Introduzione alla Chemiometria

Roberto Todeschini

Milano Chemometrics and QSAR Research Group

1

INTRODUZIONE ALLA CHEMIOMETRIA

1.1 - Introduzione

Questo libro vuole essere un contributo allo sviluppo di una strategia generale rivolta ad affrontare la complessità mediante l'utilizzo dei metodi chemiometrici. Il testo non ha la pretesa di trattare in modo esaustivo tutti i capitoli della chemiometria: non vengono infatti presi in considerazione campi pur importanti quali quelli del disegno sperimentale e dell'ottimizzazione, i cui contenuti possono essere ritrovati in altri testi.
Si è invece posto l'accento sulle strategie di analisi multivariata più tradizionalmente considerate come parti centrali della chemiometria, quali l'analisi in componenti principali, i metodi di *cluster analysis*, di classificazione e di regressione.
Nell'impostazione del testo si è cercato di trattare tutti gli argomenti fornendo sia le basi teoriche sia semplici esempi numerici. Diverse concessioni vengono fatte alla terminologia inglese, affiancando alla terminolgia italiana la dizione inglese, col preciso scopo di fornire al lettore anche l'immediato riscontro della terminologia che viene inevitabilmente incontrata in tutti gli articoli scientifici e nei testi internazionali. In qualche caso si è volutamente evitato di tradurre in lingua italiana alcune denominazioni inglesi il cui valore semantico nella lingua originale è traducibile solo a rischio di perdere alcune connotazioni fondamentali.
La lettura del testo non richiede conoscenze matematiche e statistiche particolarmente avanzate. In ogni caso, nell'*Appendice A* vengono ripresi i fondamenti dell'algebra matriciale (definizioni e operazioni) al fine di facilitarne la lettura e nell'*Appendice B* viene fornita una breve elencazione dei più importanti indici statistici, insieme ad alcuni elementi fondamentali di statistica. L'*Appendice C* contiene alcune dimostrazioni matematiche riguardanti proprietà delle componenti principali e del metodo di regressione PLS. L'*Appendice D* raccoglie, infine, i dati relativi agli esempi utilizzati nel testo. Infine, le

Appendici E e *F* riportano un elenco di siti Internet collegati ai temi trattati nel teso e una bilbiografia generale.

1.2 - Che cosa è la chemiometria

Viviamo in tempi di grandi mutamenti: mutano i sistemi geopolitici, i grandi riferimenti ideologici, le esigenze e le domande della gente. Sia il mondo industriale che gli enti preposti alla formazione tecnica e culturale non solo si trovano coinvolti in tali mutamenti, ma divengono sempre più soggetti e parte attiva del cambiamento stesso. In tale quadro, sia l'industria che le università devono essere in grado, ciascuna nell'ambito dei propri compiti, di dare risposte immediate anche di fronte a problemi di alta complessità. E proprio la complessità svolge un ruolo cruciale nel definire le linee di sviluppo, le strategie e le scelte metodologiche di discipline nuove come la chemiometria.
Ma che cosa intendiamo per problemi ad alta complessità? In molti casi reali il processo o il sistema in esame non può essere descritto o razionalizzato alla luce di una teoria ben definita; per questi problemi spesso le teorie costituiscono solo uno sfondo di conoscenza che permette l'analisi del problema, ma non consente di risolvere il problema specifico.
Si pensi, ad esempio, ai problemi di natura ambientale, ai problemi di carattere farmacologico e geologico, a quelli relativi ai processi di produzione a molti stadi, ai problemi legati all'ottimizzazione e al controllo di processi industriali, ai problemi connessi alla ricerca di risorse e al loro sfruttamento a fini economici. In problemi di questa natura, le variabili potenzialmente implicate non solo sono numerose, ma non tutte sono controllabili con la precisione desiderata, di molte non si conosce esattamente la rilevanza per il problema in esame e quanto rumore sperimentale mascheri e confonda i veri effetti delle variabili considerate.
Inoltre, nella maggior parte dei casi, non sono note le correlazioni tra le variabili e i loro effetti sinergici, cioè quegli effetti dovuti al ruolo combinato di due o più variabili, effetti che non si manifestano considerando separatamente una variabile alla volta. In molti casi, la complessità del sistema in esame può portare alla comparsa di effetti di non-linearità non prevedibili a priori o si manifesta attraverso disomogeneità del sistema stesso.
La complessità di un sistema si ripercuote necessariamente sulla complessità intrinsecamente contenuta nei dati relativi al sistema stesso. Come mostrato in Fig. 1-1, i metodi chemiometrici cercano di separare il contenuto di informazione utile da quanto altro è contenuto nei dati: la presenza di rumore

sperimentale, di informazione ridondante dovuta ad effetti di correlazione tra le variabili, la presenza di informazione di buona qualità ma non direttamente interessante per il problema studiato.
Pertanto, per affrontare problemi di questo tipo, qui definiti ad alta complessità, non potendo ricorrere a teorie esplicite, occorre rovesciare l'approccio al problema: diviene cioè necessario cercare di estrarre dai dati sperimentali, anche mediante prove sperimentali opportunamente ideate, l'informazione rilevante e pertinente in esse contenuta.
Utilizzare tutte le informazioni disponibili dall'esperienza pregressa ed ottimizzare le procedure per lo sviluppo di nuove strategie, minimizzando i tempi ed i costi e massimizzando la qualità dei risultati, sta divenendo l'unica prospettiva realistica per raccogliere la sfida ed affrontare concretamente i problemi.
La chemiometria nasce dal tentativo di rispondere a queste esigenze armonizzando competenze provenienti da settori della conoscenza molto diversi tra loro, quali la statistica, l'informatica, la matematica, le scienze sperimentali, ma sempre prestando un'attenzione primaria alla soluzione di problemi reali, da qualunque parte essi provengano.
Nonostante qualcuno abbia definito la chemiometria come la scienza dell'ovvio trattata con metodi complessi, approfittiamo dell'occasione per correggere tale definizione nella definizione più propria: la chemiometria è la scienza del complesso trattata con metodi ovvi. I metodi ovvii sono per lo più quei metodi che matematici e statistici hanno ideato per la soluzione di problemi multivariati e che, grazie anche allo sviluppo dei computer e più in generale dell'informatica, sono oggi facilmente utilizzabili. In modo altrettanto ovvio, i contesti specifici in cui queste metodologie sono state applicate hanno in molti casi fornito l'occasione di verificare, modificare, rielaborare e proporre metodi che oggi risultano anche innovativi nell'affrontare le problematiche di scienze quali la chimica, la farmacologia, le scienze ambientali, per le quali vi è una crescente richiesta di analisi sempre più complesse.
In realtà, i metodi chemiometrici, a dispetto del loro nome, hanno oggi superato l'ambito puramente chimico e vengono utilizzati come una metodologia generale in grado di estrarre informazioni da dati di qualsiasi natura.
I metodi chemiometrici vengono utilizzati per l'*esplorazione dei dati*, cioè per aprire una finestra sulla complessità al fine di gettare luce sulla struttura dei dati, sulle relazioni e correlazioni tra essi esistenti, sulla congruità, sulla rilevanza e sulla ridondanza con cui il problema è stato descritto. I dati reali si presentano comunemente come un insieme olistico in cui informazione utile e secondaria, rumore e ridondanza sono intrinsecamente mescolati (Fig. 1-1):

separare le diverse sorgenti e i diversi effetti è uno degli obiettivi dell'esplorazione dei dati.

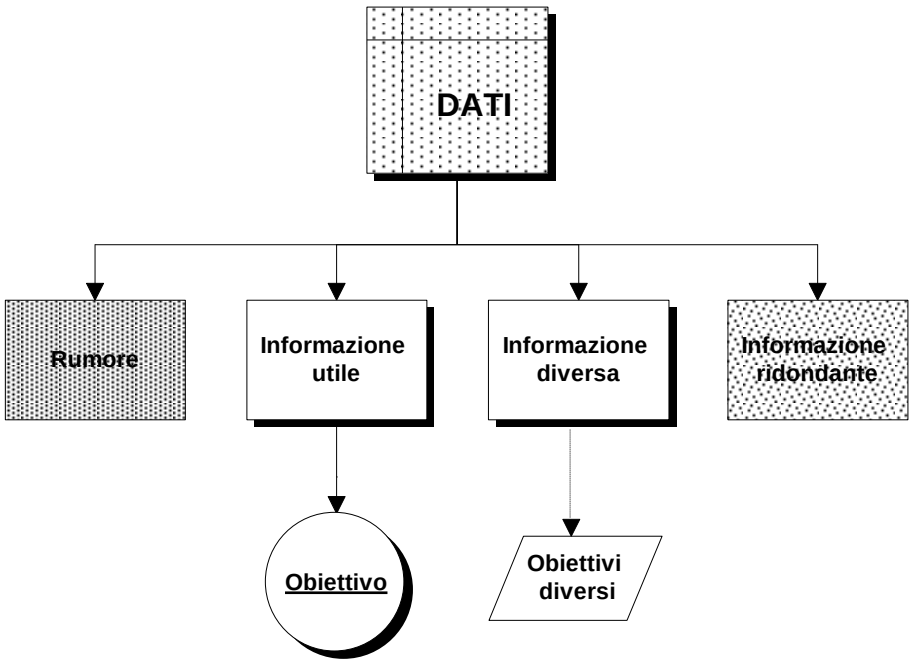


FIG. 1-1

Sono normalmente disponibili strumenti sofisticati per costruire *modelli qualitativi e quantitativi*, per evidenziare la *presenza di disomogeneità*, cioè di raggruppamenti di dati relativi a situazioni tra loro differenti, per "vedere" e rappresentare sistemi complessi con strumenti grafici efficaci.
Il grande numero, la flessibilità e l'adattabilità dei metodi chemiometrici appare a molti come un'intrinseca debolezza di questa metodologia di ricerca in quanto essa sembra poter fornire in ogni caso qualche soluzione. Tuttavia tale caratteristica è certamente poco scientifica solo per coloro i quali, non avendo affatto approfondito (o capito) né gli aspetti teorici né i limiti e il campo di

validità delle strategie chemiometriche, non sono di fatto in grado di valutare la correttezza dei risultati ottenuti.
I metodi chemiometrici appaiono dunque come una via percorribile (e in molti casi l'unica) per affrontare problemi di alta complessità per i quali non sono disponibili strategie teoriche precostituite e ben fondate.
Il grande sviluppo dei computer e la loro facile accessibilità hanno consentito lo sviluppo di metodologie come quelle chemiometriche che offrono al ricercatore una vasta gamma di metodi e soluzioni potenziali che possono essere provati e controllati in tempi brevi.
La chemiometria è una disciplina scientifica giovane, la cui nascita si deve attribuire all'iniziativa congiunta di Svante Wold, dell'Università di Umeö in Svezia, e di Bruce Kowalski, dell'Università di Seattle negli Stati Uniti. Con una lettera inviata alla rivista *Analytical Chemistry* nel giugno del 1974, Wold e Kowalski proposero di chiamare *Chemometrics* un settore scientifico che comprendesse tutte quelle tecniche matematiche rivolte a trattare, elaborare e modellare insiemi di dati chimici.
Il ***dato chimico***, diversamente da altri tipi di dati, possiede alcune peculiarità che lo rendono in qualche modo unico. In primo luogo, il concetto di ***composizione chimica***, cioè della costituzione della materia e della sua decomposizione logica nei termini dei suoi costituenti chimici puri, trascende l'ambito puramento chimico ed è conoscenza fondamentale in tutto il contesto scientifico. Dalla medicina alla biologia, dalle scienze geologiche alle scienze agrarie ed alimentari, dal controllo ambientale al controllo di qualità, ai problemi della conservazione dei beni culturali, il problema della composizione chimica di un campione e la ricerca di correlazioni con le sue proprietà è sovente un aspetto fondamentale. Il concetto di composizione chimica gioca quindi un ruolo unificante di straordinaria rilevanza scientifica e conoscitiva.
In secondo luogo, il chimico, mediante i processi di ***sintesi***, costruisce di fatto il proprio oggetto di studio, oggetto di cui, per lo più, sono note o prevedibili solo alcune delle sue caratteristiche e proprietà. La chimica è stata da sempre l'unica disciplina scientifica (oggi questa peculiarità è comune anche alla biologia molecolare) capace di *creare essa stessa, attraverso la sintesi, il suo oggetto di studio*: è fondata essenzialmente su una realtà costruita, ben diversa da quella prodotta dal costruttivismo tecnologico, ove l'oggetto costruito è teoricamente noto a priori in tutte le sue proprietà.
Come afferma Bachelard in "Il materialsmo razionale" (1972), parlando della chimica, "... si tratta di una scienza *costruttiva* della materia, di una scienza che costruisce il suo oggetto, i suoi nuovi oggetti. .... L'ora delle classificazioni in specie e generi è finita dall'istante in cui l'attività costruttiva lavora su piani razionali e moltiplica le possibilità di creazione."

Infine, il concetto di ***struttura molecolare*** ha aperto un livello di spiegazione scientifica unico nelle scienze sperimentali come terreno di mediazione tra la composizione atomica di un composto e le sue proprietà: esso è il campo d'azione in cui ricercare le correlazioni tra materia e sue proprietà.
Il dato chimico riveste quindi un ruolo unico e di grande rilevanza nello sviluppo conoscitivo di molte discipline scientifiche, a prescindere dalla chimica stessa. Lo sviluppo stesso degli strumenti analitici e spettroscopici ha reso disponibili rappresentazioni multivariate dei composti estremamente ricche di informazione, ma che per la stessa ragione richiedono per il loro utilizzo strumenti matematici molto più sofisticati.
La chemiometria è andata via via cercando di raccordare e unificare tutti gli strumenti matematici più idonei a trattare i dati chimici alla ricerca di quelle risposte teoriche o pragmatiche per la soluzione di tutti i problemi nel campo della spiegazione chimica.
La chemiometria oggi raccoglie al suo interno *i metodi di modellamento di classificazione e di regressione, l'analisi di similarità, l'analisi delle componenti principali e i diversi metodi ad essa collegati, i sistemi esperti* e *i metodi di intelligenza artificiale, le strategie basate sulle reti neurali, i metodi di disegno sperimentale e di ottimizzazione.* Tuttavia, pur nella sua ormai consolidata autonomia culturale, non si deve dimenticare che numerose metodologie chemiometriche si richiamano anche alla *statistica applicata*: anzi, possiamo dire che la chemiometria sia uno dei campi che consente maggiormente di effettuare test, verifiche e modifiche pratiche di molte tecniche proposte dagli statistici.
Dal 1974 molta strada è stata fatta: la disciplina *Chemometrics* è divenuta a buon diritto una parte rilevante di molti curricula universitari e ha trovato largo impiego nella soluzione di problemi delle industrie più diverse.
Esistono oggi due riviste specifiche di chemiometria: il *Journal of Chemometrics* (Wiley, USA) e il *Chemometrics and Intelligent Laboratory Systems* (Elsevier, Olanda). Altre riviste importanti riportano oggi numerosi lavori in cui i metodi chemiometrici non sono solo utilizzati come strumenti, ma rivestono un ruolo primario nel lavoro scientifico trattato. Tra queste riviste, citiamo *Analytical Chemistry, Analitica Chimica Acta, Trends in Analytical Chemistry, Journal of Chemical Information and Computer Sciences, Technometrics, Journal of Medicinal Chemistry, Chemosphere, Quantitative Structure-Activity Relationships, SAR and QSAR, Journal of Food Science*.
L'Italia è oggi uno dei centri europei importanti nello sviluppo della chemiometria, con i suoi tre centri universitari dove la chemiometria è particolarmente sviluppata anche nei suoi aspetti teorici: Dipartimento di Scienze Tecnologiche, Farmaceutiche e Alimentari dell'Università di Genova

(prof. Michele Forina), Dipartimento di Chimica dell'Università di Perugia (prof. Sergio Clementi) e Dipartimento di Scienze dell'Ambiente e del Territorio dell'Università degli Studi di Milano (prof. Roberto Todeschini; sito Internet: *www.disat.unimi.it/chm*).

1.3 - La chemiometria nel contesto scientifico

Per comprendere meglio il ruolo e la collocazione che la chemiometria ha nel contesto scientifico attuale, possiamo fare un breve riferimento ad alcuni concetti tipici della filosofia della scienza.
Possiamo infatti affermare, se pur in modo molto semplicistico, che la ricerca scientifica tradizionale si è sviluppata in accordo con lo schema di Fig. 1-2.
Secondo questo schema, lo sviluppo scientifico prende le mosse dai tentativi di risolvere problemi. Questo tentativo si manifesta attraverso un processo di razionalizzazione rivolto alla soluzione dei problemi, cioè attraverso lo sviluppo di un processo cognitivo che porta alla costruzione di teorie più o meno formali. Lo sviluppo di questo processo comporta la possibilità di "vedere i fatti" attraverso una mediazione teorica che li ingloba in modo coerente nella teoria e suggerisce allo stesso tempo la produzione controllata di nuovi fatti - gli esperimenti! - allo scopo di verificare le asserzioni meno ovvie che discendono dalla teoria stessa. L'interpretazione degli esperimenti consente la soluzione dei problemi evidenziati oppure la modifica di parti della teoria in modo da poter inglobare coerentemente eventuali risultati inattesi o di mettere in evidenza nuovi problemi da risolvere.

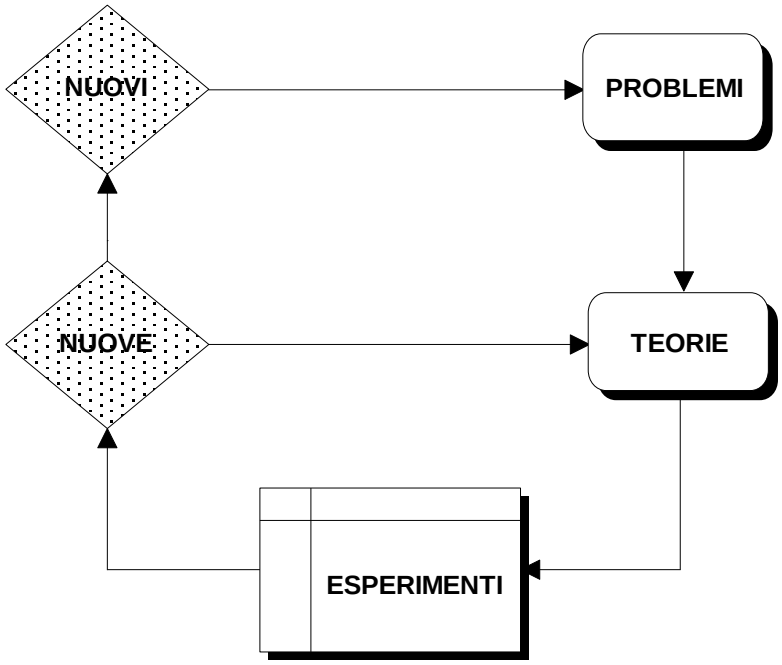


FIG. 1-2

La strategia ora delineata è sostanzialmente quella su cui si basa lo sviluppo della scienza normale.
Lo sviluppo stesso della scienza, fondato sull'evoluzione delle teorie scientifiche e della tecnologia, ha portato al presentarsi di problemi ad alta complessità, complessità che va intesa non solo come *relativa* a uno sviluppo ancora non adeguato delle conoscenze, ma una complessità che possiede anche un *carattere intrinseco* ineliminabile.
Il carattere peculiare dei problemi ad alta complessità scaturisce dalla presenza contemporanea di diversi aspetti della complessità:

