20) ed il comportamento peculiare del 19 e del 21 caratterizzati da un valore elevato della prima componente principale. Prima di investigare il senso di questa distribuzione è però opportuno chiedersi se nella parte di informazione non rappresentata in questa figura (e pari a circa al 30%) siano contenute informazioni interessanti.
Per questo scopo è opportuno valutare l'entità del residuo per ogni dato rappresentato.
L'analisi delle componenti principali infatti non è altro che lo sviluppo di una matrice X nel prodotto delle due matrici di score (S) e loading (P).

$$X_{nm} = S_{nq} \cdot P_{qm}^{T}$$

le dimensioni originali della matrice X si ottengono indipendentemente dalla dimensione interna (q) che stabilisce

l'ordine di sviluppo della PCA. Nella figura precedente, poiché la rappresentazione coincide con un piano si è considerato q=2.
Questo vuol dire che il pattern originario formato dai valori delle features è trasformato in un pattern formato dai valori delle componenti principali e dove solo i primi due termini sono considerati importanti e gli altri termini costituiscono il residuo della rappresentazione.

$$x_i = \left\{ a \cdot s_1 \, , b \cdot s_2 \, , \ldots \, , n \cdot s_n \right\}$$

$$x_i^{pca} = \left\{ \bar{a} \cdot \, pc_1 \, , \bar{b} \cdot \, pc_2 \, , residual \right\}$$

il valore del residuo (Q) si può calcolare considerando il fatto che il modulo di un vettore non dipende dal sistema di riferimento per cui il modulo di x_i coincide con il modulo di x^{pxa}.

$$\left| x_i \right| = \left| x_i^{pca} \right| \Rightarrow$$

$$\Rightarrow \sqrt{a^2 \cdot s_1^2 + b^2 \cdot s_2^2 + \ldots + n^2 \cdot s_n^2} = \sqrt{\bar{a}^2 \cdot pc_1^2 + \bar{b}^2 \cdot pc_2^2} + Q$$

$$Q = \sqrt{a^2 \cdot s_1^2 + b^2 \cdot s_2^2 + \ldots + n^2 \cdot s_n^2} - \sqrt{\bar{a}^2 \cdot pc_1^2 + \bar{b}^2 \cdot pc_2^2}$$

Il residuo può essere graficato aggiungendo un terzo asse allo score plot come raffigurato in figura 5.17.

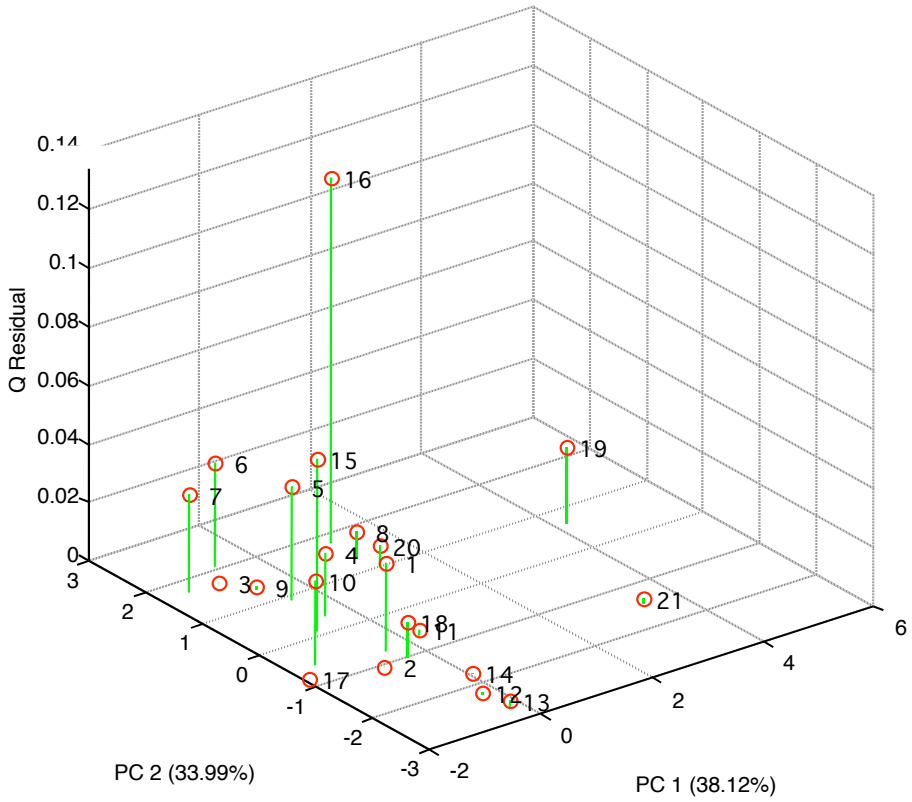


Figura 5.17

In figura 5.17 si osserva che il punto 16, che nello score plot forma un cluster con i punti 8 e 20, è caratterizzato da un grande residuo. Quindi la rappresentazione dello score plot può non essere affidabile per il punto 16 le considerazioni che faremo sullo score plot devono quindi essere considerate con cautela per questo dato.
Questa considerazione ci mostra come nonostante sia possibile quantificare la affidabilità di una rappresentazione attraverso la varianza (informazione) totale dello score plot, ogni singolo dato ha un proprio grado di affidabilità. In particolare, è necessario verificare la omogeneità del residuo per tutti i dati.
Tornano allo sviluppo alla seconda componente principale è importante valutare quale è il ruolo delle features (cioè delle variabili naturali del problema) nello score plot. Per questo è disponbile uno

strumento che rende semplice questo tipo di analisi: il biplot. Nel biplot vengono proiettati sul piano delle componenti principali sia i dati sia i versori degli assi originali. In questo modo è possibile studiare gli allineamenti tra dati e features e il comportamento reciproco tra le features stesse.
Ricordando che i versori originali sono ortogonali tra loro, quanto più questa ortogonalità è mantenuta nello score plot tanto più le features sono scorrelate tra loro e forniscono informazioni indipendenti nella rappresentazione dello score plot. Al contrario i versori che nello score tendono a diventare colineari identificano delle features ridondanti, nel senso che contribuiscono con la stessa informazione allo score plot.
Osserviamo in figura 5.18 il biplot relativo allo score plot precedente.

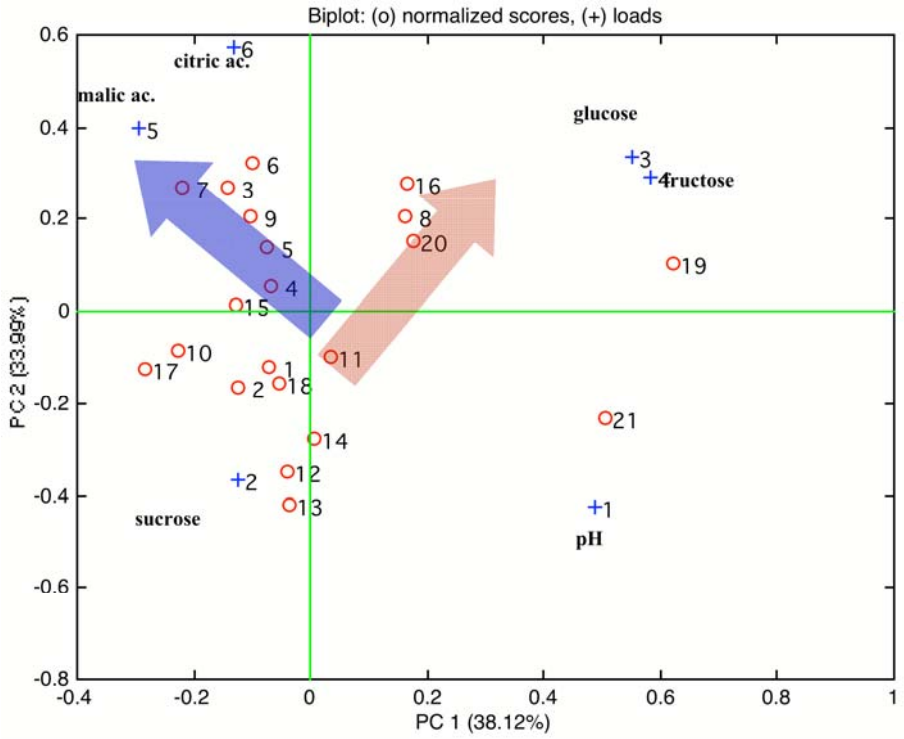


Figura 5.18

i loadings (cioè le proiezioni dei versori delle features) sono graficati con il simbolo (+). Osserviamo come la direzione degli zuccheri (fruttosio, glucosio e sucrosio) sia ortogonale alla direzione degli acidi (citrico e malico). Mentre all'interno di ogni gruppo di features, I loadings sono colineari. Si osservi che giustamente il loading del pH è opposto alla concentrazione degli acidi.
Grazie alla disposizione dei loadings, possiamo ad esempio affermare che le pesche relative ai pattern 8, 16, 20 siano pesche dolci mentre le pesche 3, 5, 6, 7 e 9 siano frutti acidi.
L'analisi combinata di score e loading consente quindi di studiare la classificazione dei singoli pattern, il ruolo delle features, il loro grado di correlazione, e di interpretare i pattern.

Classificazione delle Acque Minerali
Nel secondo esempio riconsideriamo l'esempio delle acque minerali che ha accompagnato la discussione dei metodi di pattern recognition.
Abbiamo già osservato come l'applicazione della normalizzazione modifichi in maniera sostanziale il risultato della cluster analysis e quindi la classificazione dei pattern. In questo esempio studiamo l'effetto della normalizzazione nella analisi delle componenti principali.
Innanzitutto consideriamo gli autovalori della PCA relativi ai dati non normalizzati, ai dati centrati ed ai dati autoscalati. Nelle figura 5.19 5.20 e 5.21 sono mostrati gli andamenti degli autovalori in funzione delle componenti principali per i dati originali, centrati ed autoscalati.