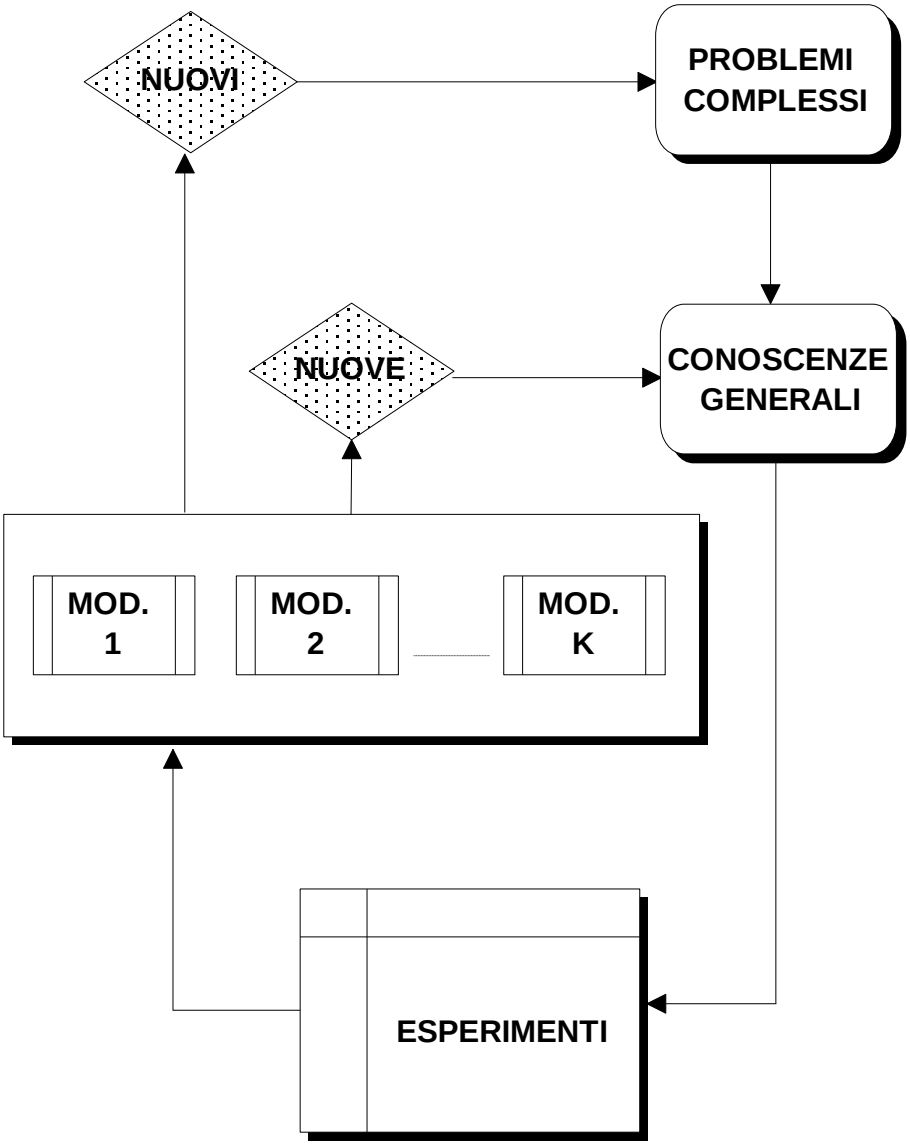

NUOVI
PROBLEMI COMPLESSI
NUOVE
CONOSCENZE GENERALI
MOD. 1
MOD. 2
MOD. K
ESPERIMENTI

FIG. 1-3

- presenza di molte potenziali variabili importanti nella descrizione del sistema
- non conoscibilità a priori della rilevanza delle variabili in gioco
- presenza di rumore sperimentale, di informazioni spuria, di correlazioni casuali
- presenza di correlazione sistematica tra alcune delle variabili considerate
- presenza di effetti sinergici e antagonisti
- presenza di effetti di non-linearità tra descrittori e risposta
- emergenza di effetti olistici e di macroproprietà del sistema
- impossibilità di un campionamento o di una sperimentazione adeguati

In questi casi, lo sviluppo delle conoscenze rivolte alla soluzione di problemi complessi non è in grado di compiersi completamente in accordo allo schema tradizionale. Quindi, se la spinta per lo sviluppo scientifico rimane la soluzione di problemi (complessi), le teorie per trattare questi problemi non appaiono, in molti casi, così ben definite e formali da consentire la soluzione diretta del problema (***hard model****Error! Bookmark not defined.*). Le teorie consolidate della fisica, della chimica, della chimica fisica, della biologia, eccetera, sono, in questo contesto, delle conoscenze di sfondo in base alle quali si cerca di progettare una sperimentazione ad alto contenuto di informazione. Ci si muove quindi alla luce di teorie di sfondo, si opera in accordo a principi e regole generali che, tuttavia, non sono in grado di fornirci i parametri effettivi per giungere ad una soluzione per il problema specifico. In questo caso, in base alle indicazioni suggerite dalle conoscenze generali, possiamo predisporre piani sperimentali dai quali si cerca successivamente di estrarre la massima informazione utile al fine di costruire dei modelli e sviluppare teorie sempre più specifiche.
In base allo studio dell'informazione contenuta nei dati sperimentali si cerca di costruire uno o più modelli (***soft model****Errore. Il segnalibro non è definito.*) che rispondano alle necessità contingenti (la soluzione del problema). In questo caso, tuttavia, poichè la soluzione viene fornita da modelli locali, la loro accettazione è subordinata al fatto che i modelli siano comunque coerenti con le conoscenze di sfondo e che si adottino delle misure precauzionali (i criteri di validazione) nella valutazione della loro affidabilità. Proprio l'utilizzo sistematico delle tecniche di validazione, tecniche capaci di stimare il potere predittivo dei modelli, cioè le prestazioni dei modelli ottenuti in predizione, contrapposte alla capacità dei modelli di descrivere i dati, sono una caratteristica qualificante dei metodi chemiometrici, consentendo un'applicabilità di queste tecniche ai più svariati problemi e nelle più diverse condizioni.

Questa strategia generale non esclude, evidentemente, che un insieme di modelli locali di provata qualità ed affidabilità possa costituire il nucleo iniziale di una nuova teoria o lo sviluppo di una teoria già consolidata in grado di affrontare in modo sistematico anche problemi ad alta complessità.
In questi casi, la chemiometria, come altre differenti metodologie basate su principi statistici, supera questi ostacoli dovuti all'assenza di teorie specifiche cercando di estrarre l'informazione utile dai dati sperimentali stessi e di produrre uno o più modelli come possibili soluzioni.
D'altra parte, proprio queste stesse considerazioni sottolineano l'incongruenza di utilizzare queste tecniche in tutti quei casi per i quali esiste una teoria ben definita e dettagliata (*hard model*) in grado di fornire risposte dirette ai singoli problemi concreti.

1.4 - L'analisi di un problema complesso

L'insieme delle strategie chemiometriche e delle relazioni tra le differenti procedure può essere visualizzato secondo lo schema di Fig. 1-4.
In generale, un problema viene affrontato attraverso una serie coordinata di passi che cominciano sempre dall'*analisi del problema*. Vengono individuati gli obiettivi e formulate le prime ipotesi di lavoro tenendo conto anche della disponibilità o meno di dati storici, cioè di precedenti dati sperimentali. Nel caso in cui questi non siano disponibili, è necessario passare alla *pianificazione degli esperimenti*. Questo è un passo di importanza cruciale che viene effettuato utilizzando i metodi del **disegno sperimentale**. Si tratta di un passo fondamentale in quanto un buon piano sperimentale consente di ottenere dei dati sperimentali con un *alto contenuto di informazione mirato alla soluzione del problema*: ci si pone così nelle condizioni ottimali per poter pervenire, mediante le successive elaborazioni dei dati, alla soluzione del problema. Il disegno sperimentale ci guida nella realizzazione degli esperimenti nelle migliori condizioni sperimentali al fine di ottenere col minimo sforzo (minimo numero di esperimenti, minimi costi, eccetera) informazione della migliore qualità.
Una volta definito il piano sperimentale, si passa all'*effettuazione degli esperimenti*: come si nota, questo è solo un passo nel contesto complessivo che deve portare alla soluzione del problema.
Una volta ottenuti i dati sperimentali, i passi successivi riguardano l'elaborazione degli stessi, utilizzando una serie di strumenti matematici che dipende dagli obiettivi prefissati.

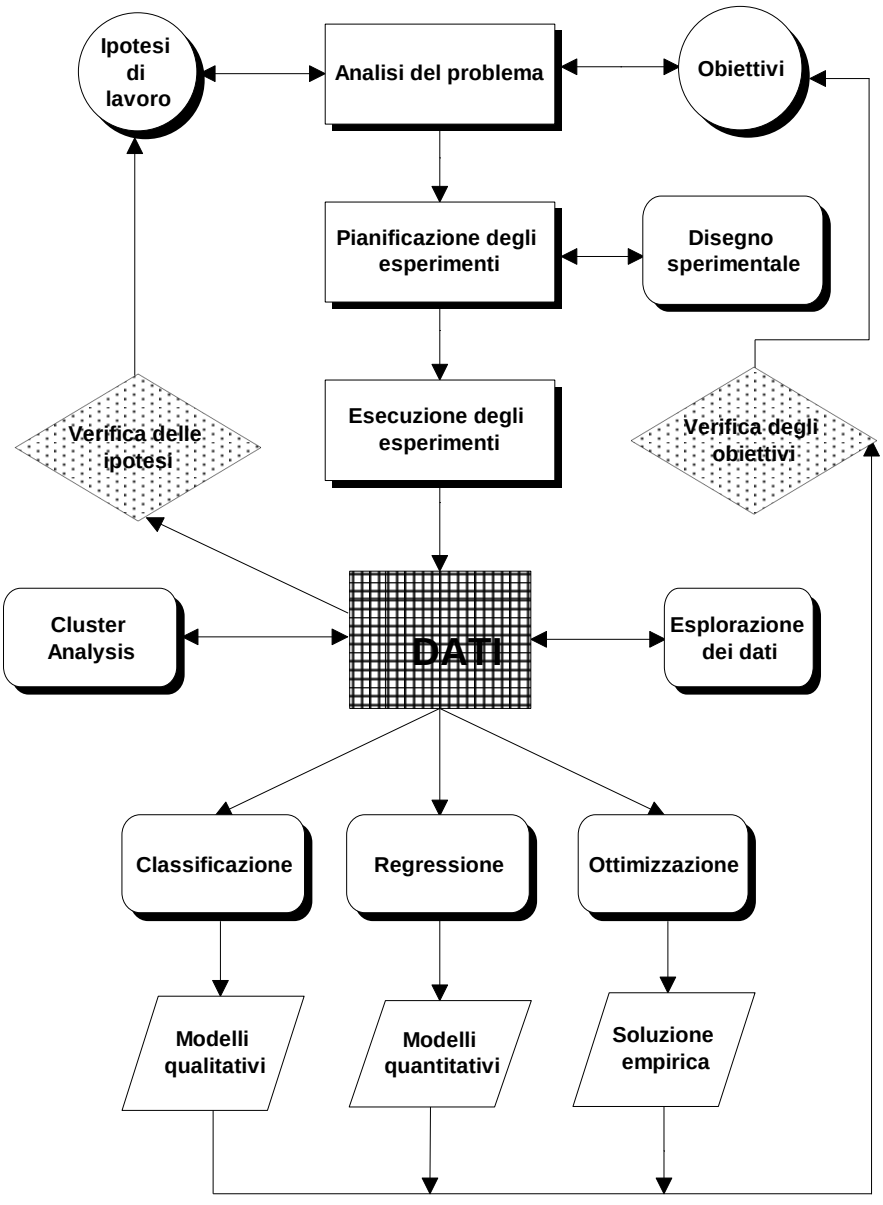

Ipotesi di lavoro
Analisi del problema
Obiettivi
Pianificazione degli esperimenti
Disegno sperimentale
Esecuzione degli esperimenti
Verifica delle ipotesi
Verifica degli obiettivi
Cluster Analysis
DATI
Esplorazione dei dati
Classificazione
Regressione
Ottimizzazione
Modelli qualitativi
Modelli quantitativi
Soluzione empirica

FIG. 1-4

Supponendo quindi che sia disponibile un insieme di dati costituito da *n* oggetti, ciascuno descritto da *p* variabili, una fase di carattere generale è **l'analisi esplorativa dei dati**, mediante la quale è possibile estrarre informazioni relative ai parametri statistici di ogni variabile, alle correlazioni tra le variabili, alle ipotesi di lavoro più interessanti, alla presenza di eventuali dati anomali ed altro ancora. In particolare, l'*analisi delle componenti principali* (ed i metodi ad essa correlati) è oggi unanimemente riconosciuta come uno dei metodi fondamentali per esplorazione iniziale di dati relativi a sistemi multivariati.
Mediante l'utilizzo dei metodi di ***cluster analysis*** è possibile valutare la similarità relativa tra i campioni studiati e quindi l'eventuale presenza di informazione addizionale dovuta alla presenza di gruppi (*clusters*). L'analisi della similarità conserva sempre, per sua natura, un margine di soggettività che si riflette sulla decisione di quanti e quali gruppi omogenei di dati sono presenti. Se questi gruppi sono interpretabili o, per qualche ragione, comunque accettati, i gruppi ottenuti sono elevati al rango di classi. Ciò significa ritenere che gli oggetti siano rappresentativi di un certo numero di classi, ovvero che i campioni rappresentano popolazioni differenti. Insieme ai dati in cui gli oggetti vengono separati in classi, esistono anche dati per i quali esistono a priori delle classi naturali alle quali appartengono. In ogni caso, una volta assegnati gli oggetti ad una classe, i **metodi di classificazione** consentono di costruire modelli di classificazione, cioè modelli matematici, funzione delle variabili utilizzate e delle classi definite, in grado di predire l'appartenenza di un nuovo campione ad una delle classi note. Ciò significa che se il modello calcolato assume un carattere di validità generalmente accettata, esso diviene indipendente dallo specifico insieme utilizzato, costituito da *n* campioni (*training set*).
Un altro tipo di problemi riguarda il caso in cui si cerca di costruire un modello matematico che permetta di predire valori quantitativi di una variabile (la risposta) dai valori, noti per *n* campioni, assunti da un insieme indipendente di variabili (i predittori).
I metodi che portano a soluzioni di questo tipo sono chiamati **metodi di regressione**. Anche in questo caso, se si ritiene che il modello possa assumere un carattere di validità più generale, esso diviene indipendente dallo specifico insieme di dati utilizzato.
Infine, esistono situazioni per le quali esiste già una soluzione (più o meno consolidata) e l'obiettivo semplice che ci si pone è quello di migliorare - ottimizzare - la precedente soluzione. In questi casi, si pone un problema molto specifico diretto all'**ottimizzazione** della precedente soluzione e per il quale si giunge di norma ad una soluzione empirica.

Dall'insieme delle analisi matematiche effettuate sui dati, dalle informazioni che di volta in volta emergono e dalla verifica delle ipotesi e degli obiettivi è possibile riformulare nuove ipotesi e di individuare nuovi obiettivi.

1.5 - Le applicazioni

L'analisi dei lavori scientifici di carattere chemiometrico di quest'ultimo decennio mostra con chiarezza che a partire dal 1990 molti lavori di carattere applicativo non sono più solo degli stessi autori che hanno proposto e sviluppato i metodi chemiometrici durante il decennio 80-90. Infatti di molti lavori sono autori ricercatori che utilizzano tecniche chemiometriche con fini applicativi nei campi più diversi. E' questo il segno che la chemiometria ha superato la prima e più delicata fase di costituirsi come disciplina e di trovare una sua precisa identità scientifica per dare inizio ad una fase applicativa i cui sviluppi sembrano estremamente incoraggianti.
Accanto ad applicazioni tradizionali (cioè di carattere più strettamente chimico) come quelle nei campi della calibrazione, dell'ottimizzazione, dell'analisi del segnale e della risoluzione, della ricerca delle relazioni tra struttura chimica ed attività (biologica, farmacologica, tossicologica) e tra struttura chimica e proprietà, del riconoscimento di modelli, dell'interpretazione di dati spettroscopici e gascromatografici, troviamo oggi un vasto panorama di applicazioni, non solo di carattere chimico, sempre più connesse al mondo industriale.
Tra le numerosissime applicazioni di questi ultimi anni, vogliamo qui segnalarne alcune particolarmente significative per dare un quadro più chiaro delle potenzialità dei metodi chemiometrici.
Metodi chemiometrici sono stati utilizzati in campo clinico nello studio delle prestazioni di test antitumorali e nella caratterizzazione, ad esempio, di patologie cardiache, tiroidee, epatiche. In campo industriale diversi sono i lavori che riguardano l'ottimizzazione del processo produttivo e del prodotto, l'analisi di materiali, la diagnostica dei processi industriali, il controllo di qualità multivariata.
Studi importanti sono stati fatti in campo geologico e minerario, in particolare con analisi esplorative di dati geochimici e geofisici, sulla caratterizzazione dei suoli, sull'individuazione di sorgenti petrolifere e giacimenti minerari.
Numerosi cominciano ad apparire gli studi in campo ambientale, anche attraverso l'applicazione di metodi chemiometrici all'analisi di immagini computerizzate; studi importanti sono stati effettuati sulla qualità delle acque e dell'aria in molte regioni, sul problema dell'individuazione delle sorgenti di

inquinamento, su problemi di ecotossicità e di ottimizzazione di processi di combustione di rifiuti solidi urbani. Vi sono anche numerose applicazioni particolari, che danno una misura della flessibilità ed adattabilità di queste tecniche, quali, ad esempio, studi sull'ottimizzazione di serie di multisensori, studi sul comportamento animale, studi sulla tipicizzazione di prodotti alimentari.

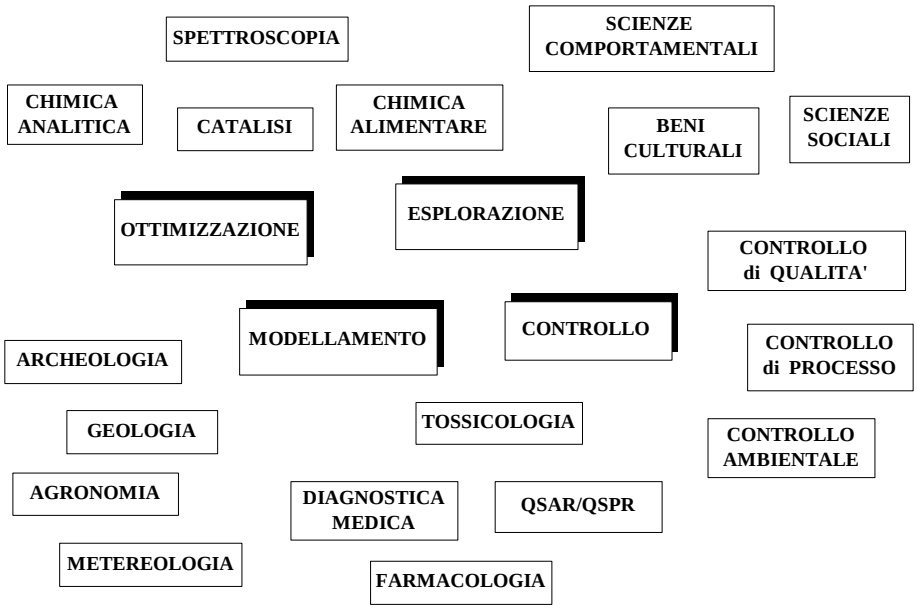


FIG. 1-5

Negli schemi seguenti sono riportate alcune applicazioni chemiometriche ormai consolidate.

❒ *Studio ed ottimizzazione di prodotti*

alimenti	vernici	cosmetici
aromi	combustibili	oli lubrificanti
farmaci	materie plastiche	materiali compositi
detersivi	additivi	miscele di reazione
emulsioni	leghe	tessuti

❒ *Ottimizzazione di impianti e di processi*

impianti produttivi	fermentatori
reattori chimici	impianti di purificazione

❒ *Sistemi di controllo e modellamento*

controllo di qualità	diagnostica clinica
sorgenti di inquinamento	modelli di classificazione
modelli di calibrazione	procedure analitiche
biosensori	sintesi organiche

1.6 - Il problema della formazione

Molte industrie europee e americane hanno già adottato questo tipo di metodologie, proprio tenendo conto dell'elevato rapporto tra benefici e costi di investimento: il notevole interesse economico e pratico legato all'applicazione di queste tecniche ha fatto sì che le grandi industrie abbiano preceduto centri di ricerca ed università nell'utilizzo di queste tecniche. Soltanto in questi ultimi anni diverse università hanno cominciato ad interessarsi in modo sistematico di queste tecniche, sviluppandone anche molti aspetti teorici nuovi.
Un piccolo centro per studi chemiometrici richiede un buon personal computer, qualche pacchetto di software e una piccola biblioteca con alcuni testi fondamentali e qualche rivista. Il costo più elevato rimane quello relativo alla formazione di personale specializzato.
A questo scopo, in Europa, un gruppo di chemiometri (universitari e non) organizza da alcuni anni una rete di scuole di chemiometria (*Progetto Eurochemometrics*) sia di carattere generale sia rivolte ad alcuni settori

particolari con obiettivi di formazione. Si sono create in questo modo le premesse non solo per affrontare il problema della formazione di base e specialistica, ma anche per realizzare un continuo scambio di informazioni e problemi tra il mondo accademico e quello produttivo, che ha già portato allo sviluppo di nuove metodologie estremamente interessanti.
Con queste prospettive emerge con chiarezza una **nuova figura professionale** che, coadiuvata nelle competenze specifiche di settore, sia capace di proporre nuove strategie e metodologie per la ricerca e il controllo di qualità in grado di far fronte alla rapidità dei cambiamenti e alle necessità di una continua innovazione. Un'ulteriore funzione importante dell'esperto chemiometra riguarda la possibilità di rivalutare, nel campo della produzione industriale, le grandi banche dati delle aziende estraendo da esse informazioni mirate a nuovi obiettivi o, nel campo sociale, di analizzare in modo efficace le banche dati degli enti pubblici ai fini del controllo della salute e del controllo ambientale.

2

LA STRUTTURA MULTIVARIATA DEI DATI

2.1 - Introduzione

Le tecniche chemiometriche si applicano normalmente a strutture di dati rappresentate da una tabella di numeri (la matrice dei dati), costituita da un certo numero di osservazioni, ciascuna delle quali è rappresentata da variabili che descrivono le osservazioni.
Un insieme di dati multivariato può essere quindi rappresentato secondo lo schema di Fig. 2-1:

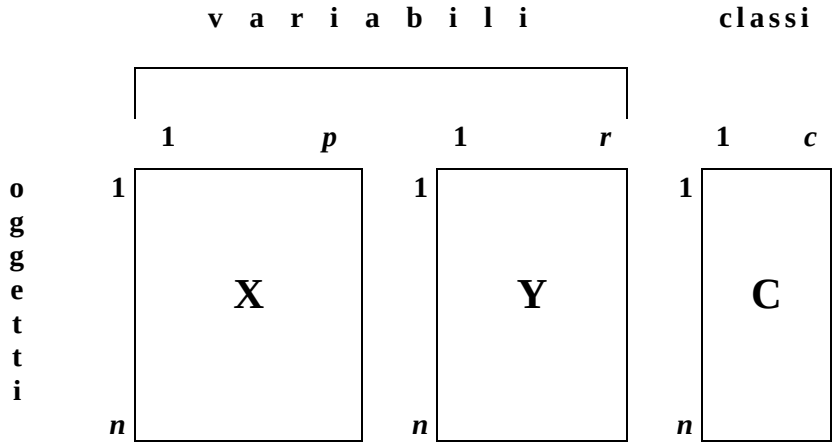


FIG. 2-1

Una comune tabella di dati è quindi rappresentabile con una matrice, le cui ***n*** righe rappresentano gli **oggetti** (**campioni, esperimenti, ecc.**) e le cui ***p*** colonne rappresentano le **variabili** (**descrittori**, **caratteri**, ecc.) con cui ciascun oggetto viene descritto.
Le variabili possono essere distinte in due gruppi logici: il blocco **X** dei **predittori** (le variabili indipendenti) e il blocco **Y** delle **risposte** (le variabili dipendenti). Naturalmente, a seconda del problema considerato, una variabile può appartenere una volta al blocco **X** e un'altra volta al blocco **Y**.
In base ad un criterio predefinito, agli oggetti può essere associato anche un vettore che contiene l'informazione dell'appartenenza di ciascun oggetto ad una delle ***G* classi** (**categorie, gruppi**) predefinite. Questo vettore è rappresentato da numeri interi, compresi tra 1 e *G*, che designano le classi, dove *G* è il numero totale delle classi: ad ogni oggetto corrisponde un numero che identifica la classe di appartenenza. Naturalmente, è possibile avere più criteri di classificazione (***c***): in questo caso, anche **C** diviene una matrice ove ciascuna colonna rappresenta un diverso criterio di classificazione.

Nota. Nel testo, se non diversamente indicato, vengono utilizzati i seguenti simboli:

Tipo	*numero totale*	*indice*	*altri indici*
oggetti	*n*	i	s, t
variabili in generale	*p*	j	k
variabili risposta	*r*	j	-
classi	*G*	g	g'
criteri	*c*	k	-
componenti principali	*M*	m	m'

Nella tabella sono riportati anche i simboli che verranno successivamente utilizzati per le componenti principali.

Una tabella di dati assume quindi la forma di una matrice **X** del tipo:

$$\mathbf{X} = \begin{vmatrix} x_{11} & x_{12} & \dots & \dots & x_{1p} \\ x_{21} & x_{22} & & & \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & & & & x_{np} \end{vmatrix}$$

dove il singolo dato x_{ij} si trova nella posizione corrispondente alla *i-esima* riga (*i*-esimo campione) e alla *j-esima* colonna (*j-esima* variabile) della matrice.
I metodi sono presentati di norma per studiare matrici **X**, ove i dati sono rappresentati da *n* osservazioni (le righe della matrice), ciascuna descritta da *p* variabili. Questa modalità di analisi dei dati viene chiamata **modo R** (***R-mode*** o ***R-analysis*****)**. Ad esempio, in questa modalità le matrici di covarianza o di correlazione di **X** sono matrici di dimensione *p* x *p*.
In molti casi, tuttavia, è interessante scambiare i ruoli di oggetti e variabili, studiando il problema inverso utilizzando gli stessi metodi: in questo caso le variabili sono *matematicamente trattate* come oggetti e gli oggetti come la descrizione delle variabili. Matematicamente questo è possibile applicando gli stessi metodi sulla *matrice trasposta* di **X** ($\mathbf{X}^{\mathbf{T}}$, di dimensione *p* x *n*). Questa modalità di analisi dei dati viene chiamata **modo Q** (***Q-mode*** o ***Q-analysis*****)**. In questo caso, ad esempio, le matrici di covarianza o di correlazione sono di dimensione *n* x *n*.

🗀 Le variabili

Le variabili sono le grandezze che utilizziamo per studiare un dato fenomeno e per descrivere complessivamente le osservazioni e possono essere sperimentali o calcolate per via teorica.
Le variabili rappresentano quindi il modo in cui si descrive un sistema relazionale empirico e le **scale di misura** rappresentano il modo con cui l'informazione empirica viene trasformata in informazione numerica.
Scopo fondamentale di una scala di misura è quello di convertire informazione empirica (sperimentale) in forma numerica nel modo più conveniente al fine di facilitare il calcolo e l'interpretazione dei risultati.

Per un dato sistema relazionale empirico, si richiede che una scala numerica preservi una o più delle sue caratteristiche, quali ad esempio l'ordinamento delle misure ottenute, l'ordinamento delle loro differenze, ecc.
Se una data scala di misura rappresenta le caratteristiche proprie del sistema relazionale empirico, qualsiasi altra scala di misura in grado di preservare le relazioni tra entità empiriche e numeriche è permessa.
Esistono 6 scale di misura fondamentali utilizzate nelle scienze naturali e sociali, basate su una trasformazione in grado di preservare sia l'ordine che gli intervalli:

$$y = a + bx \quad (b > 0)$$

- *scala assoluta*

Una scala assoluta è una scala definita in modo univoco, poichè è definibile un'unica unità di misura. Se x è una scala di misura, l'unica trasformazione ammissibile è $y = x$. Un esempio di questo tipo di scala è il conteggio del numero di elementi di un insieme.

- *scala di rapporti*

Utilizzata soprattutto per variabili continue come lunghezze, pesi, età, intervalli di tempo, è la scala di misura più comune. Nota anche come scala quantitativa, ha un origine naturale nello zero, un ordine e una gradazione di differenze.
Una scala di rapporti non è unica in quanto può essere dilatata o compressa arbitrariamente mediante un fattore di proporzionalità, come tipicamente le conversioni di unità da metri a centimetri (dilatazione) o da grammi a chilogrammi (compressione). L'unica trasformazione ammessa è $y = bx$ ($b>0$), cioè se x è una misura in una scala di rapporti, anche y è una misura in una scala di rapporti.

- *scala di differenze*

In questa scala, non comune ma comunque utilizzata, viene rimossa l'unicità dell'origine della scala, tipica della scala di rapporti (lo zero). In questa scala l'origine può essere un qualsiasi numero reale, ma le differenze rimangono invariate poichè si assume un'unica unità di misura ($b = 1$). L'unica trasformazione ammissibile è $y = a + x$. Un esempio di questo tipo di scale sono i calendari religiosi basati su differenti riferimenti temporali: l'unità di misura (l'anno) è la stessa, ma l'origine è differente.

- *scala di intervalli*

Scale di carattere ancora più generale sono le scale di intervalli, di cui le scale di differenze sono un caso particolare. Scale tipiche per le misure di temperatura, l'unico confronto significativo ammesso riguarda gli intervalli e la trasformazione ammissibile è $y = a + bx \quad (b > 0)$.

- *scala ordinale*

Questo tipo di scale non possiede un origine naturale e la distanza tra i punti della scala non è definita. Viene invece preservato l'ordinamento tra gli elementi del sistema empirico relazionale, sono cioè ammesse sono relazioni di ordine (inferiore a, maggiore di).
Se x è una scala ordinale e la funzione $f(x)$ è una funzione continua monotona strettamente crescente di x, allora anche $y = f(x)$ è una scala ordinale. Normalmente, nella pratica, queste scale sono scale discrete (seuqnze di interi) piuttosto che continue. Classifiche, ordinamente, giudizi e preferenze sono in generale variabli misurate su scale ordinali. Scale di questo tipo possono essere utilizzate al posto di scale di rapporti quando le misure continue sono affette da un significativo errore sperimentale.

- *scala nominale*

Le scale nominali sono necessariamente scale discrete rappresentate da etichette che indicano il gruppo di appartenenza di ciascun elemento dell'insieme relazionale empirico. Queste scale sono note anche come scale categoriche o qualitative, caratterizzabili esclusivamente dalle relazioni di uguaglianza e disuguaglianza (equivalenza).

In particolare, in molti problemi, è del massimo interesse studiare esplicitamente le variabili considerate nel descrivere il sistema. Studiare le variabili significa analizzarne la loro rilevanza per il problema, comprendere le correlazioni che sussistono tra di esse, evidenziare l'esistenza di gruppi di variabili simili, cioè di variabili che portano la stessa informazione o, viceversa, di variabili uniche nella descrizione multivariata di una parte del problema.
Si deve osservare che quando si parla di *variabili che portano la stessa informazione*, non si afferma necessariamente che queste variabili abbiano lo stesso significato o addirittura rappresentino la stessa cosa. Si afferma, invece, che nell'ambito del campione studiato, queste variabili, eventualmente anche con significati molto diversi tra loro, hanno lo stesso ruolo e la stessa capacità di rappresentare numericamente una parte del sistema analizzato.

🗀 Gli oggetti

Gli oggetti rappresentano gli esempi o i campioni che abbiamo a nostra disposizione per capire il fenomeno studiato, per costruire modelli matematici, per confermare le ipotesi formulate.
Un campione può essere descritto da un'unica misura (sperimentale) oppure da diverse misure (sperimentali): in quest'ultimo caso gli oggetti si definiscono *multivariati*. L'insieme delle misure effettuate su un campione è rappresentato dalle variabili selezionate per *descrivere* l'oggetto: l'insieme di valori che lo definisce costituisce il *dato*.
Per quanto il termine *misura* abbia nell'accezione più comune un richiamo sperimentale, in molti casi è di grande interesse riuscire a descrivere i campioni per via teorica, cioè con variabili calcolate sulla base di una rappresentazione teorica dei campioni. Ad esempio, questo aspetto è uno degli elementi caratterizzanti molte ricerche *QSAR* (v. Capitoli 12 e 13), dove la ricerca di modelli in grado di predire proprietà sperimentali di composti chimici può essere effettuata mediante una descrizione teorica delle strutture molecolari.
Un oggetto può essere descritto in modo completo se, per tutte le variabili selezionate per descriverlo, sono disponibili i valori ad esso relativi. In altri casi, come accade frequentemente, non sono disponibili i dati per tutte le variabili selezionate: in questo caso l'oggetto è descritto in modo incompleto e si dice che esistono dei valori mancanti (*missing values*).

🗀 Le classi

In molti casi i campioni disponibili non sono omogenei, provengono cioè a popolazioni diverse ovvero appartengono a classi o categorie differenti. Per tener conto di questa caratteristica di un sistema è necessario anche considerare la classe di appartenenza di ciascun campione. Ad esempio, un insieme di composti può essere costituito da una classe di composti tossici e da un'altra di composti non tossici. Le classi sono quindi l'espressione della presenza nella descrizione del sistema di variabili nominali o ordinali.
Quando si ricercano modelli di classificazione (v. Cap. 6), nei casi più comuni, il numero delle classi è abbastanza piccolo (da 2 a 5-6), ma non ci sono evidentemente limitazioni di principio al loro numero. L'eventuale limite è più di ordine pratico, in quanto la descrizione delle caratteristiche di ciascuna classe richiede la stima statistica dei parametri di ogni classe e quindi - di fatto - la presenza di un significativo numero di oggetti in grado di rappresentare adeguatamente ciascuna classe.

2.2 - Il pretrattamento dei dati

Prima di effettuare qualsiasi tipo di elaborazione è sempre necessaria un'analisi preliminare dei dati, analisi che ha lo scopo di controllare e predisporre i dati per le successive elaborazioni.
Nel controllo della **correttezza dei dati**, la prima fase consiste nel verificare che non vi siano errori evidenti di trascrizione dei dati, e la presenza di eventuali **dati mancanti** (*missing values*), opportunamente rappresentati da un codice numerico univoco (ad esempio, -999), scelto in modo tale che il codice non possa mai confondersi con un valore che teoricamente una variabile può assumere (non utilizzare mai lo zero come codice per i dati mancanti). Inoltre è necessario verificare che non vi sia qualche **variabile costante** (!), cioè i cui valori siano tutti uguali per gli oggetti considerati. In questo caso è necessario escludere la variabile da qualsiasi elaborazione successiva. Un controllo più generale riguarda la verifica del **tipo di variabili** che descrivono il sistema di dati: infatti, la presenza di variabili non continue, quali ad esempio le variabili binarie (che prendono solo valori 0 e 1), di variabili discrete ordinali (che possono prendere, ad esempio, valori 1, 2 ,3 e 4), o di variabili reali ma altamente degeneri (cioè che assumono pochi valori diversi per tutti i campioni considerati) può comportare alcuni problemi in quanto non tutti i metodi e gli algoritmi sono adatti a trattare variabili di questo tipo o riescono a fornire informazioni utili in queste condizioni.
In molti casi una o più variabili presentano caratteristiche statistiche distribuzionali particolari (non normalità e asimmetria accentuate), motivo per cui è opportuno operare **trasformazioni delle variabili** al fine di eliminare o attenuare le caratteristiche non gradite. Tuttavia, una volta effettuate le singole trasformazioni su una o più variabili, non siamo ancora sicuri che in un trattamento multivariato, dove tutte le variabili vengono considerate contemporaneamente, esse siano realmente confrontabili tra loro, in accordo con gli obiettivi stabiliti e i metodi utilizzati per conseguirli. In questi casi, al fine di rendere tra loro omogenee le variabili considerate, è spesso opportuno effettuare **scalature delle variabili** agendo con la stessa procedura ma separatamente su ciascuna delle variabili.
In generale, quindi, il pretrattamento dei dati è un momento delicato nella trattazione del problema in quanto 1) influenza i risultati delle elaborazioni successive, 2) ogni trasformazione o scalatura modifica in qualche modo l'informazione contenuta nei dati, 3) la non rilevata presenza di dati anomali o di errori può portare a modelli e conclusioni devianti e non generali.

2.3 - La "presenza" di dati mancanti

In moltissimi casi accade che non per tutti i campioni siano disponibili i valori corrispondenti a tutte le variabili che li descrivono. Questo aspetto poco gradevole dei dati ci obbliga a prendere alcune decisioni in merito, in quanto nessuno tra i comuni metodi matematici è in grado di trattare problemi ove manchino alcuni valori nella matrice dei dati.
Per la soluzione di questo problema esistono diverse possibilità che riportiamo qui di seguito.

🗀 Eliminazione dei campioni

Il metodo più semplice ricorre all'eliminazione dei campioni per i quali sono presenti dati mancanti. Questa è una buona soluzione solo se il numero dei campioni è elevato e quindi l'eliminazione di qualche campione non comporta la perdità di informazione rilevante. D'altra parte, questa soluzione comporta la completa eliminazione di un dato (e quindi di tutta l'informazione ad esso legata) anche nel caso in cui per un campione siano noti i valori di tutte le variabili meno uno.

🗀 Eliminazione delle variabili

Un'alternativa al caso precedente è l'eliminazione di una o più variabili per le quali vi siano numerosi dati mancanti, conservando quindi solo quelle variabili per le quali sono disponibili tutti i valori. E' evidente che in questo caso si viene a perdere completamente l'eventuale informazione legata alle variabili eliminate.

🗀 Sostituzione con il valor medio

Qualora i dati mancanti non siano troppo numerosi, è possibile sostituire il valore mancante con il valor medio, calcolato su tutti i dati restanti, della variabile per cui manca il dato. In questo caso, si attenua l'informazione utile presente nella variabile per la quale diversi campioni sono rappresentati semplicemente dal valor medio. Ciò comporta una sottostima delle varianze ed una sovrastima delle covarianze.

🗀 Sostituzione con un valore casuale

In questo caso i dati mancanti vengono sostituiti da un numero casuale estratto uniformemente nell'intervallo della variabile, calcolato su tutti i dati restanti. E' questa una soluzione analoga alla precedente, ma in questo caso si aggiunge rumore (informazione spuria) ai dati. In questo caso si ottiene l'effetto opposto al caso precedente, con una sovrastima delle varianze ed una sottostima delle covarianze.

Nota. I tre metodi successivi per la stima dei valori mancanti sono basati su tecniche che verranno esposte successivamente nel testo. Tuttavia, per motivi di completezza, si è ritenuto opportuno presentarli qui.

🗀 Sostituzione mediante regressione

In questo caso i valori mancanti per ciascuna variabile vengono predetti utilizzando un modello di regressione (v. Cap. 7) ricavato da tutti i campioni senza valori mancanti, come descrittori le variabili restanti e come risposta la variabile per la quale si vogliono predire i valori mancanti: dal modello ottenuto vengono stimati i valori mancanti per la variabile considerata come risposta. Condizione necessaria perchè questo metodo sia utilizzabile è che la qualità del modello ottenuto sia accettabile. Un limite ulteriore di questo metodo è che i valori calcolati possono essere valori estrapolati (soprattutto per campioni situati ai limiti dello spazio sperimentale) e di conseguenza valori non accettabili almeno in linea teorica (ad esempio, si possono ottenere concentrazioni negative).

🗀 Sostituzione mediante similarità locale

Questo metodo per il calcolo dei valori mancanti è basato sul metodo K-NN (*k-nearest neighbours*, v. Cap. 6, metodi di classificazione). In questo caso, gli oggetti per i quali non esistono dati mancanti vengono utilizzati per stimare la distanza da ciascun oggetto per il quale esiste un dato mancante. La distanza di ciascun oggetto da quello per cui vi è un dato mancante viene calcolata utilizzando tutte le altre variabili definite per esso definite.
Scelto un valore del parametro *k*, cioè il numero dei "*k* oggetti più vicini" da considerare, viene stimato il valore dell'*i*-esimo dato mancante per la *j*-esima

variabile dai valori assunti per la *j*-esima variabile dai *k* oggetti, pesando i valori con l'inverso delle loro distanze dall'*i*-esimo oggetto, secondo l'espressione:

$$\hat{x}_{ij} = \frac{\sum_r x_{i_r j} \Big/ d_{ii_r}}{\sum_r 1 \Big/ d_{ii_r}}$$

dove *r* scorre sui *k* intorni, i_r identifica ciascuno dei *k* oggetti più vicini e *d* è la distanza euclidea media tra il dato *i* e ciascuno dei *k* oggetti più vicini. In pratica, l'oggetto più vicino pesa di più nel determinare la stima del dato mancante.
In questo caso il valore stimato non viene mai estrapolato, evitando i problemi del metodo precedente. Tuttavia, se il campione per il quale si desidera stimare il dato mancante è molto "lontano" dai *k* campioni più vicini, il valore stimato può essere molto approssimato.

🗀 Sostituzione mediante l'analisi delle componenti principali

L'analisi delle componenti principali (v. Cap. 3) può essere utilizzata per stimare i dati mancanti. In questo caso l'analisi delle componenti principali viene effettuata utilizzando tutti i valori noti mediante l'algoritmo *NIPALS* che consente di calcolare le componenti principali anche con matrici incomplete. Una volta stimato il numero *M* di componenti significative, è possibile ricostruire i dati originali mediante il prodotto della matrice degli *scores* per la matrice dei *loadings*.

Esempio

In Tab.2-1 sono riportati i dati relativi a 20 campioni rappresentati da 4 variabili. Per il campione 4 il dato relativo alla quarta variabile è mancante; per il campione 15 è mancante il dato relativo alla prima variabile (valori -999).
In questo caso la soluzione di eliminare le variabili 1 e 4, ove esistono dati mancanti è senz'altro improponibile in quanto i dati rimarrebbero rappresentati da sole due variabili. Anche l'eliminazione dei due dati (4 e 15) non appare incoraggiante in quanto significherebbe privarsi del 10 % dei dati totali.
In Tab.2-2 sono riportati i valori calcolati per i due dati mancanti utilizzando i diversi metodi. In particolare, la regressione può essere effettuata utilizzando una delle due variabili prive di dati mancanti (la seconda o la terza): da ciascun

modello calcolato viene predetto il valore dei dati mancanti della variabile risposta considerata.