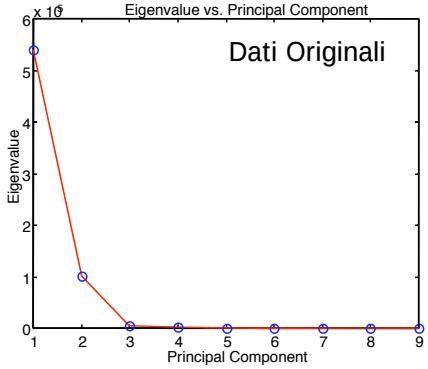


Figura 5.19

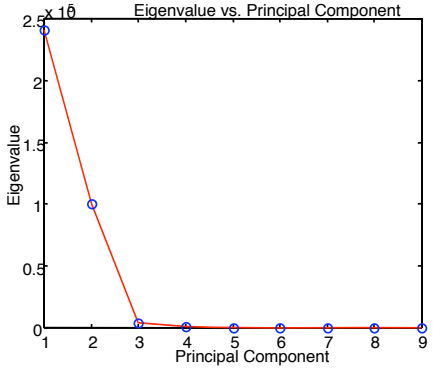


Figura 5.20

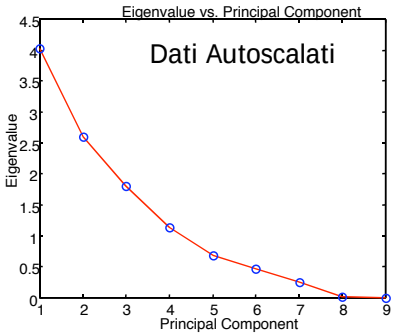


Figura 5.21

L'andamento degli autovalori per i dati centrati è ovviamente lo stesso dei dati non normalizzati. In entrambe i casi i primi due autovalori sono sufficienti a rappresentare tutta l'informazione. In questo caso ci sono due features che numericamente sovrastano le altre.
L'autoscaling invece rende le features omogenee e quindi aumenta il numero delle direzioni significative. Si osservi che il primo autovalore trascurabile in figura 5.21 è l'ottavo.
In termini di biplot questo significa considerare un numero maggiore di informazioni nella rappresentazione. La figura 5.22 mostra il biplot relativo alle prime due componenti principali della PCA calcolata sui dati centrati.

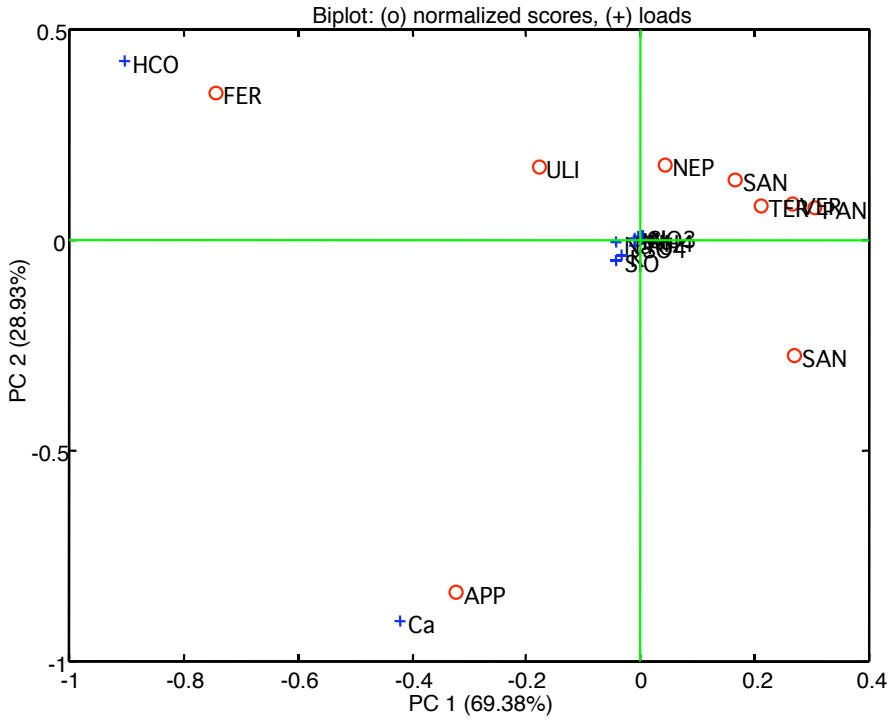


figura 5.22

Si osservi innanzitutto che circa il 98% della varianza dei dati è mostrata nello score plot. Inoltre solo le variabili numericamente più grandi (HCO e Ca) sono quelle che determinano la rappresentazione, infatti eccetto queste i loadings sono tutti graficati attorno allo zero. Si osservi inoltre che i loadings di Ca ed HCO sono circa ortogonali tra loro a significare che la rappresentazione è una proiezione sul piano identificato dai versori di queste variabili. Nella rappresentazione solo FER e APP emergono dal gruppo delle acque che mostra una certa omogeneità.
La situazione cambia se consideriamo la PCA dei dati autoscalati (figura 5.23)..

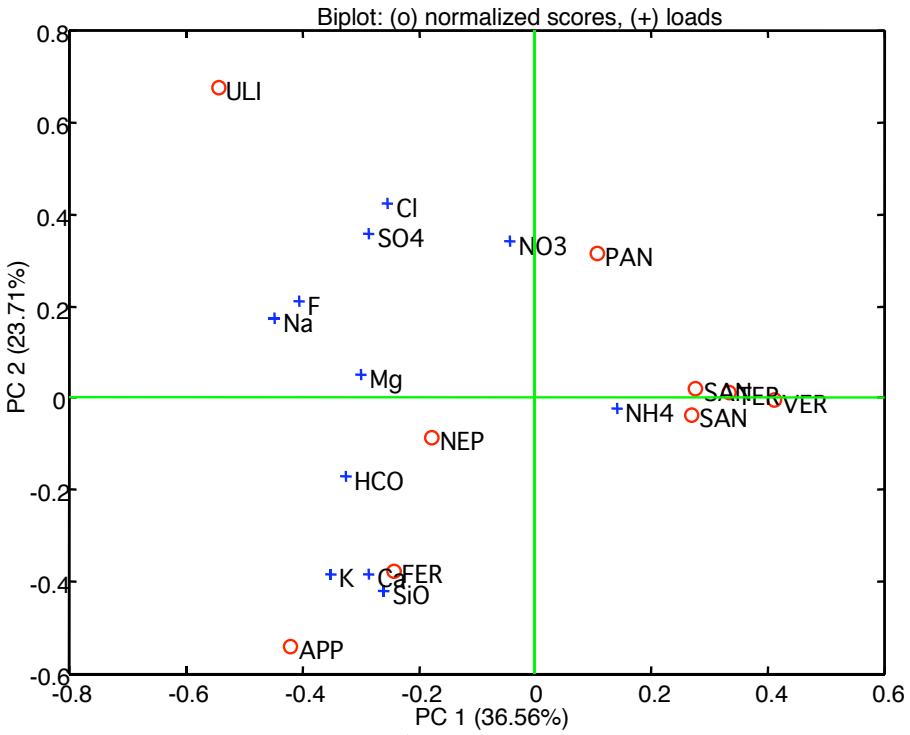


Figura 5.23

Le prime due componenti principali portano circa il 60% dell'informazione, quindi la rappresentazione è sicuramente parziale. Inoltre i loadings sono distribuiti in tutto il piano a significare che il piano delle componenti principali non è orientato verso nessuna delle variabili originali. Gli score identificano veri gruppi di acque, in particolare emerge il gruppo SAN, VER, TER, SAB, e PAN delle acque oligominerali (si osserva come i loadings sia prevalentemente orientati nel semipiano opposto). Inoltre c'è una sequenza NEP, FER, APP orientati lungo HCO, SiO, e K mentre l'acqua ULI giace in una direzione ortogonale a questi ad indicare una composizione completamente peculiare e caratterizzata da valori elevati di Cl, SO4, F, e Na.
La rappresentazione bidimensionale è comunque limitata dal fatto che circa il 40% dell'informazione non è considerato. Si può ampliare l'informazione dello score plot aggiungendo, come in figura 5.24, una terza componente principale.

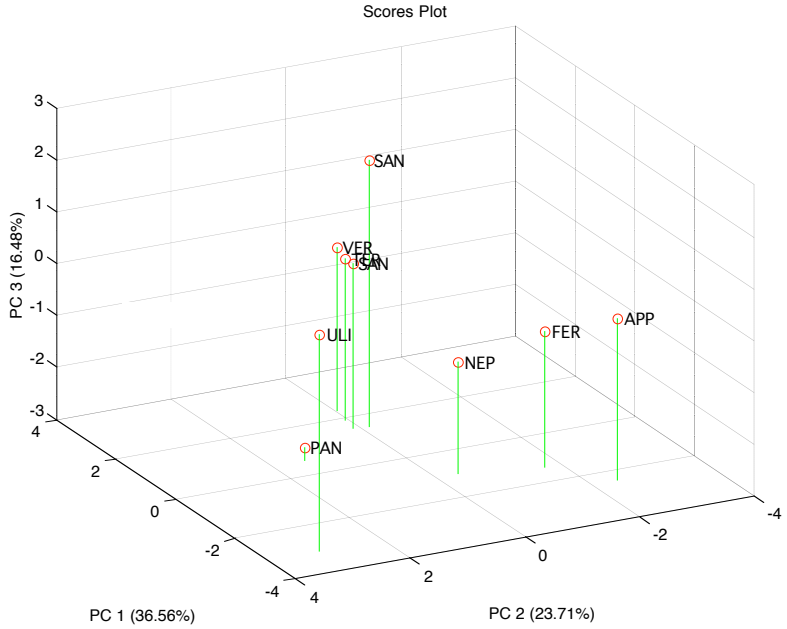


figura 5.24

In questo modo il contenuto dell'informazione sale al 76%. La dimensione aggiunta comunque non modifica di molto la distribuzione dei dati tranne per il fatto che emerge il carattere peculiare dell'acqua SAN.