ID	*X1*	*X2*	*X3*	*X4*
1	1.48	10.00	2.134	43
2	2.17	8.59	2.653	9
3	2.50	8.95	2.647	33
4	2.35	9.28	2.647	-999
5	3.19	6.54	3.171	57
6	3.08	7.77	3.165	45
7	3.20	6.44	3.165	59
8	2.86	6.89	3.171	68
9	3.37	7.48	3.165	68
10	3.52	7.02	3.160	68
11	4.07	7.75	3.685	84
12	4.21	7.46	3.684	67
13	3.88	7.23	3.684	58
14	3.72	7.53	3.685	62
15	-999	7.52	3.685	70
16	4.39	8.35	3.684	101
17	4.95	7.14	4.204	116
18	4.10	5.70	4.204	70
19	4.90	5.58	4.204	115
20	5.01	5.17	4.733	191

TAB. 2-1

L'analisi delle componenti principali viene utilizzata per predire i dati mancanti mediante il calcolo dei 4 possibili modelli (da 1 a 4 componenti), ricorrendo all'algoritmo *NIPALS*, in grado di trattare anche oggetti incompleti. Il calcolo del dato mancante vienne effettuato dal prodotto degli *scores* corrispondenti all'oggetto considerato per la matrice dei *loadings* (v. Cap. 3).

Metodo	*Parametri*	*Valore X1(15)*	*Valore X4(4)*
Media	-	3.52	72.84
Regressione	var. X2	3.46	32.88
Regressione	var. X3	3.92	31.93
PCA	1 PC	3.60	37.20
PCA	2 PC	3.69	45.99
PCA	3 PC	3.75	49.46
PCA	4 PC	3.65	49.39
K-NN	k = 1	4.21	33.00
K-NN	k = 2	4.04	26.43
K-NN	k = 3	4.07	24.26
K-NN	k = 5	3.99	31.16
K-NN	k = 7	3.90	34.90

TAB. 2-2

Infine, il metodo K-NN viene utilizzato con un diverso numero di intorni *k* (1, 2, 3, 5, 7) mediante i quali stimare i valori mancanti.

2.4 - Le trasformazioni delle variabili

In diversi casi le variabili presentano comportamenti indesiderati come quelli di non-normalità, non-additività, non-linearità rispetto ad altre variabili, eteroscedasticità (cioè la varianza campionaria non è distribuita uniformemente, opposto a omoscedasticità). Le trasformazioni vengono quindi effettuate per controbilanciare violazioni delle assunzioni richieste da molti metodi statistici quando si cerca di costruire modelli su dati sperimentali.
In generale, le **trasformazioni di una variabile** rispondono quindi ad uno dei seguenti scopi:

- stabilizzare la varianza
- linearizzare le relazioni tra le variabili
- normalizzare la distribuzione
- ottenere additività
- realizzare modelli più robusti

Riportiamo qui di seguito alcune delle trasformazioni più interessanti. Le formule sono definite per una generica variabile *x*, ma è chiaro che queste

trasformazioni si applicano anche alle variabili risposta, normalmente designate con *y*.

🗀 Trasformazione logaritmica

La trasformazione logaritmica

$$x'_{ij} = log\left(x_{ij}\right) \quad \text{o, meglio,} \quad x'_{ij} = log\left(1 + x_{ij}\right)$$

viene generalmente utilizzata allo scopo di linearizzare il comportamento di variabili il cui effetto è moltiplicativo. Questa stessa trasformazione può venire utilizzata anche in condizioni di eteroscedasticità e quando le deviazioni standard sono proporzionali alle medie (cioè, in condizioni di coefficiente di variazione costante).

🗀 Trasformazione arcoseno

La trasformazione arcoseno

$$x'_{ij} = arcsin\left(\sqrt{x_{ij}}\right)$$

viene generalmente utilizzata per normalizzare distribuzioni di tipo binomiale. Queste distribuzioni sono tipiche di dati espressi come percentuale o proporzione, dove possono esistere larghi scostamenti dalla normalità per piccole o grandi percentuali, cioè agli estremi della distribuzione.

🗀 Trasformazione radice quadrata

La trasformazione radice quadrata

$$x'_{ij} = \sqrt{x_{ij} + 0.5} \quad \text{oppure} \quad x'_{ij} = \sqrt{x_{ij} + \frac{3}{8}}$$

è applicabile quando si desidera normalizzare dati provenienti da distribuzioni poissoniane, ove le varianze sono proporzionali alle medie. Questo accade comunemente quando le variabili esprimono dei conteggi, come si verifica per molte variabili di tipo biologico.

🗀 Trasformazione inversa

La trasformazione inversa

$$x'_{ij} = \frac{1}{x_{ij}} \quad \text{oppure} \quad x'_{ij} = \frac{1}{1 + x_{ij}}$$

viene utilizzata, seppur più raramente, per normalizzare distribuzioni le cui deviazioni standard sono proporzionali al quadrato delle medie.

🗀 Trasformazione tangente iperbolica

Questa trasformazione ha un forte effetto di scedasticità, cioè un marcato effetto di stabilizzazione della varianza.

$$x'_{ij} = tanh^{-1}\left(x_{ij}\right)$$

🗀 Trasformazioni di potenza

Le trasformazioni di potenza costituiscono un'importante famiglia di trasformazioni, di cui alcuni casi precedenti sono casi particolari. Questa famiglia di trasformazioni è definita come

$$x'_{ij}(\lambda) = x_{ij}^{\lambda} \qquad -\infty < \lambda < +\infty$$

Sono casi particolari di questa famiglia di trasformazioni, la trasformazione inversa ($\lambda = -1$) e la trasformazione radice quadrata ($\lambda = 0.5$).
Alcuni grafici caratteristici sono riportati in Fig. 2-2.

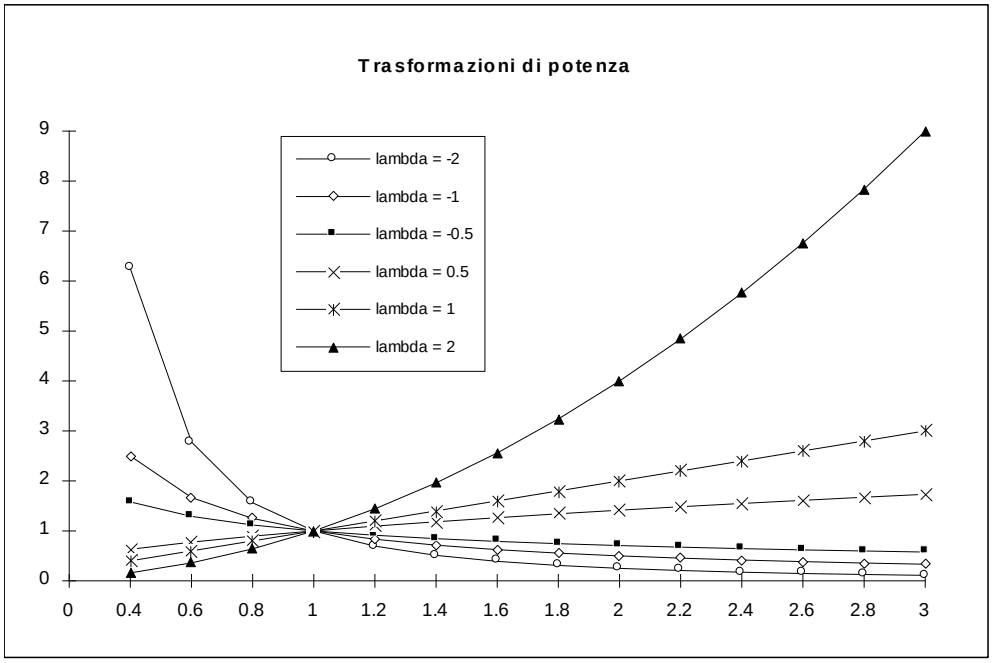


FIG. 2-2

🗀 Trasformazioni di Box-Cox

La trasformazione di Box-Cox riguarda soprattutto le trasformazioni delle risposte al fine di migliorare la loro relazione con le variabili predittrici ed è definita come:

$$y_i'(\lambda) = \begin{cases} \dfrac{y_j^{\lambda} - 1}{\lambda} & se \quad \lambda \neq 0 \\ ln(y_j) & se \quad \lambda = 0 \end{cases}$$

Questo tipo di trasformazioni richiede che la risposta sia definita positiva ($y_i > 0$). In caso contrario, è necessario semplicemente aggiungere una costante additiva alla risposte al fine di renderle tutte positive. Il termine costante -1

consente di preservare la continuità anche rispetto alla funzione logaritmica ($\lambda = 0$).
Le trasformazioni di Box-Cox costituiscono quindi una *famiglia di trasformazioni* di potenza monotone dipendenti dal parametro λ; le trasformazioni sono convesse per $\lambda > 1$ e concave per $\lambda < 1$. Casi particolari di questa famiglia di trasformazioni sono la trasformazione logaritmica $\lambda = 0$), la trasformazione inversa ($\lambda = -1$) e la trasformazione radice quadrata ($\lambda = 0.5$), la trasformazione con l'inverso della radice quadrata ($\lambda = -0.5$). Per $\lambda = 1$, non si ha nessuna trasformazione (a meno di una costante). Alcune curve caratteristiche sono riportate in Fig. 2-3.

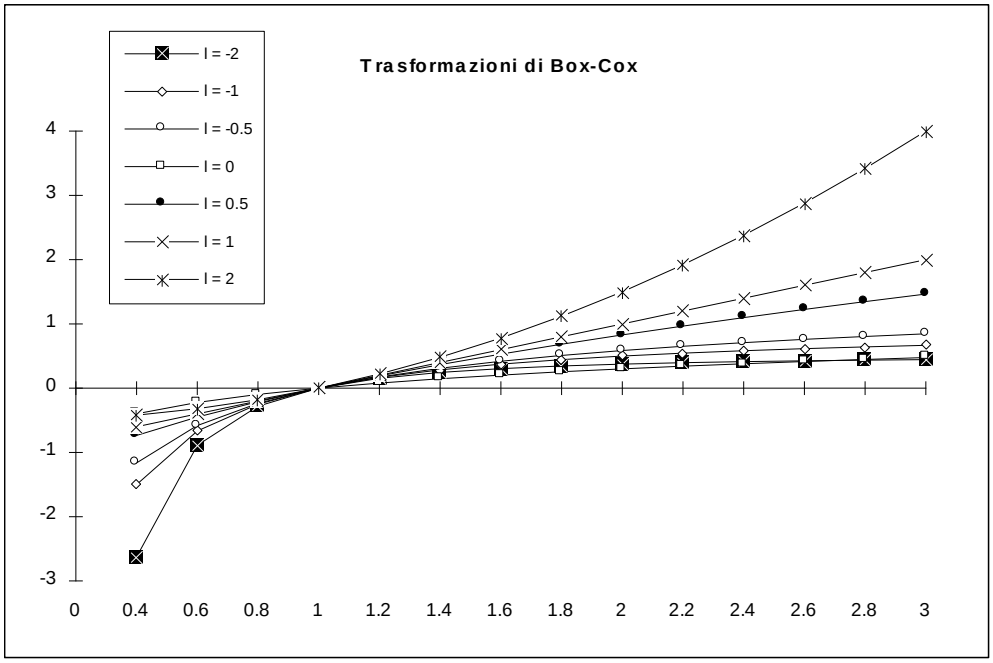


FIG. 2-3

Questa famiglia di trasformazioni può venire utilizzata per cercare la migliore trasformazione di una risposta (cioè determinare il valore del parametro λ) per ottimizzare un modello di regressione, cioè rendere minima la somma dei quadrati degli scarti (*RSS*) (v. Cap.7, metodi di regressione: le basi).

2.5 - Le scalature dei dati (o standardizzazioni, *scaling*)

In molti casi, i metodi chemiometrici richiedono che i dati originali siano pretrattati al fine di poter enucleare dai dati informazione utile e non banale ed eliminare informazione ovvia (cioè intrinseca nella singola variabile) e non interessante. In particolare, quando si ricercano informazioni sulle *relazioni tra le variabili*, obiettivo tipico dell'analisi multivariata, è importante *massimizzare la confrontabilità* tra le variabili.
Ad esempio, metodi chemiometrici fondamentali quali l'analisi delle componenti principali e tutti i metodi basati sul calcolo di distanze (metodi di *cluster analysis*) non sono invarianti alle scalature e richiedono una decisione a priori se e quale tipo di scalatura utilizzare. Si può infatti legittimamente ritenere che se lo scopo dell'analisi è quello di osservare i dati così come si presentano, nessuna scalatura dei dati deve essere preventivamente effettuata. Tuttavia, vi sono due problemi fondamentali che sono potenzialmente presenti in questa impostazione. In primo luogo, se tutte le variabili misurate rappresentano quantità molto diverse tra loro, sia le loro varianze sia le differenze tra coppie di oggetti per ciascuna variabile possono essere molto differenti tra loro. Alternativamente, anche se le variabili rappresentano tutte lo stesso tipo di quantità, lo stesse differenze possono essere riscontrate quando le varianze sono ancora molto diverse tra loro o vengono utilizzate diverse scale di misura. In entrambi i casi, le variabili che esibiscono una grande varianza (o grandi differenze tra le coppie di oggetti) risulteranno decisive nel determinare le principali sorgenti di informazione nell'analisi delle componenti principali o le distanze tra le coppie di oggetti nei metodi di *cluster analysis*. Ciò significa che il risultato dell'analisi e la sua interpretazione saranno condizionati da poche variabili che mascherano eventuali contributi di altre variabili solo in virtù del loro intrinseco significato, della loro alta varianza o della loro scala di misura.
In questi casi è quindi necessario effettuare un pretrattamento delle variabili al fine di evitare che i risultati dell'analisi multivariata siano influenzati da questi aspetti indesiderati. Al contrario, le scalature potranno essere evitate solo in quei casi in cui tutte le variabili rappresentano la stessa quantità, le loro scale di misura sono le stesse e le loro varianze sono almeno confrontabili.
Nelle trasformazioni delle variabili viste in precedenza, ogni valore trasformato dell'*i-esimo* dato dipende esclusivamente dal corrispondente valore originale del dato e dal tipo di trasformazione utilizzata. Diversamente dalle trasformazioni, il valore scalato dell'*i*-esimo dato dipende dal corrispondente valore originale del dato, dalla scalatura utilizzata e da alcuni parametri che vengono stimati dai

valori assunti dai tutti gli oggetti per quella variabile (ad esempio, la media e la deviazione standard o il minimo e il massimo). Inoltre, le scalature vengono effettuate seguendo la stessa procedura (cioè lo stesso tipo di scalatura) *su tutte le variabili contemporaneamente* se pur indipendentemente l'una dall'altra.
Qui di seguito riportiamo i più comuni tipi di scalatura. Il termine *V*(x) indica la varianza della variabile.

🗀 Centratura (*centering*)

Questo tipo di scalatura consiste nel centrare i dati rispetto al valor medio di ciascuna variabile (il centroide):

$$x'_{ij} = x_{ij} - \bar{x}_j$$

La proprietà fondamentale dei dati centrati è che il valor medio di ciascuna variabile è uguale a zero:

$$\bar{x}'_j = 0$$

Questa scalatura non modifica la varianza dei dati.

🗀 Scalatura rispetto al valor massimo (*maximum scaling*)

Questo tipo di scalatura pone un vincolo al valore massimo che ciascuna variabile può assumere, dividendo i dati relativi ad una variabile per il suo valor massimo U_j :

$$x'_{ij} = \frac{x_{ij}}{U_j}$$

dove $U_j = \max_i\left(x_{ij}\right)$. I dati così scalati hanno quindi la seguente proprietà:

$$\max_i\left(x'_{ij}\right) = 1$$

🗀 Scalatura di intervallo (*range scaling*)

Questo tipo di scalatura pone un duplice vincolo a ciascuna variabile: il valore minimo L_j e il suo valore massimo U_j sono rispettivamente uguali a zero e uno:

$$x'_{ij} = \frac{x_{ij} - L_j}{U_j - L_j}$$

dove $U_j = \max_i(x_{ij})$ e $L_j = \min_i(x_{ij})$.

Le proprietà dei dati così scalati sono quindi $\max_i(x'_{ij}) = 1$ e $\min_i(x'_{ij}) = 0$.

Nel caso in cui si desideri riscalare i dati in un intervallo, diverso da 0 e 1, definito per ciascuna variabile dai valori INF_j e SUP_j, si può utilizzare successivamente alla scalatura dei dati tra 0 e 1, una **scalatura di intervallo generalizzata**, data dalla seguente espressione:

$$x''_{ij} = (SUP_j - INF_j) \cdot x'_{ij} + INF_j$$

In questo caso le proprietà dei dati riscalati sono quindi $max_i(x''_{ij}) = SUP_j$ e $min_i(x''_{ij}) = INF_j$. Ad esempio, per riscalare i dati già scalati tra 0 e 1 nell'intervallo compreso tra -100 e +100, l'espressione diviene:

$$x''_{ij} = 200 \cdot x'_{ij} - 100$$

🗀 Scalatura a varianza unitaria (*unit variance scaling*)

Si tratta di un tipo di scalatura non molto utilizzato in chemiometria che trasforma la variabile in una variabile a varianza unitaria:

$$x'_{ij} = \frac{x_{ij}}{s_j}$$

dove s_j è la deviazione standard della *j-esima* variabile. La varianza della variabile normalizzata è quindi uguale ad 1:

$$V(\mathbf{x}'_j) = 1$$

🗀 Autoscalatura (*autoscaling*)

E' una delle scalature più utilizzate nei metodi chemiometrici e consiste in una centratura seguita da una normalizzazione a varianza unitaria:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Media e varianza delle variabili autoscalate sono: quindi 0 e 1, rispettivamente:

$$\bar{x}'_j = 0 \quad \text{e} \quad V(\mathbf{x}'_j) = 1$$

🗀 **Scalatura logaritmica (*logarithmic scaling*)**

E' una scalatura introdotta recentemente da Aitchison, particolarmente adatta per variabili che variano su molte unità di misura. Si tratta di una centratura logaritmica:

$$x'_{ij} = log\left(x_{ij}\right) - \frac{\sum_i log\left(x_{ij}\right)}{n}$$

In pratica, questa scalatura consiste nella comune centrature di una variabile *y*, una volta effettuata la trasformazione $y = log(x)$.

🗀 **Doppia centratura logaritmica (*logarithmic double centering*)**

E' questa una doppia centratura, cioè effettuata contemporaneamente sulle righe e sulle colonne, utilizzando una trasformazione iniziale logaritmica, seguita da una doppia centratura:

$$y_{ij} = log\left(x_{ij}\right)$$

$$x'_{ij} = y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}$$

🗀 **Profilo semplice (*row profile*)**

$$x'_{ij} = \frac{x_{ij}}{\sum_j x_{ij}}$$

La somma al denominatore può essere sostituita dal suo valor medio.

🗀 **Profilo semplice normalizzato (*normalized row profile*)**

$$x'_{ij} = \frac{x_{ij}}{\sum_j x_{ij}^2}$$

La somma al denominatore può essere sostituita dal suo valor medio.

🗀 **Profilo globale (*global profile*)**

$$x'_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij} \cdot \sum_j x_{ij}}}$$

La somma al denominatore può essere sostituita dal suo valor medio.

🗀 **Profilo globale normalizzato (*normalized global profile*)**

$$x'_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2 \cdot \sum_j x_{ij}^2}}$$

La somma al denominatore può essere sostituita dal suo valor medio.

Nota. Le scalature possono essere descritte da una trasformazione lineare del tipo:

$$x'_{ij} = \alpha \cdot x_{ij} + \beta$$

ove i parametri α e β sono definiti in Tab. 2-3:

Scalatura		α	β
Centratura	CS	1	$-\bar{x}_j$
Scalatura rispetto al massimo	MS	$1/U_j$	0
Scalatura di intervallo	RS	$1/(U_j - L_j)$	$-L_j/(U_j - L_j)$
Scalatura a varianza unitaria	US	$1/s_j$	0
Autoscalatura	AS	$1/s_j$	$-\bar{x}_j/s_j$
Profili	P	1/denominatore	0

TAB. 2-3

La **media** e la **deviazione standard di dati scalati** variano secondo le seguenti relazioni:

$$\bar{x}'_j = \alpha \cdot \bar{x}_j + \beta \qquad s'_j = \alpha \cdot s_j$$

Si può osservare, ad esempio, che la centratura dei dati non modifica la loro varianza ($\alpha = 1$).

Esempio

Consideriamo i seguenti 9 dati relativi ad una variabile X1. Alcune scalature di questa variabile e i parametri statistici fondamentali sono riportati in Tab. 2-4.
RS, MS, AS e US rappresentano rispettivamente la scalatura di intervallo, la scalatura rispetto al valore massimo, l'autoscalatura e la scalatura a varianza unitaria.
Consideriamo anche i altri 9 dati relativi ad una variabile X2. Alcune scalature di questa seconda variabile e i parametri statistici fondamentali sono riportati in Tab. 2-5.

ID	*X1*	*RS*	*MS*	*AS*	*US*
1	0	0	0	-1.155	0
2	2	.5	.5	0	1.155
3	4	1	1	1.155	2.309
4	0	0	0	-1.155	0
5	2	.5	.5	0	1.155
6	4	1	1	1.155	2.309
7	0	0	0	-1.155	0
8	2	.5	.5	0	1.155
9	4	1	1	1.155	2.309
media	2	0.5	0.5	0	1.155
dev.std.	1.732	0.433	0.433	1	1
min	0	0	0	-1.155	0
max	4	1	1	1.155	2.309

TAB. 2-4

ID	*X2*	*RS*	*MS*	*AS*	*US*
1	1	0	.333	-1.155	1.155
2	1	0	.333	-1.155	1.155
3	1	0	.333	-1.155	1.155
4	2	.5	.667	0	2.309
5	2	.5	.667	0	2.309
6	2	.5	.667	0	2.309
7	3	1	1	1.155	3.464
8	3	1	1	1.155	3.464
9	3	1	1	1.155	3.464
media	2	0.5	0.667	0	2.309
dev.std.	0.866	0.433	0.289	1	1
min	1	0	0.333	-1.155	1.155
max	3	1	1	1.155	3.464

TAB. 2-5

I grafici presentati nelle Fig.2-4 - Fig.2-10 illustrano la posizione dei 9 punti relativi alle due variabili X1 e X2, in funzione dei diversi tipi di scalatura.

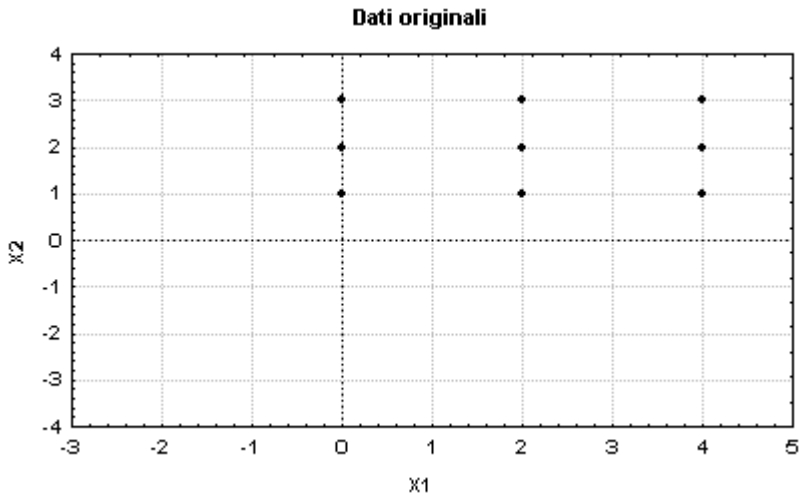

Dati originali
X2
X1
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5

FIG. 2-4

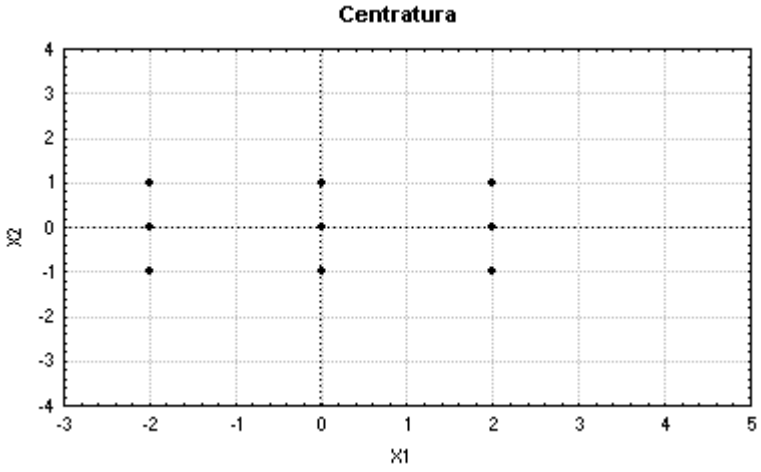

Centratura
X2
X1
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5

FIG. 2-5

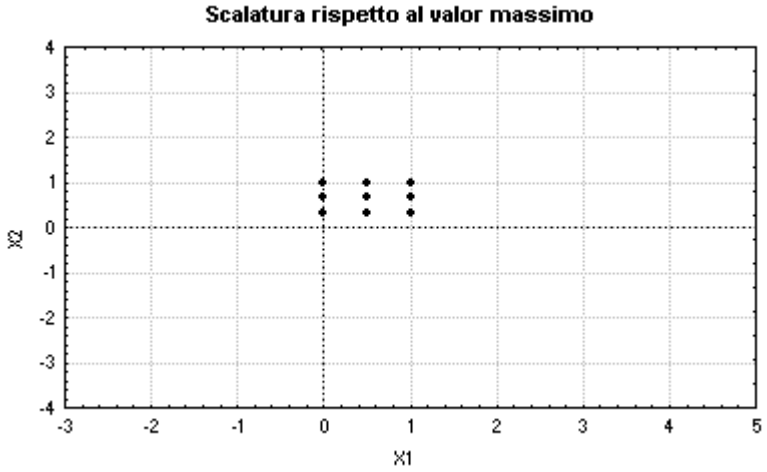

Scalatura rispetto al valor massimo
X2
X1
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5

FIG. 2-6

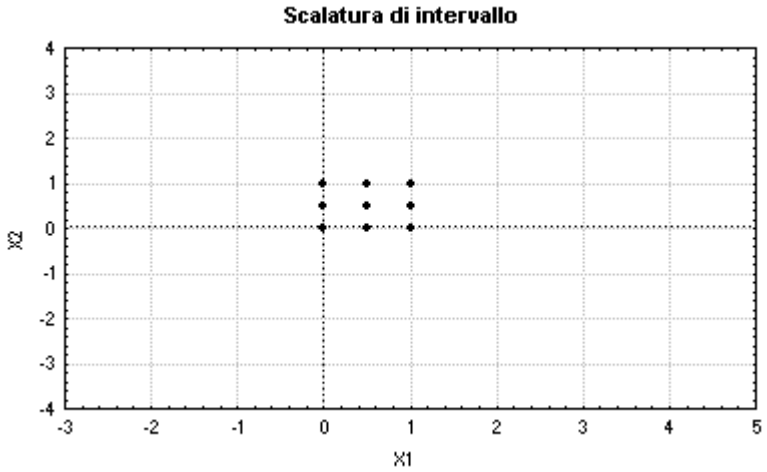

Scalatura di intervallo
X2
X1
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5

FIG. 2-7

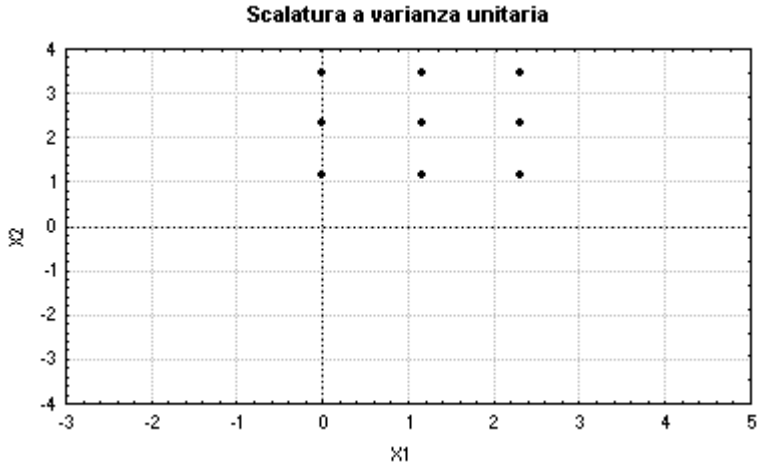

Scalatura a varianza unitaria
X2
X1
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5

FIG. 2-8

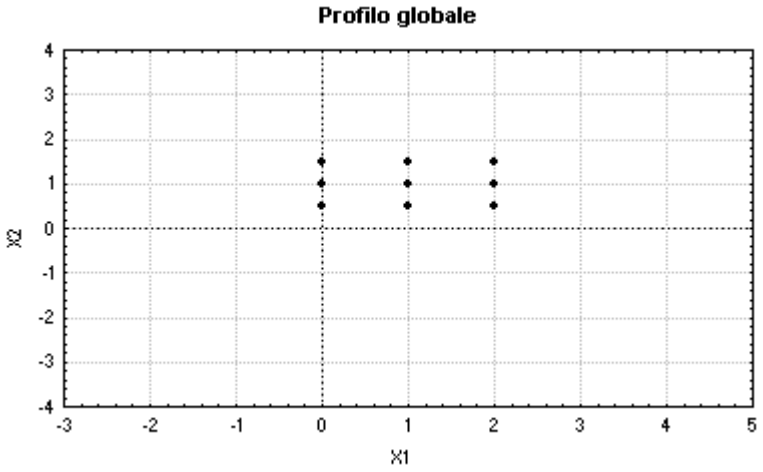

Profilo globale
X2
X1
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5

FIG. 2-9

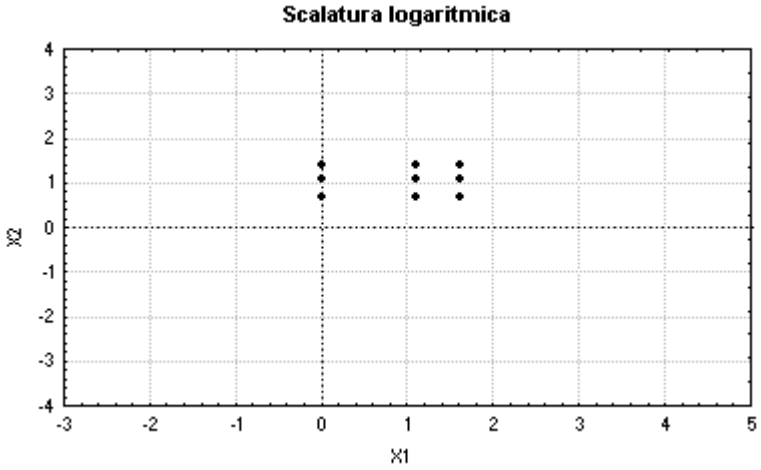


FIG. 2-10

🕮 BIBLIOGRAFIA

I.E.FRANK E R.TODESCHINI (1994). *The data analysis handbook*. Elsevier, Amsterdam.

J.H. ZAR (1984). *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, N.J.

W.J. KRZANOWSKI (1988). *Principles of Multivariate Analysis. A User's Perspective*. Oxford Univ. Press, Oxford.

3

L'ANALISI DELLE COMPONENTI PRINCIPALI

3.1 - Introduzione

L'Analisi delle Componenti Principali è una tecnica di analisi multivariata, proposta da Karl Pearson nel 1901 e sviluppata nella sua forma attuale da Harold Hotelling nel 1933, di fondamentale importanza per l'esplorazione dei dati.
In generale, un certo numero di variabili che descrivono i dati sono trasformate in nuove variabili, chiamate **componenti principali**, che sono delle combinazioni lineari delle variabili originali e la cui caratteristica più importante è quella di essere tra loro ortogonali.
L'analisi delle componenti principali è la più importante tra le diverse tecniche per l'esplorazione dei dati basate sulla decomposizione della matrice dei dati in **fattori** (che vengono anche chiamati, a seconda del metodo utilizzato, **componenti principali** o **variabili latenti**). Alcuni altri metodi della famiglia dei metodi di decomposizione della matrice dei dati in fattori sono presentati nel Capitolo 9.

3.2 - L'Analisi delle Componenti Principali

L'Analisi delle Componenti Principali (PCA, *Principal Component Analysis*) è una delle tecniche fondamentali per l'analisi multivariata dei dati.
Mediante questa tecnica è possibile:

- valutare le correlazioni tra le variabili e la loro rilevanza
- visualizzare gli oggetti (individuazione di outliers, di classi, eccetera.)
- sintetizzare la descrizione dei dati (eliminazione di rumore o informazione spuria)
- ridurre la dimensionalità dei dati
- ricercare proprietà principali

- definire un modello di rappresentazione dei dati in uno spazio ortogonale

Inoltre, PCA è un passaggio intermedio per molte tecniche multivariate.
La PCA consiste in un processo di rotazione dei dati originali definiti da una matrice **X** di dimensione *n* x *p*, effettuato in modo che il primo nuovo asse sia orientato nella direzione di massima varianza dei dati, il secondo sia perpendicolare al primo e sia nella direzione della successiva massima varianza dei dati, e così di seguito per tutti i *p* nuovi assi.
Nella Fig. 3-1 viene mostrato un esempio in due sole variabili. Come si può osservare dalla figura, la prima componente principale (PC 1) è nella direzione di massima vrianza dei e la sua origine è situata nel valor medio della variabile. La varianza residua viene rappresentata dalla seconda componente principale (PC 2), nella direzione perpendicolare alla prima componente. Poichè in questo caso abbiamo in tutto due sole variabili, le due componenti descrivono interamente i dati iniziali. Ciascuna delle due componenti è una combinazione lineare delle due variabili originali.

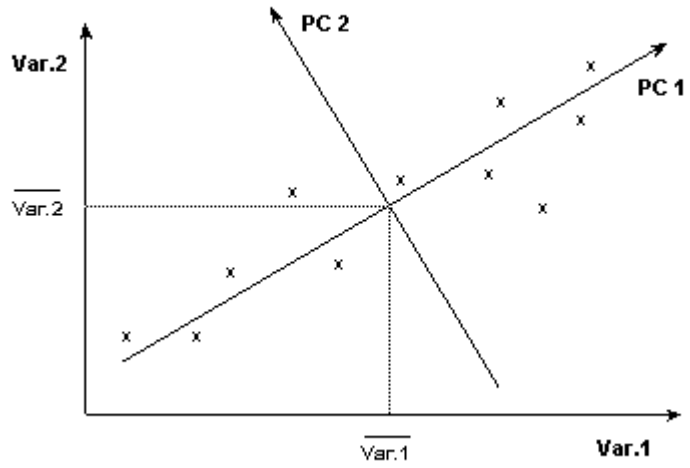


FIG. 3-1

La procedura matematica per la determinazione delle componenti principali consiste nel calcolo di autovalori e autovettori della matrice di covarianza (o di

correlazione) dei dati **X**, cioè nella diagonalizzazione della matrice di covarianza **S** di **X**, definita come:

$$\mathrm{diag}(\mathbf{S}) = \mathrm{diag}\left[\frac{\mathbf{X}_c^T \mathbf{X}_c}{n-1}\right]$$

dove $\mathbf{X}_C$ è la matrice dei dati centrata.

Nota. Di norma, l'analisi delle componenti principali viene effettuata sulla matrice di covarianza o, più comunemente, su quella di correlazione. Si osservi che la matrice di covarianza ottenuta dopo avere autoscalato i dati originali coincide con la matrice di correlazione dei dati originali.

La diagonalizzazione della matrice di covarianza comporta la determinazione di una matrice diagonale $\mathbf{\Lambda}$ (p, p), detta **matrice degli autovalori**, i cui elementi diagonali sono gli **autovalori** λ_m, ordinati in modo decrescente, e di una **matrice dei loadings L** (p, M), le cui colonne sono gli **autovettori** $\mathbf{l}_m$ della matrice di covarianza, cioè ciascuna colonna contiene i coefficienti dell'autovettore corrispondente e M è, in generale, minore o uguale a p. Gli autovettori sono i versori nel nuovo spazio. Gli assi del nuovo spazio (componenti principali, detti anche fattori o autovettori) sono gli assi relativi alle direzioni di massima varianza, in ordine via via decrescente.
La matrice di covarianza può essere decomposta nelle due matrici **L** e $\mathbf{\Lambda}$ mediante la tecnica di decomposizione a valore singolo (SVD) come:

$$\mathbf{S} = \mathbf{L} \cdot \mathbf{\Lambda} \cdot \mathbf{L}^T = \sum_p \lambda_p \mathbf{l}_p \mathbf{l}_p^T$$

Risulta quindi possibile rappresentare la matrice dei dati **X** in un nuovo spazio ortogonale, secondo la seguente relazione:

$$\mathbf{T} = \mathbf{X}\,\mathbf{L}$$

$$(n, M) = (n, p)\,(p, M)$$

dove **L** ha la funzione di una matrice di rotazione e **T** è chiamata **matrice degli scores**. Nel caso in cui $M = p$, l'operazione consiste in una semplice rotazione dei dati originali in un nuovo sistema di coordinate, senza alcuna modifica dell'informazione complessiva inizialmente contenuta nella matrice dei dati **X**.
Poichè gli autovalori λ_m rappresentano la varianza associata a ciascun autovettore (componente principale), è in generale probabile che gli autovalori più piccoli siano associati a variabilità dovute a rumore o a informazione non rilevante. In questi casi è possibile eliminare questa parte di variabilità dei dati prendendo in considerazione solo un numero M di componenti minore di p. Questo aspetto della PCA è del tutto fondamentale e sono stati proposti molti metodi per determinare il numero M di componenti principali significative (v. oltre).

Le proprietà fondamentali delle componenti principali $\mathbf{t}_m$ sono:

1. $E(\mathbf{t}_m) = 0$ se i dati originali sono almeno centrati.

2. $V(\mathbf{t}_m) = \lambda_m$

3. $C(\mathbf{t}_m, \mathbf{t}_{m'}) = 0$ per $m \neq m'$

4. $V(\mathbf{t}_1) \geq V(\mathbf{t}_2) \geq .... \geq V(\mathbf{t}_p) \geq 0$

5. $\Sigma_m V(\mathbf{t}_m) = \text{tr}(\mathbf{S})$

6. $\Pi_m V(\mathbf{t}_m) = |\mathbf{S}|$

E, V e C indicano, rispettivamente, il valore atteso, la varianza e la correlazione.

Dopo il calcolo delle componenti principali, la procedura inversa consente di rappresentare la matrice originale dei dati **X** come prodotto di due matrici:

$$\hat{\mathbf{X}} = \mathbf{T}\,\mathbf{L}^{\mathsf{T}}$$

$$(n, p) = (n, M)\,(M, p)$$

dove **T** e **L** sono rispettivamente le matrici degli *scores* e dei *loadings*.
Se il numero *M* di componenti considerate coincide col numero totale *p* di variabili, la matrice originale dei dati **X** viene interamente riprodotta nella decomposizione [**T**, **L**]; se il numero *M* di componenti considerate significative è minore di *p*, la *matrice* dei dati *riprodotta* $\hat{\mathbf{X}}$ dal prodotto della scomposizione è un'approssimazione di **X** da cui, in generale, è stata eliminata l'informazione spuria o il rumore.

3.3 - *Loadings* e *scores* in PCA

La matrice **L** dei *loadings* è la matrice le cui colonne rappresentano gli autovettori della matrice di covarianza (o di correlazione); le righe rappresentano le variabili originali: ciò significa che, selezionato un autovettore, in ciascuna riga troviamo i coefficienti numerici che rappresentano l'importanza di ciascuna variabile originale in quell'autovettore (Fig. 3-2).
I *loadings* sono *coefficienti lineari standardizzati*, cioè la somma dei quadrati dei *loadings* di un autovettore è uguale a 1 ovvero gli autovettori hanno varianza unitaria; valgono quindi le relazioni:

$$-1 \le \ell_{jm} \le +1 \qquad \sum_j \ell^2_{jm} = 1$$

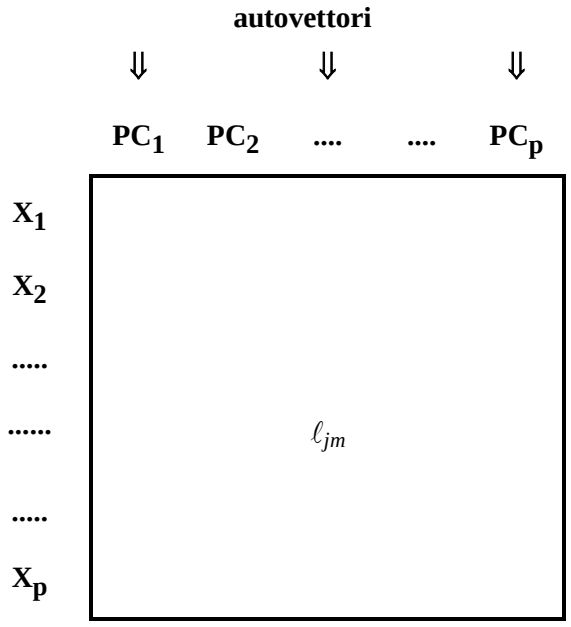


FIG. 3-2

Un valore di ℓ_{jm} vicino a 1 in valore assoluto indica che la componente *m*-esima è rappresentata soprattutto dalla *j*-esima variabile originale; viceversa, un valore di ℓ_{jm} vicino a zero indica che la *j*-esima variabile non è rappresentata (non è importante) nella *m*-esima componente.
Il valore degli *scores* è il risultato di una combinazione lineare, in cui le variabili sono le variabili originali (generalmente scalate) e i cui coefficienti moltiplicativi sono i *loadings* della *m*-esima componente:

$$t_{im} = x_{i1} \cdot \ell_{1m} + x_{i2} \cdot \ell_{2m} + \ldots + x_{ip} \cdot \ell_{pm} = \sum_j x_{ij} \ell_{jm}$$

ovvero $t_{im} = \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{l}_m$, ove entrambi i vettori hanno lunghezza *p*, cioè il numero di variabili originali.