Matrice di sensori di gas non selettivi
Consideriamo in questo esempio una matrice di quattro sensori di gas esposti a miscele binarie di n-ottano e toluene. Le miscele considerate nello spazio delle variabili sono equispaziate come mostrato in figura 5.25.

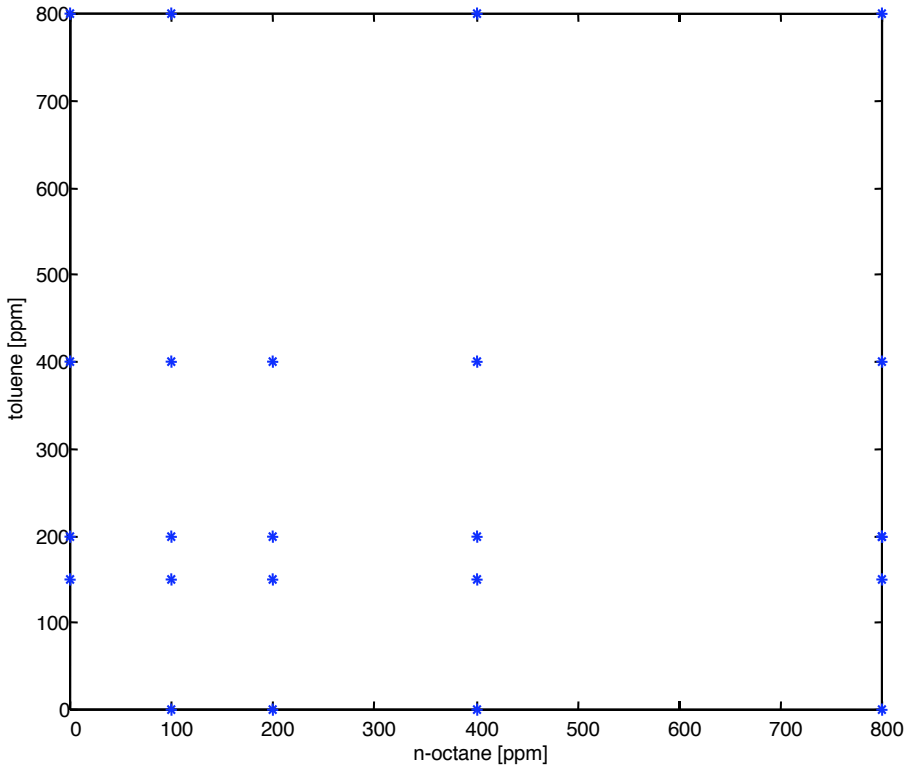


Figura 5.25

Come abbiamo già visto nel capitolo relativo ai sensor array, un sistema di sensori lineari in cui cioè la risposta è combinazione lineare delle concentrazioni dei gas della miscela trasforma una distribuzione quadrata di punti dello spazio delle variabili in una parallelogramma la cui apertura è direttamente proporzionale al grado di correlazione tra i sensori.
Consideriamo in figura 5.26 le prime due componenti principali della matrice dei sensori.

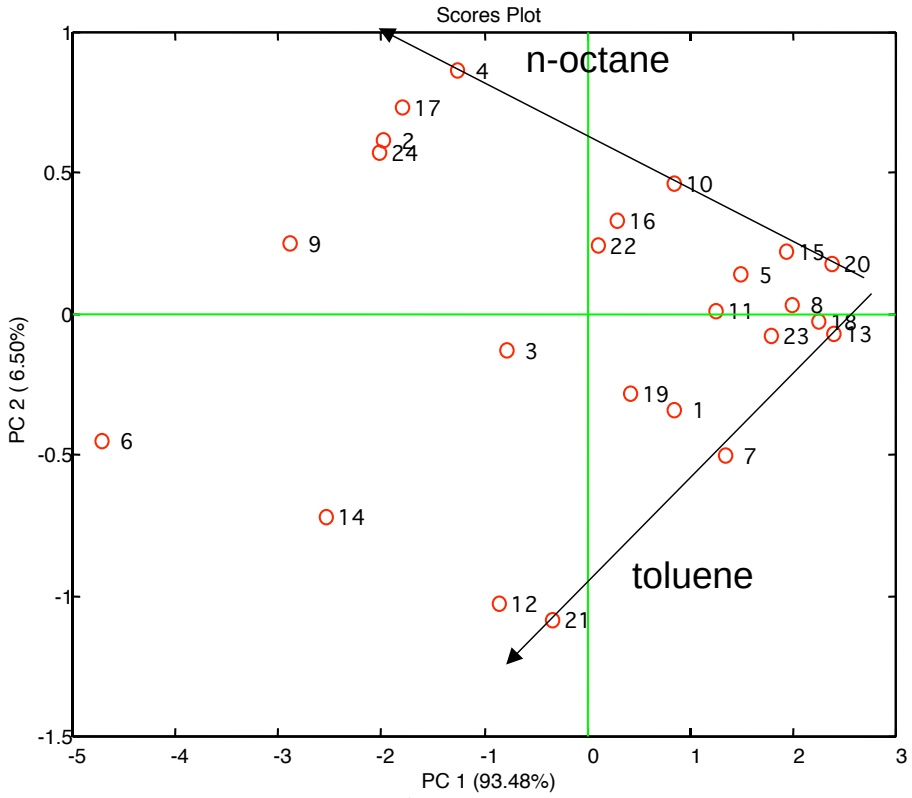


Figura 5.26

Il piano contiene circa il 99% della varianza ad indicare una grande correlazione tra i sensori. Osserviamo come in questo piano ritroviamo il parallelogramma dell'esempio bidimensionale discusso nel capitolo relativo agli array di sensori.
E' opportuno considerare il fatto che la correlazione tra i sensori è dovuta al fatto che la risposta sensoriale dipende sia dalla qualità (tipo di gas) che dalla quantità (concentrazione) e che poiché in questo esempio la qualità è fissata (due gas) ma la quantità è variabile ne nasce una elevata correlazione delle risposte sensoriali.
Questo esempio ci fornisce la possibilità di introdurre un tipo di normalizzazione dei dati che consente di eliminare la dipendenza dalla quantità nel caso di gas puri e di ridurla nel caso di miscele e di sensori non perfettamente lineari.

Considerando infatti che la risposta di ogni singolo sensore ad un gas è data dal prodotto della concentrazione del gas moltiplicata la sensibilità del sensore a quel gas, se si divide la riposta del sensore per la somma delle risposte di tutti i sensori si ottiene che la risposta normalizzata è indipendente rispetto alla concentrazione.

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

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

Se applichiamo la PCA ai dati normalizzati secondo la formula precedente otteniamo lo score plot dell prime due componenti principali di figura 5.27.

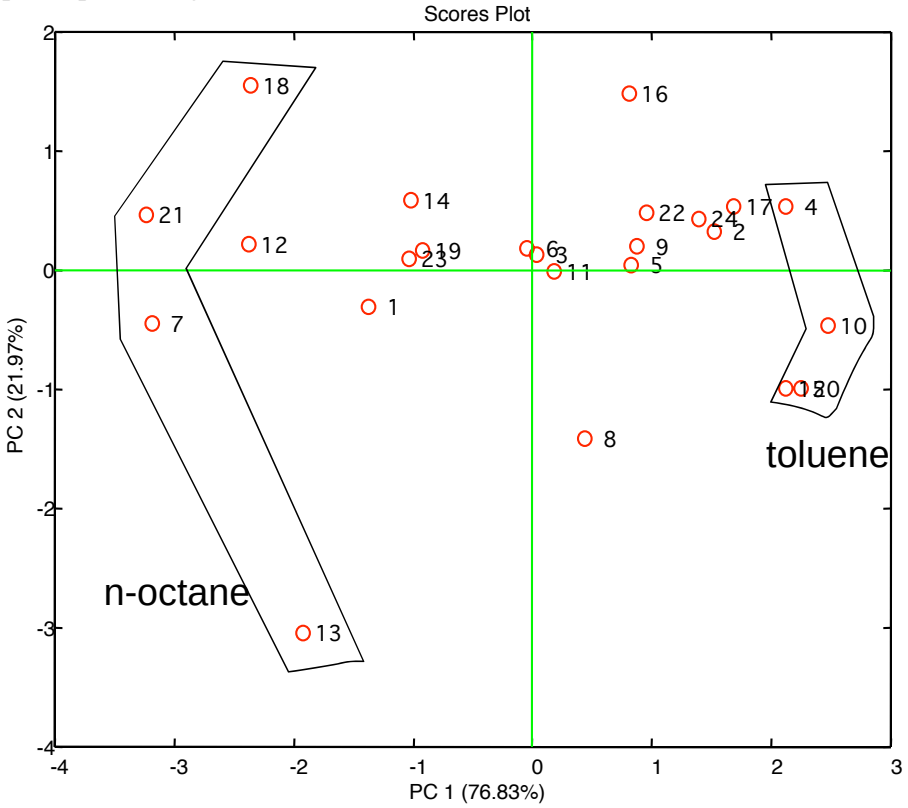


Figura 5.27

La varianza dello score plot è calata di poco (97%) ma stavolta è distribuita nelle due direzioni principali. Inoltre il parallelogramma non esiste più ma i dati vanno dall'n-ottano puro al toluene puro passando per i vari titoli delle miscele. Il contenuto di informazione quantitativa è quasi scomparso lasciando prevalentemente una descrizione di tipo qualitativo dei dati.

Analisi del welfare nelle regioni italiane

Concludiamo questi esempi di pattern recognition con la analisi delle componenti principali discutendo un problema econometrico cioè lo studio del welfare nelle regioni italiane. Mostriamo cioè come la PCA consente di studiare qualunque fenomeno multivariato indipendentemente dalla sua natura.