Nota. Note le componenti principali $\mathbf{t}_m$, i *loadings* possono essere calcolati in base alle seguenti espressioni:

$$w_{jm} = \frac{x_{ij} \cdot t_{im}}{\sum_i t_{im} \cdot t_{im}} \qquad \ell_{jm} = \frac{w_{jm}}{\sqrt{\sum_j w_{jm}^2}}$$

e, in notazione vettoriale,

$$\mathbf{w}_m = \mathbf{x}^{\mathrm{T}} \mathbf{t}_m / \left(\mathbf{t}_m^{\mathrm{T}} \mathbf{t}_m \right) \qquad \boldsymbol{\ell}_{\mathbf{m}} = \frac{\mathbf{w_m}}{\left(\mathbf{w_m^T w_m} \right)^{1/2}}$$

$$(p, 1) = (p, n)\ (n, 1)$$

dove $\mathbf{w_m}$ e $\boldsymbol{\ell}_{\mathbf{m}}$ sono, rispettivamente i *loadings* non normalizzati e normalizzati.

Diversamente dai *loadings* i cui valori sono limitati tra ±1, gli *scores* hanno valor medio uguale a zero, ma possono assumere valori numerici qualsiasi. Gli *scores* rappresentano le nuove coordinate degli oggetti nello spazio delle componenti principali.
Un aspetto di grande rilevanza nello studio di problemi multivariati riguarda la possibilità di "vedere" graficamente i dati. L'analisi delle componenti principali ci fornisce una soluzione algebrica che ci consente anche rappresentazioni grafiche molto efficaci sia dei soli oggetti (***scores plot***) sia delle sole variabili (***loadings plot***) sia di oggetti e variabili contemporaneamente (***biplot***).
Per ogni coppia di componenti principali la quantità di varianza totale rappresentata nel grafico è data dalla somma delle varianze spiegate dalle singole componenti.

🗀 I grafici dei loadings (*loadings plot*)

Questo tipo di grafico consente di analizzare il *ruolo di ciascuna variabile* nelle diverse componenti, le loro correlazioni dirette e inverse, la loro importanza.

Una volta scelte le due componenti che costituiscono gli assi coordinati, le coordinate di ciascuna variabile sono definite dalla coppia di *loadings* (matrice **L** dei loadings) che ciascuna variabile ha nelle due componenti considerate. Per questo tipo di grafico, ciascuna variabile sarà sempre necessariamente compresa tra -1 e +1 (l'intervallo di definizione dei *loadings*). Ad esempio, variabili che si trovano vicino al centro (il punto 0,0) sono variabili che non sono rilevanti per nessuna delle due componenti; variabili che si collocano agli estremi di una delle componenti sono invece variabili importanti in questa componente (*loadings* grandi in valore assoluto). Quindi, *loadings* grandi positivi o negativi per alcune variabili indicano che queste variabili sono significativamente rappresentate nella componente.
Gruppi di variabili che compaiono vicine nel grafico dei loadings indicano che, limitatamente all'informazione portata da queste componenti, esse portano un'informazione comune o simile (sono cioè correlate): se questo accade per tutte le componenti prese in considerazione come modello, è possibile rappresentare il contenuto di informazione portato da questo gruppo di variabili con una sola di esse. Questo vale anche per variabili che appaiono in posizione opposta le une alle altre rispetto all'origine (il punto 0,0): anche in questo caso queste variabili sono correlate o, più correttamente, inversamente correlate.

🗀 I grafici degli scores (*scores plot*)

Questo tipo di grafico consente di analizzare il *comportamento degli oggetti* nelle diverse componenti e le loro similarità.
Una volta scelte le due componenti che costituiscono gli assi coordinati, le coordinate di ciascun oggetto sono definite dalla coppia di *scores* (matrice **T** degli scores) che ciascun oggetto ha per le due componenti considerate. Il grafico degli *scores* consente di analizzare i comportamenti degli oggetti "visti" alla luce delle componenti considerate, cioè alla luce del loro significato e dei valori delle variabili che maggiormente le caratterizzano. In questo modo è possibile notare raggruppamenti di oggetti simili (clusters), la presenza di oggetti particolari (outliers), il manifestarsi di particolari regolarità e distribuzioni.

🗀 I grafici *Biplot*

Il grafico biplot consente di rappresentare contemporaneamente oggetti e variabili al fine di poter valutare le relazioni che tra essi intercorrono. Un

confronto ottimale tra oggetti e variabili avviene attraverso l'analisi delle corrispondenze e l'analisi delle mappe spettrali.
Più semplicemente, nell'analisi delle componenti principali, il confronto consente di attribuire la posizione dei campioni nello *score plot* ai valori delle variabili importanti nel corrispondente *loading plot* (v. anche par. 9.2).

3.4 - La correlazione nei dati

Poichè la presenza di correlazione nei dati ha una grande rilevanza sull'esito finale dell'applicazione di molti metodi chemiometrici, è di grande interesse riuscire ad avere una stima della quantità di correlazione contenuta nei dati.
La **matrice di correlazione R** contiene esplicitamente tutte le informazioni riguardanti le correlazioni tra le variabili considerate. La diagonale principale della matrice di correlazione è costituita da valori tutti uguali ad uno (ogni variabile è perfettamente correlata con se stessa). Gli elementi non diagonali ci informano sulla correlazione tra tutte le coppie di variabili: questi valori sono compresi tra -1 e +1. Nel primo caso si ha una perfetta correlazione inversa (quando una variabile cresce, l'altra diminuisce); nel secondo caso si ha una perfetta correlazione diretta (quando una variabile cresce, anche l'altra cresce).
Attraverso la decomposizione della matrice dei dati in autovalori e autovettori, effettuata su dati autoscalati (cioè diagonalizzando la matrice di correlazione), è possibile stimare la quantità di correlazione contenuta nei dati.
Per comprendere meglio sia le proprietà degli autovalori calcolati da PCA sia il concetto di correlazione, consideriamo i quattro casi teorici A-D.

CASO A

Tutte le *p* variabili originali che descrivono i dati sono perfettamente ortogonali e sono equivarianti (cioè la quantità di varianza spiegata da ciascuna variabile è la stessa). In questo caso, ricordando che la somma degli autovalori è uguale a *p*, gli autovalori sono i seguenti:

$$\lambda_1 = \lambda_2 = \ldots = \lambda_p = 1$$

La varianza spiegata da ciascuno di essi è quindi 1/*p*. La quantità di correlazione contenuta nei dati è uguale allo 0%.

CASO B

Tutte le *p* variabili originali sono uguali, cioè i dati sono in realtà rappresentati da una sola variabile. In questo caso, gli autovalori sono i seguenti:

$$\lambda_1 = p \qquad \lambda_2 = \lambda_3 = \ldots = \lambda_p = 0$$

La varianza spiegata dal primo autovalore è 1 (il 100%), mentre quella spiegata dagli altri *p* -1 autovalori è zero. La quantità di correlazione contenuta nei dati è il 100%.

CASO C

Per un insieme di dati definito da *p* variabili, la matrice di correlazione è così definita:

$$\mathbf{R} = \begin{vmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \ldots & 1 \end{vmatrix}$$

cioè tutte le variabili sono equicorrelate e la loro correlazione è $\rho > 0$.
Per questa matrice, il primo autovalore è:

$$\lambda_1 = 1 + (p-1)\rho$$

e la quantità di varianza spiegata dalla prima componente è data da:

$$ev_1 = \frac{\lambda_1}{p} = \rho + \frac{1-\rho}{p}$$

Tutti gli altri *p* - 1 autovalori sono:

$$\lambda_2 = \lambda_3 = \ldots = \lambda_p = 1-\rho$$

La quantità di correlazione per questa matrice è quindi $\rho \times 100$.

CASO D

Sia data una matrice di p variabili, con q variabili uguali ($q < p$). Le restanti $p - q$ variabili hanno correlazione zero tra loro e con le q variabili. Ovviamente, le q variabili hanno tra loro correlazione 1.
In questo caso, il primo autovalore è uguale a q, i successivi $p - q$ autovalori hanno ciascuno valore 1, gli ultimi $q - 1$ autovalori hanno autovalore uguale a 0.
La quantità di correlazione presente nei dati è pari a $(q - 1)/(p - 1) \times 100$.

Nota. La presenza di q autovalori uguali a zero (ottenuti da una matrice di covarianza o di correlazione) indica che gli oggetti $\mathbf{x_i}$ sono in realtà descritti da uno spazio $p - q$ dimensionale. Questo significa che esistono q differenti funzioni lineari delle p variabili originali che hanno valori costanti per ciascuno degli n oggetti.

L'indice più comune per valutare la correlazione nei dati è il cosiddetto ***condition number*** calcolato dal rapporto tra il primo e l'ultimo autovalore:

$$cn = \frac{\lambda_1}{\lambda_p} \qquad (3 - 1)$$

Più grande è il valore di cn, maggiore è la correlazione presente nei dati. Tuttavia appare subito evidente che in tutti i casi in cui anche solo l'ultimo autovalore è nullo, il *condition number* diviene infinito e non è quindi in grado di distinguere tra loro casi diversi in cui l'ultimo autovalore è nullo. In questo senso, cn fornisce una stima semi-quantitativa della correlazione dei dati. Per i 4 casi definiti, cn vale 1 nel caso A, infinito nei casi B e D, e

$$cn = \frac{1 + (p - 1) \cdot \rho}{1 - \rho}$$

nel caso C.
In modo più quantitativo, la correlazione nei dati può essere valutata da indici che forniscono una *misura di ridondanza*.
Un indice di questo tipo (**indice di correlazione *K***), proposto dall'autore, si basa sulla differenza tra la quantità di varianza spiegata da ciascun autovalore e la quantità di varianza teorica spiegata da ciascuna variabile ($1/p$). La

correlazione multivariata contenuta nei dati può quindi essere valutata utilizzando l'espressione:

$$K\% = \frac{\sum_m \left| \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{p} \right|}{\frac{2(p-1)}{p}} \times 100 \qquad (3 - 2)$$

La sommatoria scorre su tutte le *p* componenti; il suo valore minimo è 0; il valore massimo è dato da un contributo (1 - 1/*p*) e da (*p* - 1) contributi pari a 1/*p*. Il suo valore massimo è quindi dato da:

$$\frac{2(p-1)}{p}$$

K% è quindi una grandezza compresa tra 0 e 100 e fornisce una stima del contenuto percentuale di correlazione nei dati. Come si può facilmente notare, per i due casi A e B i valori di *K%* sono, rispettivamente, 0% e 100%. Per i casi C e D, le quantità di correlazione calcolate da questa espressione sono ρ % e $\frac{q-1}{p-1}\%$, rispettivamente.

Un altro indice di ridondanza utilizzato per valutare la quantità di correlazione in una matrice di dati è l'**indice di Gleason-Staelin**, definito come

$$\varphi = \sqrt{\frac{\|\mathbf{R}\|^2 - p}{p \cdot (p-1)}} \qquad (3 - 3)$$

dove

$$\|\mathbf{R}\|^2 = \sum\sum r_{jk}^2 = \sum_m \lambda_m^2$$

e **R** è la matrice di correlazione.
Questo indice, chiamato anche misura di ridondanza, si comporta come l'indice sopra definito nei casi A, B e C, ma fornisce risultati diversi nel caso D, ove l'espressione diviene:

$$\varphi_D = \sqrt{\frac{q^2 + (p-q) - p}{p \cdot (p-1)}} = \sqrt{\frac{q \cdot (q-1)}{p \cdot (p-1)}}$$

Ad esempio, per $p = 5$ e $q = 2$, i due indici assumono i valori:

$$K\% = 20\% \qquad \varphi\% = 31.6\%$$

Diversamente da quest'ultimo caso, su insiemi di dati reali, in generale l'indice di Gleason-Staelin fornisce valori di correlazione più bassi dell'indice *K*.

3.5 - Il numero di componenti significative

Uno dei problemi fondamentali che l'analisi delle componenti principali comporta è la determinazione del numero M di componenti (fattori) significativi, con $M < p$. La ricerca del numero di componenti significative è nota come ***rank analysis***. Infatti, se i dati contengono una struttura informativa (cioè, non casuale), la separazione tra la variabilità dovuta a rumore sperimentale o informazione spuria e informazione utile avviene mediante una delimitazione opportuna del numero di componenti principali significative. Qualsiasi procedura di selezione di un numero ridotto di componenti significative presuppone che la variabilità dell'informazione utile sia più grande della variabilità associata al rumore sperimentale o ad informazione secondaria.
Poichè ciascun autovalore rappresenta la varianza associata alla corrispondente componente principale e la somma di tutti gli autovalori coincide con la varianza totale presente nei dati, la **varianza percentuale spiegata** dalla prima componente principale ($EV_1\%$) rispetto alla varianza totale è data da

$$EV_1\% = \frac{\lambda_1}{\sum_{m=1}^{p} \lambda_m} \cdot 100$$

Ad esempio, la varianza spiegata da un grafico degli scores (o dei loadings) per le prime due componenti (PC1 e PC2) è

$$EV\% = \frac{\lambda_1 + \lambda_2}{\sum_{m=1}^{p} \lambda_m} \cdot 100$$

Ciò significa che il grafico bidimensionale che stiamo esaminando ci mostra una quantità *EV%* della varianza totale dei dati.
Se *M* è il numero di componenti significative, la varianza spiegata dalle prime *M* componenti (**varianza cumulata**) è definita come:

$$Cum.E.V.\% = \frac{\sum_{m=1}^{M} \lambda_m}{\sum_{m=1}^{p} \lambda_p} \times 100$$

Naturalmente, selezionate *M* componenti signficative, la **varianza residua** *R.V.*, cioè la varianza non spiegata, è

$$R.V.\% = \frac{\sum_{m=M+1}^{p} \lambda_m}{\sum_{m=1}^{p} \lambda_p} \times 100$$

Esistono molte tecniche per l'accertamento del numero di componenti significative, ma è bene sottolineare che, in generale, la risposta non è totalmente univoca.

🗀 *grafico degli autovalori (scree plot)*

Il numero *M* di autovalori da ritenere viene valutato in base all'analisi grafica degli autovalori riportati contro il numero di fattori. In questo tipo di grafico, noto come ***scree plot***, si riportano sull'asse delle ascisse il numero delle componenti e sull'asse delle ordinate gli autovalori corrispondenti.
Vengono scelti i primi *M* fattori per i quali l'abbassamento di varianza residua risulta più accentuato. Questo è un metodo sostanzialmente grafico-visuale, consigliabile solo per i casi più semplici. Se gli autovalori hanno valori molto diversi tra loro, con i primi autovalori molto grandi, il grafico viene spesso costruito utilizzando il logaritmo degli autovalori.

In modo analogo, è possibile esaminare i rapporti tra autovalori successivi, cioè costruire un grafico della sequenza:

$$R_m = \frac{\lambda_m}{\lambda_{m+1}}$$

Anche in questo caso non si tratta di un vero e proprio criterio per la selezione del numero di componenti significative, ma l'analisi dei rapporti tra autovalori successivi può consentire di individuare rapidamente la presenza di salti che stanno ad indicare un mutamento della struttura latente dei dati. Con questo criterio, di carattere qualitativo, vengono ritenuti i primi *M* fattori corrispondenti al valor massimo del rapporto.

🗀 *criterio dell'autovalore medio (AEC e CAEC)*

Secondo questo criterio di scelta, sono significativi tutti i fattori i cui autovalori sono maggiori del valor medio degli autovalori.

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq \frac{\sum_{m=1}^{p} \lambda_p}{p} = \bar{\lambda}$$

Quando le variabili sono autoscalate, o, il che è la stessa cosa, la diagonalizzazione avviene sulla matrice di correlazione, il valor medio degli autovalori è 1. In questo caso, vengono ritenuti i primi *M* fattori il cui autovalore è maggiore di 1. Poichè molti ritengono che questo criterio fornisca normalmente un numero di componenti signifcative troppo piccolo, è stato proposta una semplice variante secondo cui il livello oltre il quale le componenti vengono considerate significative viene abbassato, moltiplicando il valore medio per 0.7 (CAEC).

🗀 *funzione indicatrice di Malinowski* (*MIF*)

E' una funzione empirica definita come:

$$MIF = \frac{\left(\sum_{m=M+1}^{p} \frac{\lambda_m}{n(p-M)} \right)^{1/2}}{(p-M)^2}$$

Il numero *M* di fattori significativi viene determinato in corrispondenza del minimo di questa funzione.

Per il calcolo delle componenti significative, recentemente, l'autore ha proposto due funzioni dell'indice di correlazione *K* precedentemente definito.
Le due espressioni sono:

$$KL = 1 + \text{cint}\ (p-1)\cdot(1-K)$$

$$KP = cint\left(p^{(1-K)}\right)$$

dove *cint* è il valore arrotondato all'intero più vicino del prodotto considerato.
Il criterio *KL* consente di definire un *limite superiore per le componenti principali significative*, mentre il criterio *KP* fornisce un limite inferiore.
In entrambi I casi, si può osservare il corretto comportamento della funzione nei due casi teorici A ($K = 0 \rightarrow KL = KP = p$,) e B ($K = 1 \rightarrow KL = KP = 1$), cioè nel caso in cui non vi sia correlazione, il numero di componenti significative coincide col numero totale di variabili, mentre, nel caso in cui vi sia perfetta correlazione (ad esempio, tutte le variabili sono uguali tra loro), esiste un'unica componente significativa.

Nota. Da prove effettuate su molti insiemi di dati reali, il criterio *KP* fornisce mediamente un numero di componenti significative confrontanibili con quelli ottenuti con i criteri *AEC* e *BS*, mentre il criterio *KL* fornisce mediamente risultati di poco superiori a quelli forniti dal criterio *CAEC*.

🗀 *validazione incrociata* (*Double-Cross Validation, DCV*)

Il criterio si basa sulla valutazione di come la matrice originale dei dati **X** viene riprodotta utilizzando un crescente numero di componenti *M*, in accordo con l'espressione:

$$\hat{\mathbf{X}} = \mathbf{T}\,\mathbf{L}^{\mathsf{T}}$$

Per valutare il numero di componenti significative, viene calcolata la grandezza:

$$PRESS(M) = \sum_i \sum_j \left(\hat{x}_{ij}(M) - x_{ij} \right)^2$$

dove *PRESS* rappresenta uno scarto quadratico in predizione tra I dati originali e i dati riprodotti utilizzando le prime *M* componenti.
La decisione sul numero *M* di componenti significative viene presa in relazione al valore del rapporto:

$$R = \frac{PRESS(M)}{PRESS(M-1)}$$

Le componenti sono significative fino a che il rapporto *R* si mantiene inferiore a uno.

Altri metodi a volte utilizzati per la stima del numero di componenti significative sono la *funzione-errore di Malinowski* (*Imbedded Error, IE*) e il *test F di Malinowski* (*Malinowki F-ratio test, MFR*).

3.6 - Le proprietà principali

Uno degli elementi che caratterizza maggiormente l'analisi in componenti principali è il concetto di **proprietà principale**. Come si è visto, l'analisi delle componenti principali produce delle *combinazioni lineari dei descrittori originali* che rappresentano, in ordine decrescente, le direzioni di massima varianza dello spazio sperimentale considerato. I descrittori che sono tra loro correlati sono descritti dalla *stessa* componente principale poichè portano lo stesso tipo di informazione. Le componenti principali sono vettori mutuamente ortogonali e differenti componenti descrivono variazioni indipendenti e non correlate dei descrittori.