Consideriamo ogni regione italiana caratterizzata da un pattern di 21 feature relativi ad informazioni sulla geografia, la popolazione, i consumi, l'occupazione, etc.

Lo score plot delle prime due componenti principali è mostrato in figura 5.28.

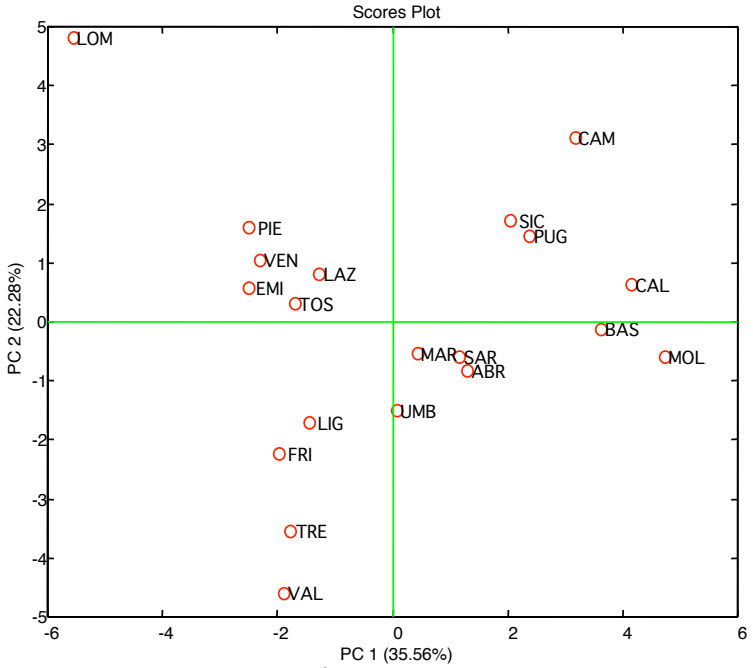


Figura 5.28

Il plot rappresenta solo il 58% dell'informazione globale e mostra il carattere peculiare della Lombardia ed alcuni raggruppamenti interessanti come quello tra Piemointe, Veneto, Emilia, Lazio e Toscana che identifica regioni con un tenore di vita simile. Inoltre sono in evidenza le regioni a minore sviluppo (Campania, Sicilia, Puglia, Molise, Calabria e Basilicata). Le altre regioni italiane seguono invece un diverso orientamento ad indicare un terzo tipo di modello sociale. Queste sono le regioni più piccole caratterizzate da un economia basata sulla piccola impresa.
Il biplot di figura 5.29 mette in evidenza le peculiarità di questi tre gruppi.

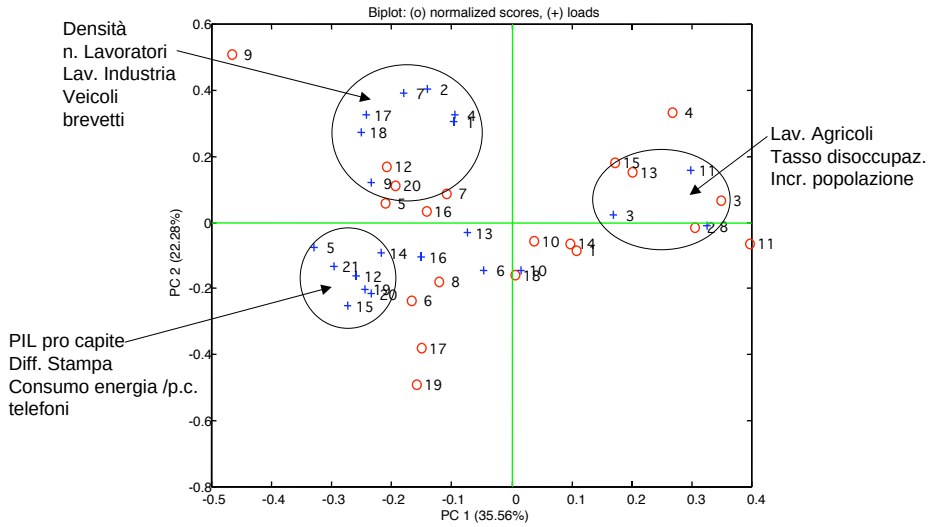


figura 5.29

Il gruppo che ha comprende le regioni grandi e che converge verso la Lombardia è caratterizzato da alta densità, da elevato numero di lavoratori nell'industria, da elevato numero di occupati, dal numero dei veicoli e dei brevetti. Le regioni a minore sviluppo sono invece in direzione opposta rispetto agli indicatori sopra citati (e quindi hanno basso valori di queste variabili) ed invece valori elevati di lavoratori agricoli, tasso di disoccupazione e di incremento della popolazione. Infine ortogonale rispetto a questa direzione, che identifica il percorso verso una società industriale, c'è il gruppo delle piccole

regioni che si caratterizza per un elevato PIL pro-capite, diffusione stampa, consumi di energia pro capite e di telefoni. Quindi per una ricchezza ed una qualità della vita maggiori.

Analisi degli outliers di una distribuzione multivariata: esempio sorgenti inquinate

Una delle applicazioni più frequenti nel controllo di processo è quella mirata alla ricerca dei campioni anomali. Si pensi ad esempio ad una linea di produzione nella quale si produce uno stesso campione, in questo caso è importante avere a disposizione un metodo che consenta di valutare se la differenze tra un singolo prodotto ed il prodotto "ideale" rientrano nella natura variabilità propria del processo di produzione oppure se il prodotto in questione è anomalo e quindi deve essere scartato. Un dato la cui probabilità è molto piccola è detto outlier. In pratica si ragiona dicendo che se la probabilità di osservare quel dato è molto piccola allora il fatto che il dato sia stato osservato è probabilmente dovuto al fatto che il dato non appartiene alla distribuzione di probabilità. Cioè che il fenomeno che ha generato quel dato non sia descritto dalla distribuzione di probabilità calcolata con l'intero insieme dei dati. In altre parole, il dato non è conforme con il restante insieme, un dato di questo genere si chiama outlier.

Abbiamo già affrontato questo problema parlando delle applicazioni della distribuzione normale univariata. Lo stesso esempio può essere esteso al caso multivariato. In particolare, la PCA ci consente di visualizzare le superfici di isoprobabilità proiettate sul piano delle comoponenti principali consentendo così il confronto tra i singoli campioni e la distribuzione normale che rappresenta l'intero insieme di dati. Le superfici isoprobabili per una distribuzione multivariata sono iper-ellissoidi che proiettati sul piano delle componenti principali diventano ellissi.

Consideriamo un esempio relativo alla composizione chimica di un certo numero di sorgenti d'acqua. Supponiamo di avere misurato per ogni sorgente le seguenti quantità: concentrazioni di U, As, B, Ba, Mo, Se, V, Solfati, alcalinità totale, bicarbonati, ed i valori di conducibilità e pH.

Il calcolo della PCA si basa sulla ipotesi fondamentale che che l'intero set di dati sia distribuito normalmente, quindi le superfici di iso-probabilità sono ellissi che nella base delle componenti principali sono rappresentate in forma canonica.

Sullo score plot possono quindi essere rappresentate le curva di isoproabilità. Per ricercare i dati che sono probabili outlier della distribuzione si confronta la posizione dello score con l'ellisse che racchiude il 95% di probabilità. I dati che giacciono all'esterno sono appunto gli outliers.
La figura 5.30 mostra lo score plot dei dai patterns relativi alle sorgenti. Nello score plot è evidenziata la ellisse del 95% di probabilità.

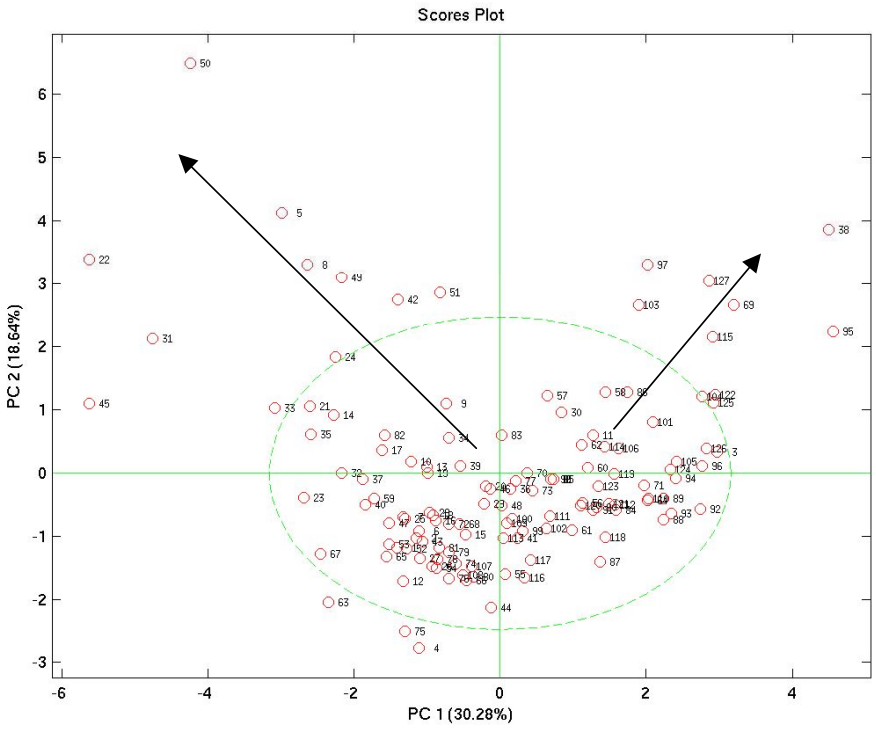


Figura 5.30

Nello score plot sono evidenziate due direzioni lungo le quali troviamo gli outliers. Osservando in un biplot quali variabili giacciono lungo le direzioni di deviazione dalla popolazione normale si può giungere alla identificazione della tipologia di anomalia.