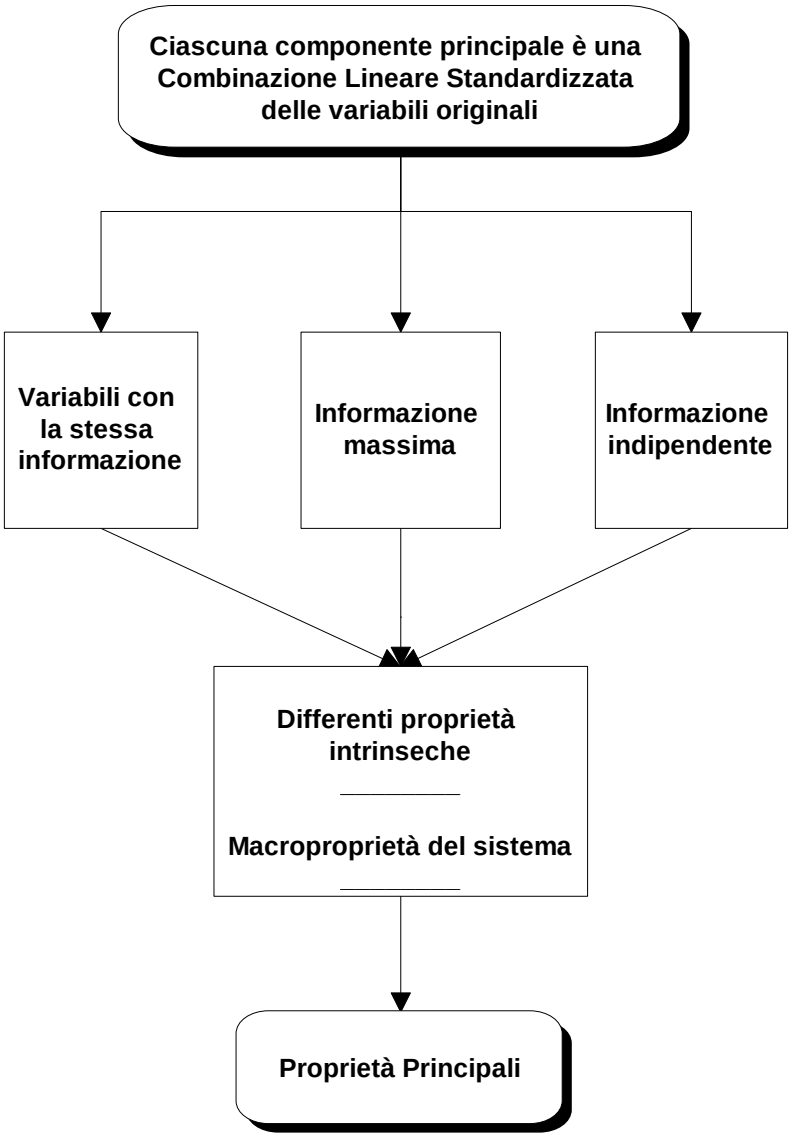

Ciascuna componente principale è una Combinazione Lineare Standardizzata delle variabili originali
Variabili con la stessa informazione
Informazione massima
Informazione indipendente
Differenti proprietà intrinseche
Macroproprietà del sistema
Proprietà Principali

FIG. 3-3

Ciò significa che componenti differenti rappresentano variazioni nei dati dovute a **differenti proprietà intrinseche**.
Queste proprietà intrinseche si manifestano quindi come le variazioni più significative dei descrittori sperimentali originali: esse rappresentano in qualche modo delle *proprietà principali*, ovvero **macroproprietà del sistema** non direttamente misurabili.
Per quanto sia un aspetto spesso trascurato, il fatto di *interpretare*, di *comprendere il significato* delle proprietà principali è, a mio giudizio, di grande rilevanza per consentire un salto qualitativo nella conoscenza profonda del sistema studiato. Si passa infatti da una descrizione - conoscenza del sistema in termini di variabili macroscopiche originali, ciascuna normalmente nota nei suoi significati, ad una descrizione ad un livello semantico superiore - *una metadescrizione* - il cui ruolo può consentire di cogliere le **proprietà emergenti del sistema**, cioè proprietà *nuove* rispetto alle conoscenze originali, proprietà che emergono da una *visione olistica del sistema* o dagli *effetti sinergici o antagonisti delle variabili originali* che lo descrivono.
Ad esempio, se dei composti chimici sono descritti contemporaneamente dal loro peso molecolare, dal numero di atomi e di legami che li compongono, dalla rifrazione molare, dal loro volume molecolare e dalla superficie totale, oltre che da altri tipi di descrittori, possiamo aspettarci di trovare una specifica componente principale in cui ciascuno di questi descrittori sia ampiamente rappresentato (cioè, i loro *loadings* sono significativamente diversi da zero). Se ne può dedurre che questa componente rappresenta una *nuova macrovariabile di carattere dimensionale* il cui significato va oltre quello dei singoli descrittori considerati: essa può essere pensata come un nuovo concetto di dimensione molecolare, di carattere più generale.
In un certo senso, abbiamo una nuova definizione operativa di variabile dimensionale: essa non corrisponde più a nessuna delle tipiche grandezze dimensionali con cui rappresentiamo le molecole, ma è una combinazione lineare di queste ultime che rappresenta l'emergere di un nuovo piano semantico dimensionale, ad un livello di complessità superiore e più generale.
A tutti gli effetti pratici, queste nuove variabili - le proprietà principali - siano esse interpretate o meno, costituiscono una nuova descrizione sintetica del sistema considerato. Questo aspetto apre la via a potenzialità applicative di straordinaria rilevanza.
Caratteristica importante delle proprietà principali è il fatto che ciascuna di esse raccorda solo la parte più informativa di ciascuna delle variabili originali che contribuiscono alla combinazione lineare. Ciò significa che soltanto la variazione sistematica delle variabili originali contribuisce a definire le componenti principali più significative: la variazione non rilevante o la

variazione causata da rumore sperimentale o la variazione sistematica non correlata con la componente considerata non vengono rappresentate nella componente.
Diversi metodi chemiometrici sono stati ideati per sfruttare l'informazione essenziale contenuta nelle proprietà principali (Fig.3-4). In particolare, nei metodi di regressione, questo obiettivo viene realizzato con la regressione in componenti principali (*Principal Component Regression, PCR*) e nel metodo di regressione ai minimi quadrati parziali (*Partial Least Squares regression, PLS*).

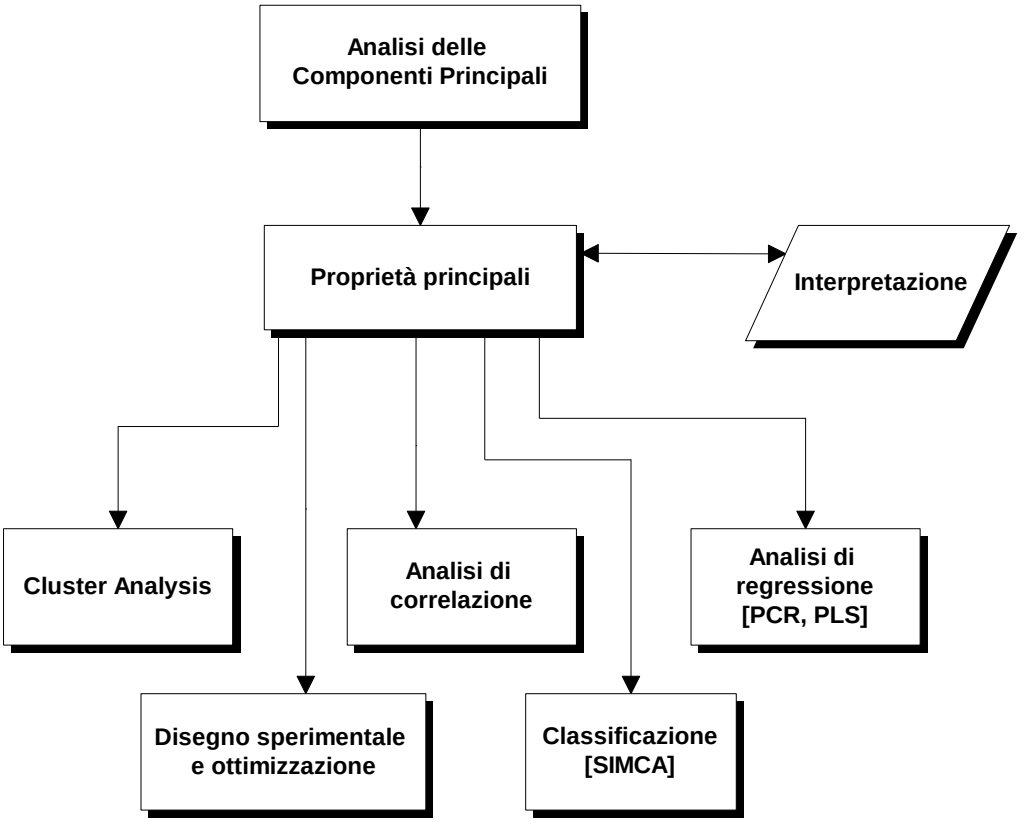


FIG. 3-4

Nei metodi di classificazione, il metodo SIMCA utilizza esplicitamente le componenti principali significative di ogni classe per costruire modelli di classe

che sfruttano l'informazione significativa per definire le caratteristiche di unicità di una classe.
In maniera ancora più generale, ciascuna proprietà principale può essere utilizzata *al posto delle variabili originali* in tutte le tecniche chemiometriche che conosciamo, compresi i metodi di ottimizzazione e di disegno sperimentale.

3.7 - Le rotazioni

Le rotazioni costituiscono normalmente l'ultimo stadio dell'analisi delle componenti principali. Lo scopo delle rotazioni degli assi principali ottenuti dall'analisi delle componenti principali è di carattere esclusivamente interpretativo: per questa ragione una rotazione va effettuata esclusivamente *dopo* aver stabilito il numero *M* di componenti significative da ritenere.
Una volta stabilito il numero delle componenti significative che rappresentano il sottospazio informativo dei nostri dati, la rotazione viene effettuata sulle *M* componenti, consentendo di *semplificare* le componenti principali ritenute, in modo da renderne più semplice l'interpretazione e più diretta l'analogia di qualcuna di esse con qualche grandezza fisica esplicita.
Normalmente infatti una componente principale è rappresentata da molte delle variabili originali utilizzate nell'analisi delle componenti principali, cioè i *loadings* di una componente sono significativamente grandi per molte variabili. Mediante le rotazioni è possibile ottenere una semplificazione della componente principale, esaltando il ruolo delle variabili più importanti in quella componente e riducendo il ruolo delle variabili meno importanti. I nuovi assi principali ottenuti dopo la rotazione sono generalmente chiamati *assi primari del modello*.
Un primo tipo di rotazioni sono quelle effettuate *soggettivamente* - **rotazioni grafiche** -, basandosi esclusivamente su un riallineamento grafico rispetto a punti ritenuti interessanti. Un altro tipo di rotazioni sono quelle denominate **target rotations** che vengono effettuate mediante il metodo chiamato **Target Transformation Factor Analysis** (**TTFA).** Con questo metodo si cerca di far coincidere il più possibile o due modelli calcolati di fattori oppure un modello calcolato di fattori con un modello ipotetico predefinito teoricamente (*target*).
Le rotazioni più comuni sono le **rotazioni analitiche**, cioè rotazioni che vengono effettuate mediante opportune procedure matematiche su un modello calcolato di fattori.
Le operazioni di rotazione comportano la parziale perdita di alcune proprietà matematiche delle componenti principali originali. Se le rotazioni sono effettuate

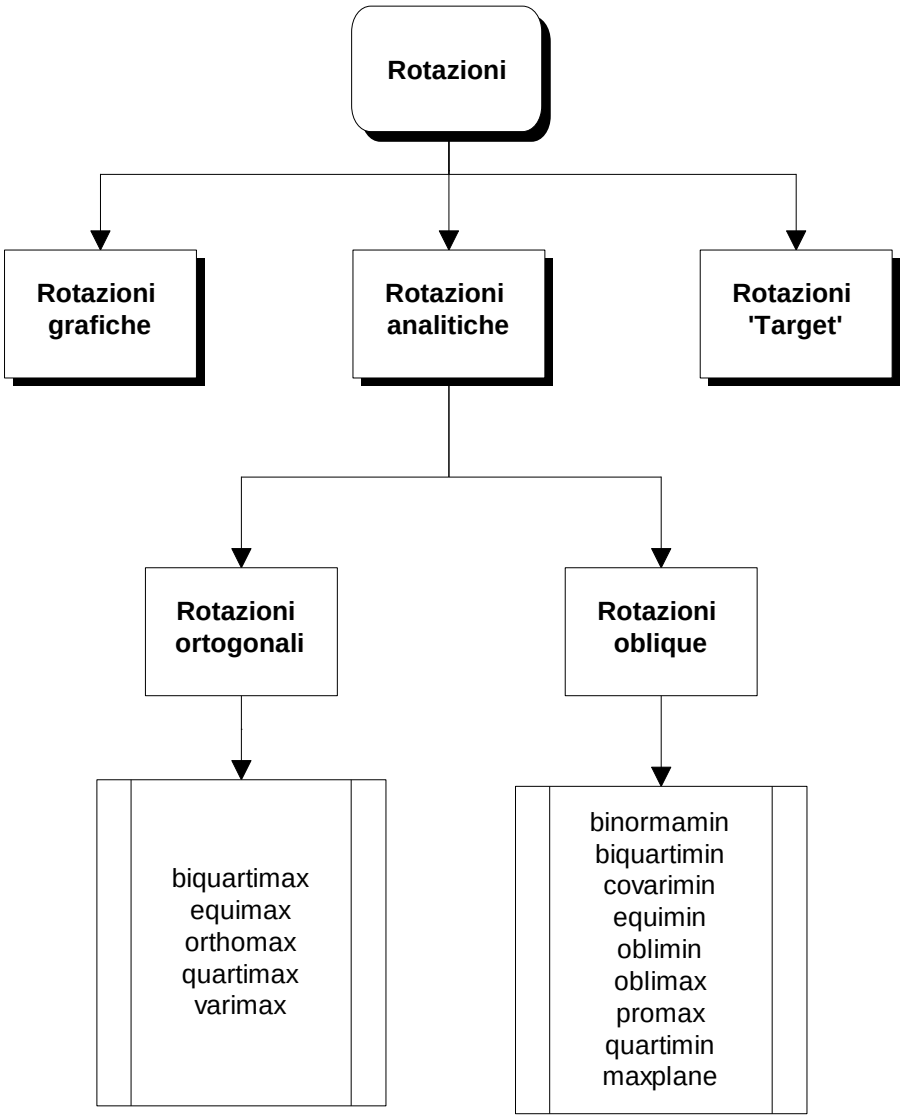


FIG. 3-5

conservando l'ortogonalità delle componenti principali - **rotazioni ortogonali** - ciò che viene meno è la corrispondenza delle nuove componenti ruotate con gli

assi di massima varianza: ciò significa, ad esempio, che la prima nuova componente non coincide più con la direzione di massima varianza, ma con un'altra direzione, spesso non troppo lontana da quella. Se si effettuano invece **rotazioni non-ortogonali**, un'ulteriore conseguenza è quella di perdere anche la perfetta ortogonalità tra le componenti principali ritenute. Ciò significa che non vengono preservati i valori delle communalità (v. Cap. 9) nè la varianza totale spiegata dalle componenti ritenute.
Per ognuna delle due tipologie di rotazioni analitiche, esistono numerosi metodi (Fig.3-5) che differiscono tra loro per i criteri utilizzati nella "semplificazione" dei *loadings* che definiscono ciascuna componente principale. Per le rotazioni ortogonali, viene comunemente consigliato di utilizzare il criterio *varimax* (in particolare, la variante *raw varimax*); il suo equivalente nel contesto delle rotazioni non-ortogonali è il criterio *covarimin*.

Esempio

Sono stati analizzati 38 campioni di vino, determinando il contenuto di 17 metalli in ciascuno di essi. (Dati: WINES)
L'analisi delle componenti principali, dopo autoscalatura dei dati, fornisce il risultato riportato in Tab.3-1.
In primo luogo, si può osservare che, essendo i dati autoscalati, la somma degli autovalori è uguale a 17, pari al numero delle variabili. Accanto alla colonna degli autovalori, sono riportate le varianze spiegate da ciascuna componente principale (*EV*%) e le varianze cumulate spiegate dalle prime *m* componenti principali (*CEV*%).
Le ultime 5 colonne successive della tabella rappresentano le analisi per il calcolo delle componenti significative con il metodo dell'autovalore medio (*AEC*, 6 PC), con i criteri *KL* (9 PC) e *KP* (4 PC) (v.oltre), del segmento spezzato (*BS*, 5 PC) e di Malinowski (*MIF*, 7 PC).
Il *condition number* è 216.5, mentre il contenuto di correlazione multivariata dei dati, calcolato utilizzando le due espressioni precedenti (par. 3-4) è: $K\% = 47.4\%$ $\varphi\% = 28.2\%$

ID	*autovalore*	*E.V.%*	*C.E.V.%*	*AEC*	*KL*	*KP*	*BS*	*MIF*
1	4.1785	24.6	24.6	*	*	*	*20.233*	*0.00908*
2	2.7468	16.2	40.7	*	*	*	*14.350*	*0.00886*
3	2.2098	13.0	53.7	*	*	*	*11.409*	*0.00868*
4	1.9349	11.4	65.1	*	*	*	*9.448*	*0.00843*
5	1.4355	8.4	73.6	*	*		*7.978*	*0.00827*
6	1.0813	6.4	79.9	*	*		6.801	*0.00821*
7	0.8527	5.0	84.9		*		5.821	*0.00821*
8	0.6082	3.6	88.5		*		4.981	0.00839
9	0.5129	3.0	91.5		*		4.245	0.00860
10	0.4287	2.5	94.1				3.592	0.00881
11	0.3711	2.2	96.2				3.003	0.00883
12	0.2542	1.5	97.7				2.469	0.00901
13	0.1682	1.0	98.7				1.978	0.00945
14	0.1151	0.7	99.4				1.526	0.00998
15	0.0495	0.3	99.7				1.106	0.01315
16	0.0333	0.2	99.9				0.714	0.02254
17	0.0193	0.1	100.0				0.346	-

TAB. 3-1

In Fig. 3-6, è riportato il grafico degli autovalori (*scree plot*), dove si può osservare che i primi sei autovalori sono maggiori di uno (il valor medio) e che dall'ottavo-nono autovalore in poi l'andamento è quasi piano.

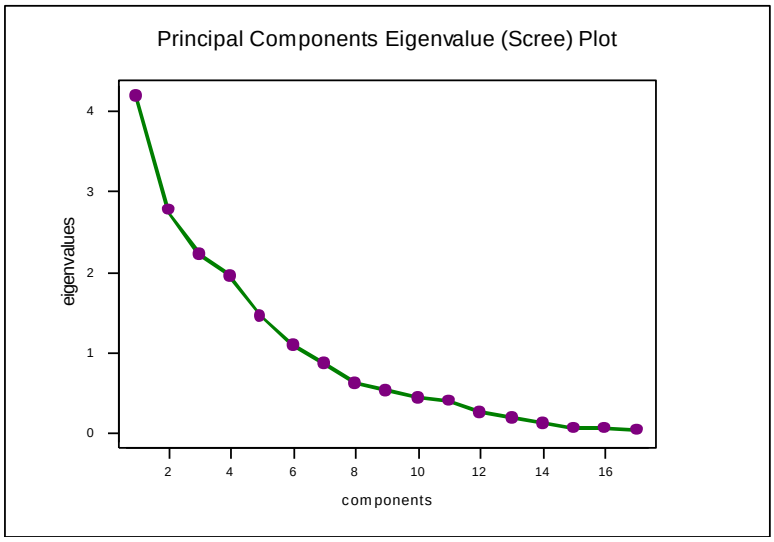


FIG. 3-6

Nella Tab. 3-2 sono riportati i *loadings* delle prime 6 componenti principali. Ogni riga rappresenta una delle variabili originali e le colonne rappresentano le componenti. In grassetto sono evidenziati i *loadings* più importanti (selezionando arbitrariamente quelli con valori assoluti superiori a 0.300) di ogni componente, cioè le variabili che più caratterizzano ciascuna componente.