Concludiamo la discussione sulla PCA mettendo in evidenza i limiti della analisi. In particolare si deve tenere presente che la rappresentazione prodotta dalla PCA è guidata dalle caratteristiche della matrice di covarianza dei dati e dalla ipotesi che i dati siano, nel senso multivariato, distribuiti nel loro complesso normalmente.
Quindi, 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. Nel caso della pattern recognition questa eventualità è tutt'altro che remota essendo presenti dati provenienti da diverse classi e quindi da diverse distribuzioni. Si ricordi che il teorema del limite centrale assicura che solo un numero elevato di distribuzioni qualunque forma una distribuzione normale.
Inoltre si deve sempre tenere presente che lo score plot della PCA è una proiezione lineare da uno spazio a dimensione N ad uno spazio a dimensione 2 o 3 e che quindi sono possibili effetti di falsa proiezione che comportano errori di classificazione. Si pensi ad esempi alle costellazioni celesti che mettono in relazione stelle lontanissime tra loro ma le cui proiezioni sulla volta celeste sono vicine.

5.5 Metodi Supervised

Finora abbiamo analizzato i metodi unsupervised. Lo scopo della analisi è stata quindi quella di fornire una rappresentazione adeguata dello spazio dei pattern in cui sia conservato il più possibile il carattere statistico dell'insieme dei dati. In particolare, la PCA è il metodo con cui, assumendo che i dati siano distribuiti normalmente, possiamo rappresentare pattern multidimensionali in spazi di dimensione 2 o 3. Abbiamo anche visto come, identificando la similitudine tra patterns con la distanza nello spazio dei pattern, sia possibile determinare come dato un insieme di dati questi si raggruppino in classi.
Lo scopo della pattern recognition è però un poco più ampio, ed è quello di attribuire un pattern ad una classe "pre-definita"; cioè ad una categoria pre esistente. Questa classificazione può in pratica essere differente rispetto a quella definita in via unsupervised dai metodi sopra discussi.
Consideriamo ad esempio di aver selezionato una serie di dati appartenenti a due classi (per chiarezza del disegno supponiamo in figura 5.31 uno spazio dei pattern bidimensionale).

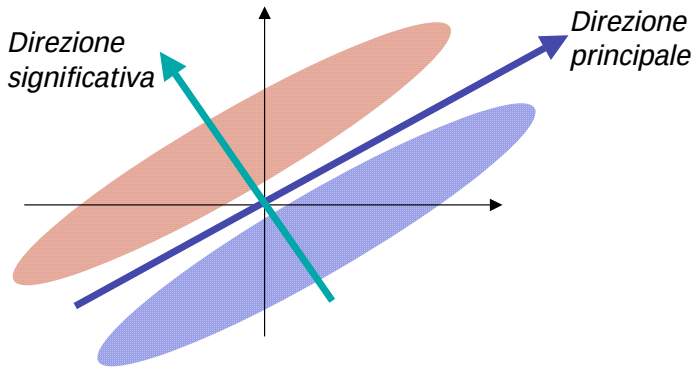


Figura 5.31

E' evidente come in questo caso sia la analisi dei clusters sia la PCA non ano non in grado di distinguere tra queste due classi. Per la cluster analysis il criterio della distanza non è in grado di assegnare due punti giacenti distanti alla stessa classe e due punti prossimi ma di classi diverse a due classi.
La PCA invece considera le direzioni di massima correlazione come direzioni importanti e nella figura precedente la direzione di massima correlazione è quella lungo la quale le due classi si mescolano.
Le direzioni principali non sono quindi in grado di descrivere opportunamente le due classi. Per far questo occorre, come nel caso della PLS discussa nel capitolo 4, un criterio che consenta di ruotare le componenti principali fino ad ottimizzare la soluzione del problema. La rappresentazione ottimale delle due classi è quindi un problema di punto di vista. Esiste cioè un sottospazio dove la separazione delle classi è massima, come nella figura 5.32.

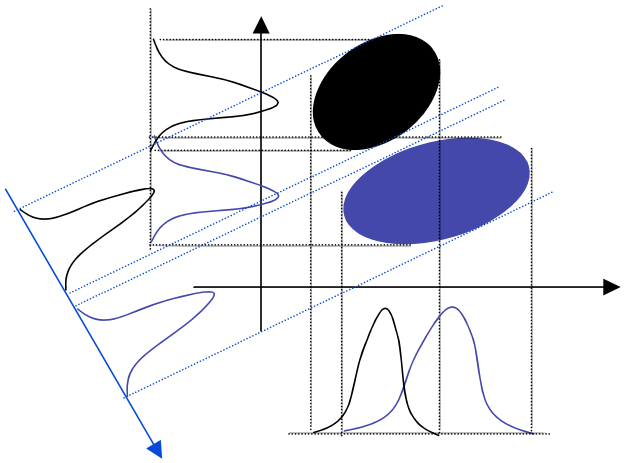


figura 5.31

Vediamo ora come si può affrontare in maniera sistematica il problema di assegnare un pattern ad una classe predefinita.
La differenza sostanziale sta nella natura della classe. Abbiamo infatti due possibilità, nella prima la classe è definita da un elemento mentre nella seconda ogni classe contiene più elementi.
Un esempio del primo caso è quello del "Character Recognition" ad esempio i metodi che consentono dalla immagine digitalizzata di un testo manoscritto di ricavare il testo in formato digitale (ad esempio come documento di testo). In questo caso ogni classe del problema è una lettera dell'alfabeto mentre il pattern è rappresentato dal valore dei pixel del carattere digitalizzato.
Il secondo caso è molto più generale e riguarda ad esempio la separazione degli individui in uomini e donne, oppure di una popolazione di frutti a seconda della cultivar e via dicendo. In questo caso ogni insieme contiene molti elementi differenti tra loro tranne per il particolare che definisce la classificazione. Gli elementi di questi insiemi avranno quindi una certa distribuzione definita dai differenti valori delle features.
In pratica nel primo caso la classe sarà identificata da un pattern univoco e i dati saranno delle deviazioni rumorose dai pattern che definiscono le classi. Nel secondo caso i pattern che descrivono gli elementi che costituiscono la classe definiscono una distribuzione statistica (PDF). Per cui ogni classe è identificata da una distribuzione di probabilità (PDF) dei pattern.

Classficatore K-nearest neigbhour (KNN)
Nel primo dei due casi la pattern recognition può essere agevolmente definita attraverso il cosiddetto "Template matching". Questo metodo è efficiente quando ogni classe contiene solo un pattern ed i pattern misurati (quello da assegnare alle classi) sono affetti da rumore additivo (no traslazione, rotazione, deformazione del pattern). Un esempio tipico di questo problema è l' Optical Character Recognition (OCR).
Per ogni classe, viene quindi definito, come in figura 5.32, un pattern tipo (template) che rappresenta il carattere ideale senza rumore aggiuntivo.
Per assegnare un pattern ad una classe è quindi sufficiente contare il numero di corrispondenza (matching) tra il pattern misurato ed i template delle classe ed assegnare il pattern a quella classe dove il numero di variabili del pattern è massimo (questo metodo è detto della correlazione massima). Altrimenti si può seguire la via opposta ed assegnare il pattern alla classe per il quale risulta minimo il numero di disaccordi (metodo dell'errore minimo).

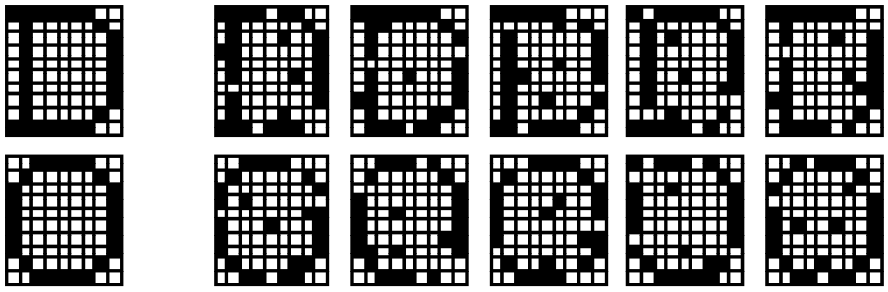


Figura 5.32

Nel caso precedente il pattern è rappresentato dai pixel dell'immagine, il valore di ogni pixel è binario (0-1) corrispondente ad una immagine bianco-nero.
Le immagini della figura precedente possono essere descritte in termini di pattern multidimensionali secondo la definizione data nell'introduzione di questo capitolo. La differenza tra pattern quindi può essere calcolata come differenza tra vettori. I templates di J classi

sono quindi i vettori m_j. Dato un pattern misurato x la distanza tra il pattern ad il template della classe K è data da:

$$\varepsilon_k = \left| \left| \vec{x} - \vec{m}_k \right| \right|$$

Geometricamente la situazione è raffigurata in figura 5.33.

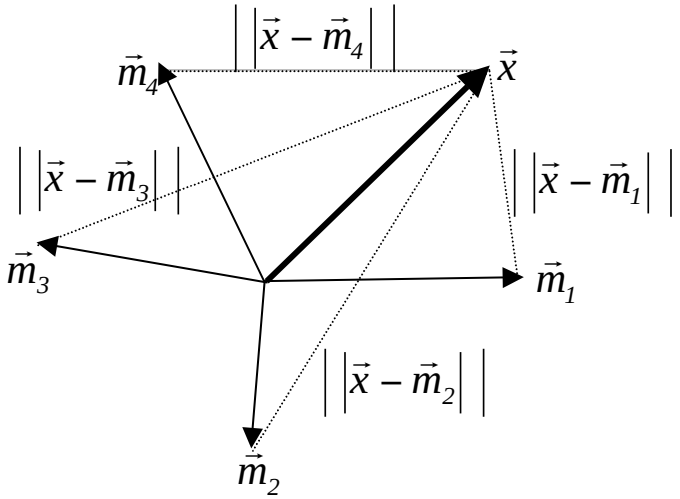


Figura 5.33

Si osservi che la dimensione dello spazio è fissata dalle dimensioni dei pattern, quindi dal numero delle features che definiscono la base-ortonormale dello spazio dei patterm.
L'errore nel caso del classificatore di minimo errore (minimum distance classifier). L'errore per la classe K si può scrivere come:

$$\varepsilon_k^2 = \left| \vec{x} - \vec{m}_k \right|^2 = \left(\vec{x} - \vec{m}_k \right)^T \left(\vec{x} - \vec{m}_k \right)$$
$$= \vec{x}^T \vec{x} - \vec{m}_k^T \vec{x} - \vec{x}^T \vec{m}_k + \vec{m}_k^T \vec{m}_k =$$
$$= -2 \left[\vec{m}_k^T \vec{x} - \frac{1}{2} \vec{m}_k^T \vec{m}_k \right] + \vec{x}^T \vec{x}$$

Dove l'ultimo termine essendo uguale per ogni classe può essere trascurato.
L'errore quindi è minimo se la seguente funzione è massima:

$$g(\vec{x}) = \left[\vec{m}_k^T \vec{x} - \frac{1}{2} \vec{m}_k^T \vec{m}_k \right]$$

la funzione $g(x)$ è detta funzione discriminante, tale funzione è pari alla correlazione del pattern x con il template m_k a cui si sottrae il quadrato del vettore di template.
Il metodo del classificatore minimo quindi corrisponde ad assegnare il pattern alla classe per la quale la correlazione tra pattern e template risulta massima. Come espresso dallo schema di figura 5.34.

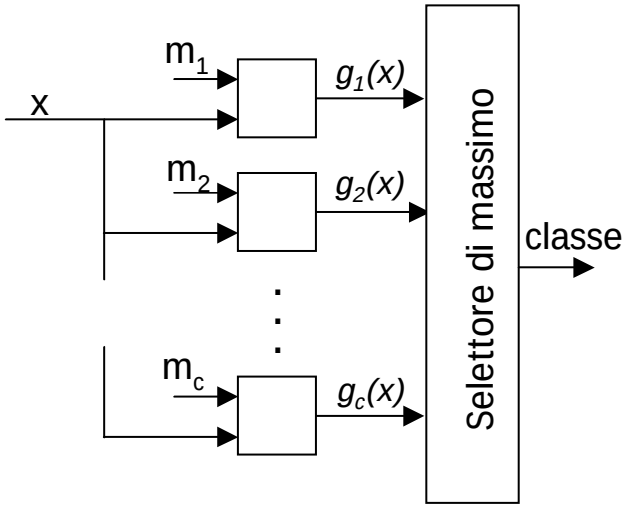


Figura 5.34

E' importante sottolineare che il template vector deve essere noto a-priori oppure deve essere stimato attraverso un processo di calibrazione del classificatore.
Nel caso di stima del template vector si procede considerando un numero N di misure di elementi di cui si conosce sicuramente la classe di appartenenza. Per ogni serie di pattern che descrivono campioni appartenenti alla stessa classe si calcola il vettore medio. In

questo modo si sottintende che le deviazioni tra pattern sono dovute ad errori di misura e non a caratteristiche proprie degli elementi del pattern. Attribuendo le deviazione ad errori di misura si può anche concludere che all'aumentare delle misure diventa sempre più certa la definizione dei template vectors.
Il metodo appena descritto è anche detto *k-nearest neighbour* in questo modi si mette in evidenza come questo metodo consenta di calcolare, tra un insieme di k template, il template a distanza minima dal pattern.
Da un punto di vista geometrico si può osservare come le k funzioni discriminanti suddividono lo spazio dei pattern in k regioni ciascuna pertinente ad una sola classe.
Come conseguenza della definizione del k-nearest neighbour esista un serie di punti dello spazio dei pattern che risultano equidistanti da almeno due template vectors, pattern che giacciono su questi punti sono quindi non attribuibili ad una unica classe. I luoghi di punti di non decisione sono i contorni delle regioni di pertinenza delle singole classi. L'operazione di suddivisione dello spazio in regioni corrispondenti alle classi è detta tessellazione. Le funzioni discriminati lineari (quelle cioè che descrivono la distanza euclidea tra vettori) danno luogo ai contorni di forma poligonaledi figura 5.35.
.

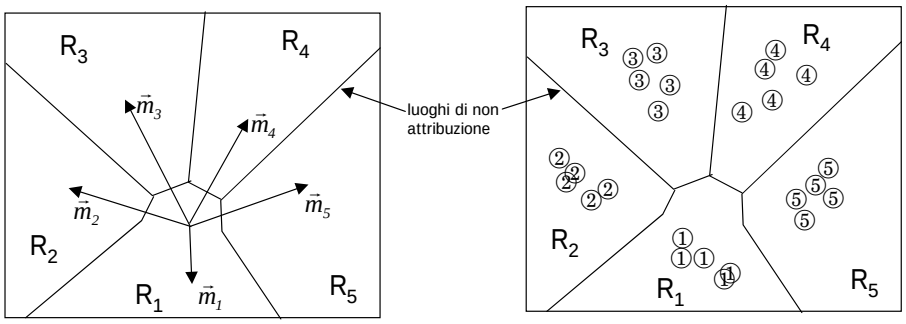


figura 5.35

Nella figura precedente vediamo a sinistra un esempio di tesselazione definita da 5 template vectors che danno luogo a cinque regioni. A destra vediamo la posizione di pattern originati da misure che sono assegnati alle classi di pertinenza della regione dello spazio in cui giacciono. I contorni delle regioni corrispondono all'inviluppo dei luoghi di punti equidistanti dai template vectors.

Classificatore Statistico

Il secondo caso enunciato precedentemente riguarda il caso, molto più generale, in cui gli elementi della classe sono descritti da pattern intrinsecamente variabili. Per cui ad ogni classe corrisponde una distribuzione di pattern piuttosto che un template. La distribuzione può in principio essere qualsiasi, proprio come qualsiasi può essere lo schema di classificazione. Vedremo comunque che nel caso in cui le classi siano distribuite gaussianamente è possibile definire un criterio di classificazione.

In termini generali, un pattern è assegnato alla classe verso la quale ha la massima probabilità di appartenere. Quindi il concetto di distanza metrica (ad esempio euclidea) tra pattern e classe (o meglio tra pattern e template) è sostituito dal concetto di distanza statistica per la quale un pattern è tanto più prossimo ad una distribuzione tanto maggiore è la probabilità di appartenere alla distribuzione stessa.

Per chiarire tale concetto centrale consideriamo il caso univariato e la distribuzione normale.

Supponiamo di avere due classi in ognuna delle quali la variabile x è distribuita normalmente ma con media e varianza differenti, come mostrato in figura 5.36.

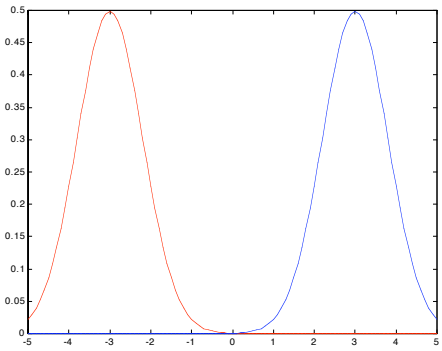


figura 5.36

In pratica possiamo dire che pattern caratterizzato da x>0 appartengono alla classe B, mentre pattern caratterizzato da x<0 appartengono alla classe A. In questo caso le distribuzioni sono disgiunte e l'appartenenza di un pattern ad una classe è ovvia (anche

se teoricamente la gaussiana si estende da -∞ a +-∞ la consideriamo comunque confinata nella porzione di spazio in cui la probabilità da luogo ad osservabili).
Nel caso invece in cui le distribuzioni siano parzialmente sovrapposte, come in figura 5.37, la situazione e leggermente differente.

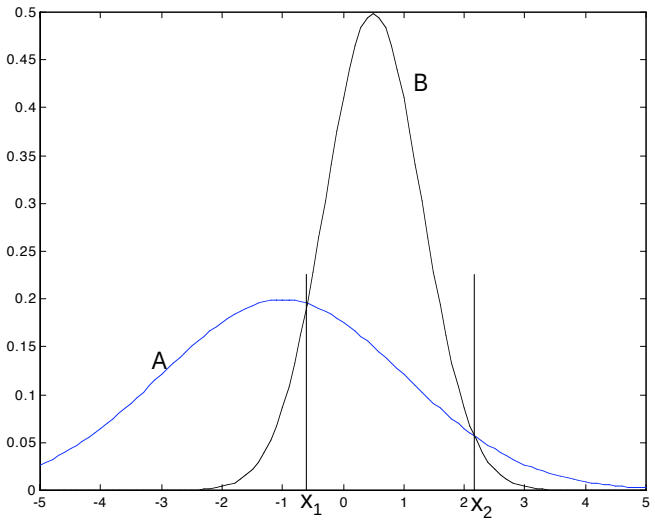


Figura 5.37

Nel caso mostrato in figura infatti le distribuzioni relative alle classi A e B sono sovrapposte e considerando la probabilità più grande possiamo definire tre regioni di appartenenza: per $x<x_1$ e per $x>x_2$ il pattern probabilmente appartiene alla classe A, mentre per $x_1<x<x_2$ il pattern appartiene alla classe B. Inoltre per $x=x_1$ ed $x=x_2$ la probabilità di appartenere alle due classi è uguali e quindi il pattern non può essere assegnato.
Lo stesso ragionamento può essere applicato al caso multivariato, quando cioè il pattern è multidimensionale e la distribuzione normale è multivariata. Il luogo di punti isoprobabile è in questo caso un iper-ellissoide definito attraverso la seguente forma quadratica che è pari alla probabilità che un vettore x appartenga ad una distribuzione normale definita da un vettore medio μ e matrice di covarianza Σ.

$$P(x) = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

la quantità precedente è detta *distanza di Mahalanobis* o distanza statistica.
Il metodo del classificatore KNN si generalizza al caso in cui le classi siano definte statisticamente sostituendo la distanza euclidea tra pattern e template vector con la distanza di Mahalanobis tra pattern e distribuzione statistica della classe.
In pratica, valutando la distanza di Mahalanobis di un pattern verso ogni classe di distribuzione statistica nota si assegna il pattern alla classe verso la quale la distanza di Mahalanobis è minore, cioè verso la quale è maggiore la probabilità di appartenenza secondo lo schema di figura 5.38.

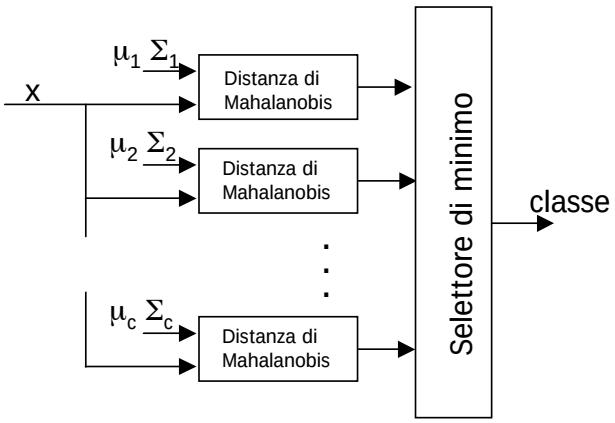


Figura 5.38

Come nel caso del KNN per template vectors la funzione di probabilità della classe è o nota a priori o stimabile dai dati sperimentali. Anche in questo caso la calibrazione del calibratore consiste nel misurare, per ogni classe, un numero N di elementi e di calcolare il pattern medio e la matrice di covarianza dei pattern. A differenza dei template vectors in questo caso gli errori di misura si possono confondere con la varianza propria della distribuzione della classe. In ogni caso anche in questo caso un numero elevato di campioni di calibrazione assicura una migliore stima della funzione di distribuzione.
E' opportuno a questo proposito richiamare il problema del bias già discusso nel capitolo 3. Soprattutto nel caso di classificazione di

campioni complessi bisogna porre estrema attenzione al fatto che i campioni di calibrazione siano effettivamente rappresentativi della classe e non siano una sottoclasse dotata di valori particolari dei patterm.

Come esempio di applicazione dei classificatori KNN e statistico consideriamo due cultivar di pesche. Ogni frutto è descritto da un pattern bidimensionale formato da due features: gradi brix ed acidità totale. La scelta di un pattern bidimensionale rende possibile la rappresentazione grafica dei due metodi.
La figura 5.39 mostra i pattern nel piano delle features.

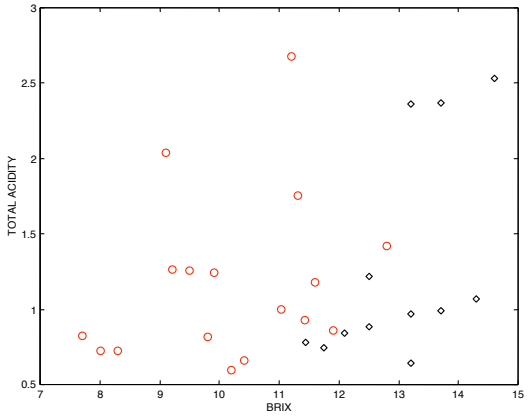


Figure 5.39

le due classi corrispondono a valori in genere differenti per quanto riguarda i gradi brix mentre per l'acidità non c'è una tendenza che separa le due classi.
Applichiamo il metodo del KNN. Per prima cosa bisogna calcolare il pattern medio delle due classi. Tali pattern fungono da template vectors per la classificazione. Da un punto vista grafico la costruzione della linea di separazione tra le due classi è semplice. Infatti ricordando che i luoghi equidistanti da un punto sono dei cerchi i punti di intersezione di cerchi equidistanti dai due template vectors formano la linea retta che separa il piano nelle due regioni di competenza delle due classi come mostrato in figura 5.40.

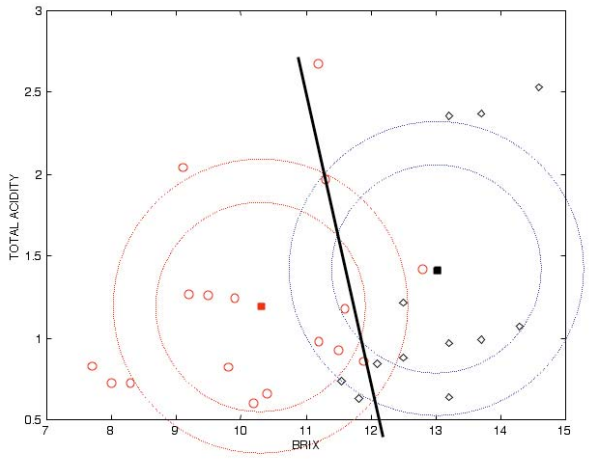


Figura 5.40

Come si può osservare in figura 5.40 alcuni pattern appartenenti ad una classe giacciono nel semipiano di competenza dell'altra. Ci sono in particolari due pattern della classe B che giacciono nel semipiano di A e un pattern della classe A che giace nel semipiano di B. In totale, 3 pattern sono mal classificati.
Per quanto riguarda il metodo del classificatore statistico basato sulla distanza di Mahalanobis il luogo dei punti di isoprobabilità giace sulle intersezioni delle ellissi luogo di punti di improbabilità delle due classi. L'intersezione di tali punti, mostrata in figura 5.41, da luogo ad una linea di separazione non lineare del secondo ordine.

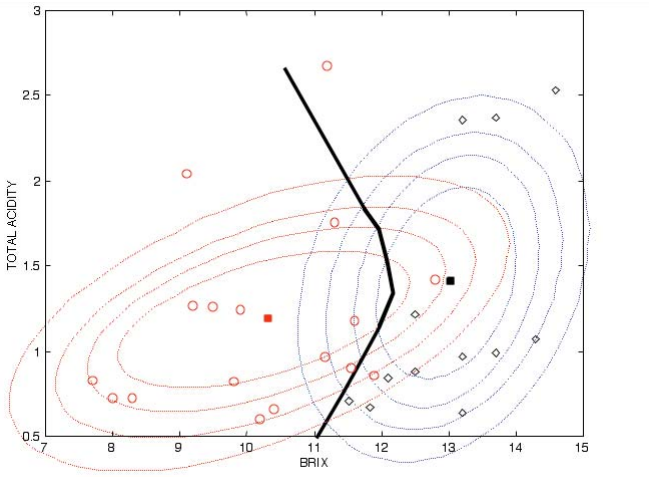


Figura 5.41

In questo caso, il numero dei pattern non correttamente classificati si riduce a 1, avendo solo un pattern della classe che giace nel semipiano di competenza della classe B.

5.6 Analisi discriminante e regressione

Il metodo della distanza di Mahalanobis richiede la conoscenza della distribuzione di probabilità degli elementi della classe. In molti casi, la distribuzione di probabilità dei patterns di ogni classe non è nota oppure non è gaussiana.

In questi casi non è possibile applicare un metodo come il KNN ma un classificatore non-parametrico indipendente dalla statistica delle classi e nelle quali si cerca una combinazione delle variabili che consente l'identificazione delle classi. Questa operazione è detta analisi discriminante. A seconda delle funzioni che si utilizzano si può definire una analisi discriminante lineare o nonlineare.

Nel caso lineare, il calcolo è simile alla multiple linear regression, nella quale vengono messe in relazione lineare la matrice X (insieme dei patterns dei campioni misurati) e una matrice Y (codifica numerica della class membership).

Ci sono vari metodi per rappresentare la class membership (concetto qualitativo) in termini numerici.

Il modo più utilizzato è quello della codifica "one-of-many". La matrice Y è costruita con un numero di colonne pari al numero delle classi nel problema per cui ogni colonna di Y quindi identifica una classe.
Dato un pattern X_i la corrispondente riga Y_i è quindi formata ponendo a zero tutti gli elementi tranne quello che corrisponde alla classe a cui appartiene Xi che viene fissato a 1.
In fase di identificazione il modello di regressione avrà una accuratezza finita, quindi il pattern viene assegnato alla classe il cui valore corrispondente risulta più grande.
Come abbiamo illustrato nel capitolo 4 il metodo ottimale per la soluzione dei problemi di regressione multipla lineare è la PLS. Quindi PLS, anche se non ottimale per la soluzione di problemi di pattern recognition, può essere usato come un classificatore. Il metodo è detto PLS-DA (PLS Discriminant Analysis).
Il metodo PLS-DA consente di sfruttare le caratteristiche di PLS in termini di scores e dei loadings la cui analisi è simile a quella vista nel caso della PCA.

Validazione in PLS-DA
La validazione in PLS-DA segue la stessa modalità discussa per la PLS ordinaria. In pratica si usano i metodi già discussi ed in particolare la leave-one-out.
La differenza più rilevante consiste nel fatto che la quantità da minimizzare non è l'RMSECV delle colonne della matrice Y ma la percentuale di classificazione, definita precedentemente come la percentuale rispetto al totale dei campioni dei pattern correttamente classificati.
Questa accortezza conduce a risultati spesso diversi in termini di numero ottimale di variabili latenti e consente di stimare la percentuale classificazione attesa.
Per illustrare il metodo consideriamo alcuni esempi:

Classificazione dei fertilizzanti per la coltura delle mele
Supponiamo di avere eseguito misure delle caratteristiche di mele ottenute con quattro differenti fertilizzanti: urea, nitrati e potassio, ammonio e solfati, più un metodo di controllo. Questo da luogo ad un problema a quattro classi. Ogni mela è stata caratterizzata da un pattern a sette dimensioni formato dal contenuto di azoto totale, azoto nei semi, fosforo, potassio, calcio, magnesio e dal peso.
Le matrici X e Y del problema sono quindi le seguenti:

Y				X						
control	urea	potassium nitrates	ammonium and sulfate	total nitroger	pit nitrogen	phosphorous	potassium	calcium	magnesium	weight
1	0	0	0	3240	1663	836	8747	218	388	97.3
1	0	0	0	3077	1663	891	8460	249	372	75.8
1	0	0	0	3205	1770	031	0575	261	376	70.5
1	0	0	0	3330	1755	889	8330	209	367	77.5
0	1	0	0	3755	1915	842	10375	145	408	108.2
0	1	0	0	5037	2180	930	10047	172	420	103.2
0	1	0	0	4753	2137	945	10447	160	421	95.3
0	1	0	0	4453	1967	850	9677	206	396	93.5
0	0	1	0	4200	2063	869	11190	184	398	111.8
0	0	1	0	5915	2050	1016	12060	184	461	109.7
0	0	1	0	5193	2210	958	11733	179	428	117.3
0	0	1	0	5347	2167	919	11910	191	420	99.6
0	0	0	1	5157	2357	1062	10210	136	401	86.7
0	0	0	1	7440	2975	1261	10820	122	446	85.5
0	0	0	1	6950	2527	1137	9710	191	408	70.8
0	0	0	1	4445	2075	846	8820	161	334	81.1

Il modello PLS-DA valuta la matrice Y con un certo errore che fa si che i valori predetti di Y non siano rigorosamente 0 e 1 ma valori continui. In questo esempio si ottiene:

Y vero				Y stimato			
1	0	0	0	**0.7839**	0.4456	-0.0168	-0.2128
1	0	0	0	**1.1728**	-0.3482	0.1035	0.0718
1	0	0	0	**0.8882**	0.1623	0.0387	-0.0892
1	0	0	0	**0.8729**	0.0584	-0.2114	0.2801
0	1	0	0	0.0515	**0.9332**	0.0552	-0.0398
0	1	0	0	0.0116	**0.9322**	-0.0086	0.0648
0	1	0	0	0.0748	**0.7485**	0.0089	0.1678
0	1	0	0	0.1801	**0.7226**	0.0919	0.0053
0	0	1	0	0.0482	-0.1203	**0.9887**	0.0835
0	0	1	0	0.0820	0.2404	**0.8671**	-0.1895
0	0	1	0	-0.0390	-0.0669	**1.0746**	0.0313
0	0	1	0	-0.1771	0.0942	**0.9599**	0.1230
0	0	0	1	0.1673	-0.0846	0.1036	**0.8137**
0	0	0	1	-0.2136	0.1390	-0.0245	**1.0990**
0	0	0	1	0.1372	-0.0736	0.0300	**0.9064**
0	0	0	1	-0.0410	0.2174	-0.0608	**0.8844**

Il metodo assegna il pattern alla classe per cui si ottiene il valore più grande. In questo modo possiamo osservare che i pattern sono stati tutti pienamente classificati. Cioè che PLS ha determinato un modello che assegna i pattern degli esempi alla classe corretta. La validità di tale modello a predirre pattern ignoti potrà essere valutata in un processo di validazione basato ad esempio sul metodo *leave-one-out*.
Il vantaggio nell'uso di PLS consiste nel fatto che ad esempio studiando i loadings del modello è possibile collegare le variabili con

le classi. Ad esempio consideriamo in figura 5.42 il biplot delle prime due variabili latenti del modello PLS-DA.

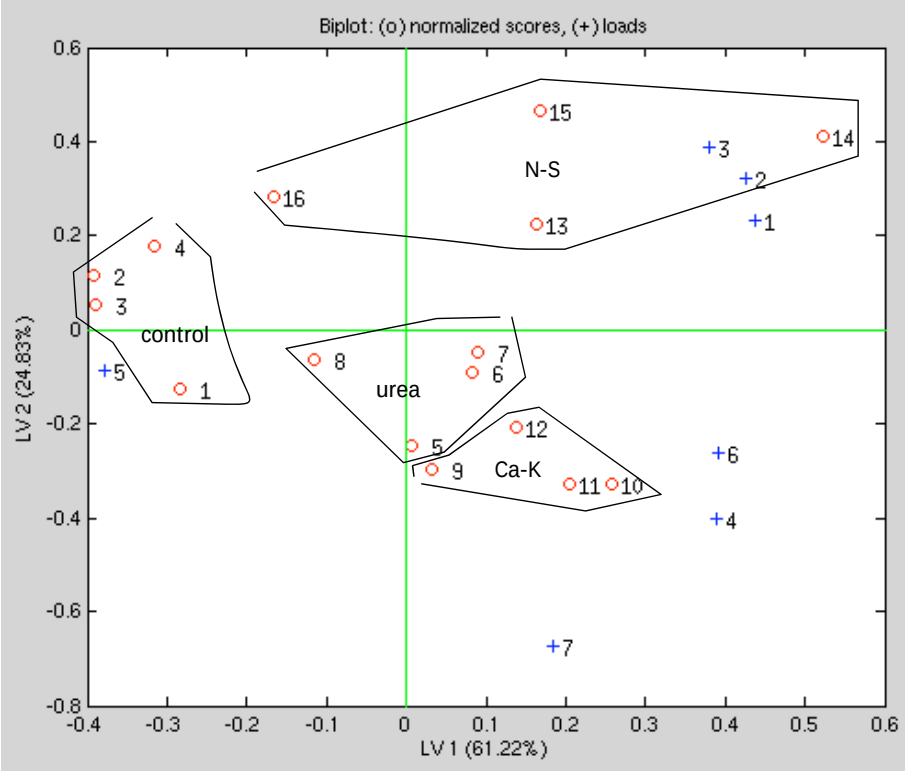


Figura 5.42

Gli Scores mostrano la separazione tra i quattro gruppi che identificano due direzioni di allontanamento dal fertilizzante di controllo.
I Loadings del grafico mostrano invece come il trattamento ammonio-solfati (NS) comporta un aumento dell'azoto totale e nei semi e del contenuto di fosforo. D'altro canto i trattamenti con urea e nitrati Potassio comportano un incremento nel contenuto di potassio, magnesio e del peso del frutto.
Inoltre, la quantità di calcio più elevata si trova nelle mele coltivata con il fertilizzante di controllo.

Classificazione del tipo di nocciole dallo spettro NIR
Consideriamo lo stesso set di dati descritto nell'esempio della PLS discusso nel capitolo 4. In questo caso siamo interessati a classificare gli spettri nelle cinque classi relative al tipo di nocciole. Il modello PLS-DA risulta ottimizzato da 18 variabili latenti. Anche in questo caso la matrice degli spettri è stata centrata (media nulla).
Il risultato della classificazione viene di solito espresso attraverso una matrice, detta matrice di confusione. Lungo la diagonale il numero dei campioni correttamente classificati. Mentre i campioni fuori diagonale rappresentano i campioni non correttamente classificati. La percentuale dei campioni correttamente classificati rappresenta la percentuale di classificazione. In questo caso si hanno solo 2 spettri non correttamente classificati. Entrambe corrispondono alla classe 5 e sono stati attribuiti alle classe 2 e 3. in totale la percentuale di dati correttamente classificati è del 98%.

```
confmat =

    45     0     0     0     0
     0    15     0     0     1
     0     0    23     0     1
     0     0     0    10     0
     0     0     0     0     9


percvin =

   98.0769
```

La figura seguente mostra lo score plot delle prime due variabili latenti. Poiché il modello PLS-DA che raggiunge la pressoché totale classificazione dei dati è formato da 18 variabili latenti, le prime due di queste mostrano solo una classificazione parziale dove solo due classi sono correttamente separate dal resto. Questo nonostante lo score plot, mostrato in figura 5.43, rappresenti circa l'86% del totale della varianza.

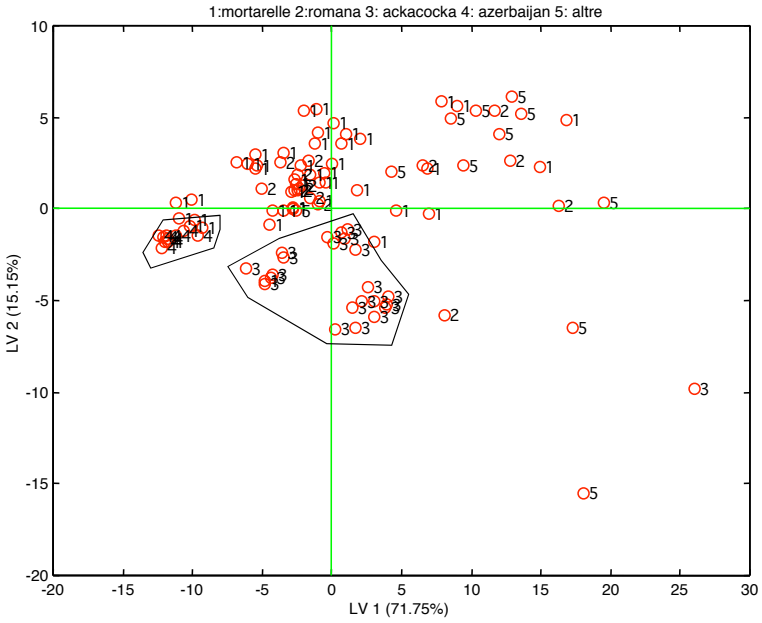

1:mortarelle 2:romana 3: ackacocka 4: azerbaijan 5: altre
LV 2 (15.15%)
LV 1 (71.75%)
10
5
0
-5
-10
-15
-20
-20
-15
-10
-5
0
5
10
15
20
25
30

Figura 5.43