MATRICE DEI LOADINGS (L)

ID	*Var.*	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
1	**Cd**	0.125	-0.285	**0.351**	0.055	**-0.369**	-0.233
2	**Mo**	-0.034	**-0.546**	0.150	-0.125	0.132	-0.096
3	**Mn**	-0.056	0.118	**0.571**	0.021	0.011	0.072
4	**Ni**	-0.109	-0.247	-0.268	-0.140	-0.107	**0.552**
5	**Cu**	-0.004	-0.122	-0.219	-0.065	**0.496**	-0.061
6	**Al**	0.039	0.130	-0.278	**-0.420**	0.047	**-0.352**
7	**Ba**	**-0.353**	0.080	0.061	-0.229	**-0.348**	-0.013
8	**Cr**	-0.271	-0.118	0.266	0.101	**0.394**	-0.087
9	**Sr**	**-0.415**	0.187	0.134	-0.166	-0.085	0.168
10	**Pb**	-0.030	**-0.537**	0.168	-0.161	0.064	-0.091
11	**B**	0.020	-0.034	-0.091	**0.618**	-0.052	-0.224
12	**Mg**	**-0.405**	-0.048	0.075	-0.084	-0.111	0.115
13	**Si**	-0.239	-0.142	-0.282	**0.308**	-0.276	-0.123
14	**Na**	**-0.303**	0.161	-0.019	-0.194	0.228	**-0.438**
15	**Ca**	-0.233	**-0.333**	**-0.339**	-0.022	-0.140	-0.116
16	**P**	-0.256	-0.024	-0.015	0.289	**0.368**	**0.342**
17	**K**	**-0.403**	0.097	-0.011	0.243	-0.029	-0.231

TAB. 3-2

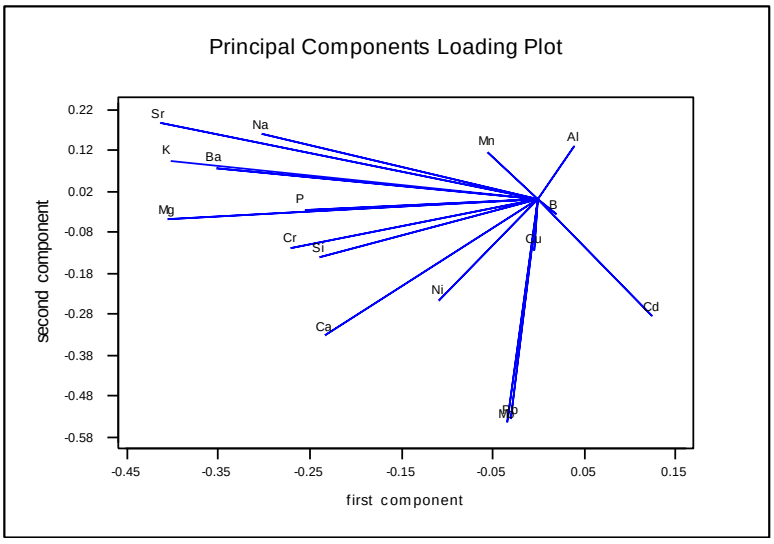


FIG. 3-7

In Fig. 3-7 è riportato il grafico dei *loadings* per le prime due componenti principali. Come si può osservare, la prima componente principale è soprattutto caratterizzata da un gruppo di variabili (i metalli alcalini e alcanino-terrosi K, Na, Mg, Ba, Sr), i cui coefficienti sono tra loro simili (valori relativamente alti negativi). La seconda componente principale è soprattutto caratterizzata da Pb e Mo (in basso al centro, sovrapposte), con piccoli contributi anche da parte delle variabili Ca, Cd e Ni.
Un aiuto nell'interpretazione delle componenti principali viene dato, oltre che dai grafici dei loadings, anche dai profili di colonna della matrice dei loadings. Nella Fig. 3-8, sono riportate sull'asse delle ascisse le 17 variabili originali e sull'asse delle ordinate i valori dei loadings, separatamente per ciascuna delle prime 6 componenti principali.

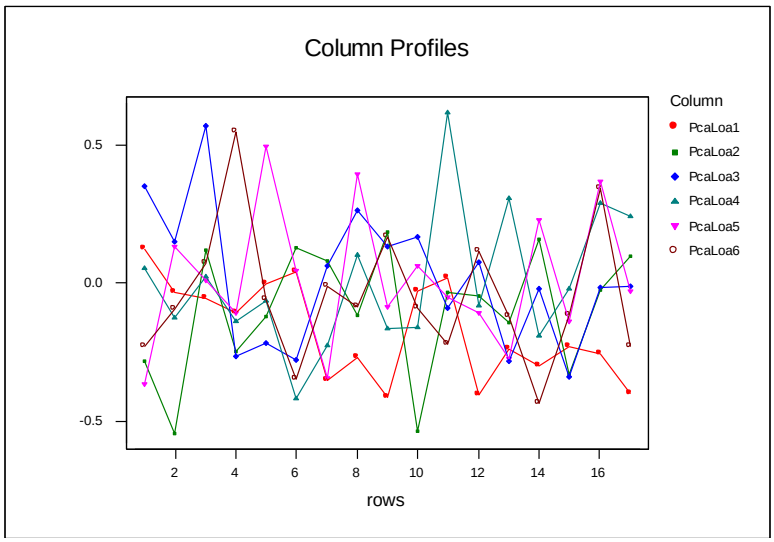


FIG. 3-8

I valori degli *scores* sono riportati in Tab. 3-3.
Nella Fig. 3-9 sono riportati i 38 oggetti (i campioni di vino) nello spazio delle prime 2 componenti principali. Il grafico consente di visualizzare il 41% della varianza totale dei dati (*C.E.V.%* in Tab. 3-1). Si può subito osservare la particolarità del campione 7 (isolato in basso). Riesaminando il corrispondente grafico dei *loadings*, si può ritenere che la posizione di questo campione sia dovuta a valori particolarmente alti di Pb e/o Mo. Alla sinistra del grafico si può

osservare la presenza di un piccolo gruppo costituito dai campioni 28, 12 e 24, facilmente caratterizzabili da valori alti di uno o più metalli alcani e alcalino-terrosi.
Si osservi che per entrambi i tipi di valutazione si è parlato di valori alti delle variabili, nonostante gli scores di questi campioni siano i più negativi. Questo è dovuto al fatto che i *loadings* (cioè i coefficienti moltiplicativi delle variabili originali) sono negativi: quindi valori alti (e positivi) delle variabili originali moltiplicati per valori alti negativi dei *loadings* portano a valori grandi negativi degli scores.

Nota. A secondo del metodo di calcolo utilizzato, è possibile ottenere autovettori i cui segni dei *loadings* sono scambiati. Questo non comporta nessun cambiamento del significato intrinseco delle componenti, ma esclusivamente un'inversione della rappresentazione grafica di *loadings* e *scores* e, quindi, un'inversione nell'interpretazione della posizione degli *scores*.

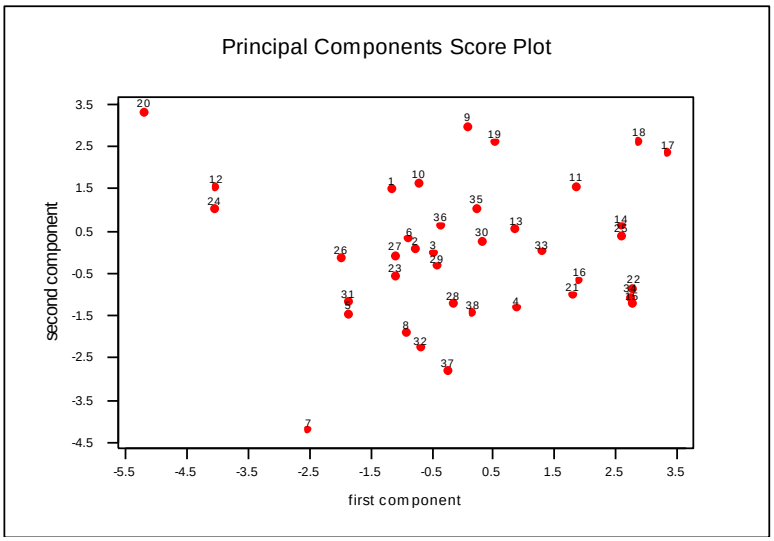


FIG. 3-9

Oggetto	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
1	-1.12737	1.47775	-2.07827	-1.60979	1.34919	0.33895
2	-0.74411	0.04378	-0.70583	-0.31868	-1.51970	-0.27384
3	-0.45637	-0.06368	-1.54691	-1.24533	0.59191	-0.36445
4	0.90533	-1.38870	-1.72905	-1.81780	-0.94783	1.71682
5	-1.84861	-1.48547	-0.39191	-2.23765	0.66357	-0.22071
6	-0.85769	0.26658	-1.84901	0.43238	-2.05345	-0.39229
7	-2.50120	-4.27559	-1.44333	-0.92856	3.03295	-0.78811
8	-0.89326	-1.92253	-2.02803	-1.16567	-0.25411	3.55823
9	0.11223	2.97094	-1.58397	-0.09269	0.32436	-0.86251
10	-0.68225	1.61484	-1.09427	-1.28924	-0.53641	-0.18019
11	1.86721	1.53469	-1.05746	0.91525	0.25538	-0.72598
12	-4.00869	1.52198	2.60220	0.18733	1.11371	0.76132
13	0.88773	0.51436	-0.76877	0.74297	0.06888	-1.17683
14	2.62597	0.55428	0.79257	1.19698	0.10511	-0.36835
15	2.78663	-1.23884	1.24125	0.49887	1.13414	-0.09682
16	1.93094	-0.70128	0.49137	1.04281	2.09485	0.14567
17	3.38489	2.32509	-0.34436	0.52488	0.33271	0.77677
18	2.90658	2.60230	-0.93507	-0.17260	0.65179	0.16273
19	0.55253	2.56926	-1.32733	-1.62324	-1.01050	-0.33094
20	-5.18763	3.30698	2.00948	0.33081	1.52376	1.93926
21	1.81760	-1.05009	0.93882	0.81639	1.67258	0.38285
22	2.81248	-0.87367	-0.30380	-1.19980	1.25243	-0.92465
23	-1.08735	-0.59116	-0.36245	2.70700	0.23244	1.09910
24	-4.03346	0.99145	0.50881	1.35307	0.95037	-2.22752
25	2.60677	0.34869	0.47992	1.00260	1.02919	0.52375
26	-1.95168	-0.18003	-1.15548	2.50805	-0.91358	-0.47762
27	-1.07694	-0.09926	-0.12849	-1.38016	-0.46632	-0.43074
28	-0.13782	-1.25803	-0.39044	2.25006	-0.91780	0.26529
29	-0.40955	-0.39299	-0.51226	1.97254	-1.37524	0.74613
30	0.35128	0.25006	2.19852	0.51655	-1.21976	0.97403
31	-1.83394	-1.21102	-0.33982	2.37453	-1.03509	-0.74639
32	-0.65744	-2.34089	-0.64259	-0.41022	-0.83874	-1.28531
33	1.30743	-0.03699	1.57417	-1.23176	-0.68224	1.02167
34	2.77393	-1.09653	1.36075	0.63944	0.32744	-0.04162
35	0.25917	1.02169	1.67138	-1.29601	-2.22115	-0.52869
36	-0.34384	0.57799	1.67300	-2.37965	-0.10732	-1.49292
37	-0.23227	-2.83764	0.54363	0.10219	-1.61786	-0.17677
38	0.18278	-1.44835	4.63302	-1.71586	-0.98962	-0.29931

TAB. 3-3

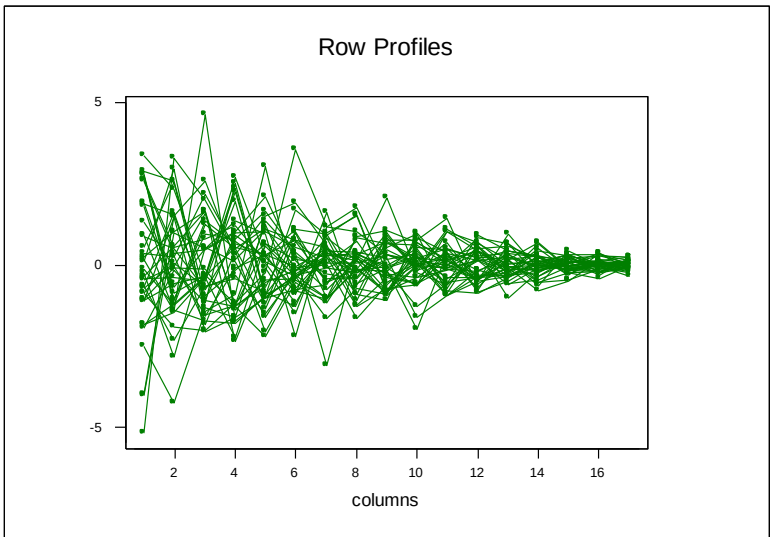


FIG. 3-10

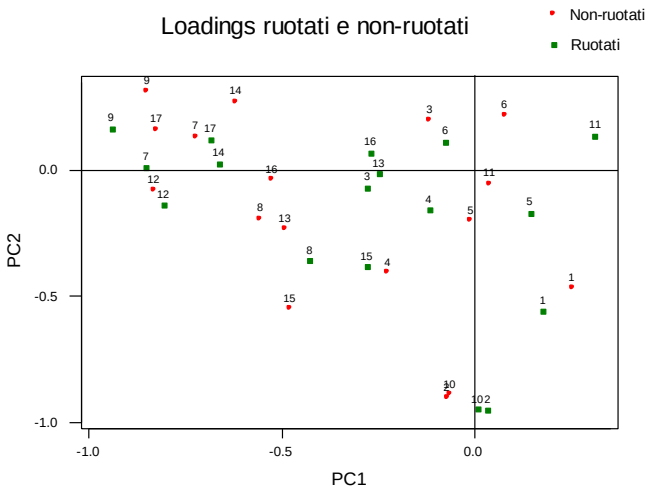


FIG. 3-11

Nel grafico di Fig. 3-10 sono riportati, per tutte le 17 componenti principali (in ascisse), gli scores di tutti i 38 oggetti (in ordinate). Si può osservare in questo grafico il tipico andamento a cuneo della soluzione in componenti principali: cioè, la progressiva diminuzione della varianza spiegata da ciascuna componente, partendo dalla prima e arrivando all'ultima. Si può anche osservare che la varianza di qualche componente rimane relativamente alta solo per la presenza di qualche campione il cui comportamento (in quella componente) si distacca in parte da tutti gli altri.
Nelle Fig.3-11 e 3-12 sono riportati i grafici dei *loadings* e degli *scores* per le prime due componenti principali, confrontando i rispettivi valori prima e dopo la rotazione. In questo caso è stata effettuata la rotazione ortogonale *raw varimax* sulle prime 6 componenti principali.

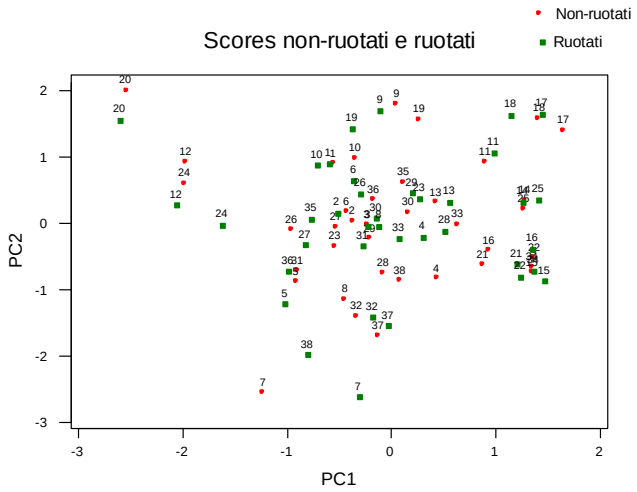


FIG. 3-12

🕮 BIBLIOGRAFIA

J.E. JACKSON (1991). *A User's Guide to Principal Components.* Wiley, N.Y.

I.T. JOLLIFFE (1986). *Principal Component Analysis.* Springer-Verlag, N.Y.

A. BASILEVSKY (1994). *Statistical Factor Analysis and Related Methods.* Wiley, N.Y.

W.J. KRZANOWSKI (1988). *Principles of Multivariate Analysis. A User's Perspective*. Oxford Univ. Press, Oxford.

R.J. RUMMEL (1970). *Applied Factor Analysis.* Northwestern Univ. Press, Evanston (IL)

R. TODESCHINI (1997). *Data correlation, number of significant principal components and shape of molecules. The K correlation index***.** Analytica Chim. Acta. 348, 419-430.

4

L'ANALISI DEI *CLUSTER*

4.1 - Introduzione

Esaminando la Fig.4-1, è legittimo porsi la seguente domanda: quanti gruppi distinti riconosciamo nella figura?

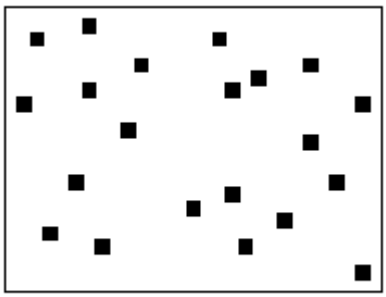


FIG. 4-1

Se persone diverse dessero una risposta a questa domanda, sicuramente otterremmo un certo numero di risposte diverse: alcuni risponderebbero 2, altri 3, 4 o più, altri uno solo. Questo risultato ambiguo non deve sorprendere poichè il problema non ha, in generale e in linea di principio, un'unica soluzione. Il problema della ricerca di gruppi è quindi un problema che, pur nella sua semplicità concettuale, risulta non banale in quanto non ha di norma una risposta unica, ma dipende sia dal soggetto che affronta il problema, sia dagli obiettivi stessi che il soggetto si pone al riguardo.
L'analisi di problemi di questo tipo è nota come ***cluster analysis***, o ricerca di gruppi.

Il principio guida nella ricerca di informazione utile è, in generale, quello di ricercare la presenza nei dati di *strutture non casuali*. Questo obiettivo è comunemente perseguito, in questo caso, associando il concetto di struttura non casuale a quello di raggruppamento e ricercando la presenza di raggruppamenti nello spazio dei dati, in contrapposizione all'ipotesi di completa omogeneità dei dati (isotropia). I metodi di *cluster analysis* forniscono possibili risposte sulla presenza di raggruppamenti (*clusters*) utilizzando il concetto di **similarità**.

Nota. I metodi di *cluster analysis* non devono essere confusi con i **metodi di classificazione**. In questi ultimi i gruppi sono chiamati ***classi*** e preesistono agli oggetti che consideriamo; ciascun oggetto è a-priori assegnato alla sua classe di appartenenza. Nella *cluster analysis* le classi non sono note a-priori: al contrario siamo alla *ricerca dell'esistenza di gruppi*. Se al termine dell'analisi riteniamo di poter dare un significato ai gruppi individuati, allora possiamo definire questi gruppi come classi, con un passaggio puramente concettuale di carattere semantico che eleva i gruppi al rango di classi.

4.2 - Distanze e similarità

Il concetto di similarità è un concetto di grande rilevanza scientifica e pratica. Esso, infatti, è la trasposizione matematica del concetto di **analogia**, concetto che noi utilizziamo in ogni momento della nostra vita nel riconoscere, distinguere, classificare.
Dal punto di vista matematico, il concetto di similarità è il complemento del concetto di dissimilarità. Per misurare quantitativamente quest'ultimo concetto è possibile utilizzare il concetto di distanza: tanto più due oggetti sono "distanti", rispetto ad un quadro di riferimento in cui compaiono molti altri oggetti, tanto più essi sono dissimili. Viceversa, tanto più essi sono "vicini", tanto più possono essere considerati simili.

Alcune tra le distanze più utilizzate in chemiometria sono riportate qui di seguito.

1. $d_{st} = \sqrt{\sum_{j=1}^{p} (x_{sj} - x_{tj})^2}$ distanza Euclidea

2. $d_{st} = \sum_j \left| x_{sj} - x_{tj} \right|$ distanza di Manhattan

3. $d_{st} = \max_j |x_{sj} - x_{tj}|$ distanza di Chebyshev o Lagrange

4. $d_{st} = \sum_{j=1}^{p} \frac{|x_{sj} - x_{tj}|}{(x_{sj} + x_{tj})}$ distanza di Camberra

5. $d_{st} = \frac{\sum_j \left| x_{sj} - x_{tj} \right|}{\sum_j \left(x_{sj} + x_{tj} \right)}$ distanza di Lance-Williams

6. $d_{st} = \sqrt[r]{\sum_j \left| x_{sj} - x_{tj} \right|^r}$ distanza di Minkowski

7. $d_{st} = \sqrt{\left(\mathbf{x_s} - \mathbf{x_t}\right)^{\mathbf{T}} \mathbf{S}^{-1} \left(\mathbf{x_s} - \mathbf{x_t}\right)}$ distanza di Mahalanobis

8. $d_{st} = \sqrt{\sum_j \frac{\left(x_{sj} - x_{tj} \right)^2}{s_j^2}}$ distanza di Pearson

dove $\mathbf{S}^{-1}$ è l'inversa della matrice di covarianza e s_j^2 è la varianza della *j*-esima variabile.
La distanza euclidea e la distanza di Manhattan sono casi particolari della distanza di Minkowski, per $r = 2$ e $r = 1$, rispettivamente. La distanza di Camberra può assumere come valore massimo *p*, il numero di variabili, poichè

ciascun elemento della sommatoria è compreso tra 0 e 1. Nella distanza di Mahalanobis, l'effetto dell'inverso della matrice di covarianza sulle distanze è quello di comprimere le distanze tra punti definiti in uno spazio di variabili tra loro correlate e diminuisce il peso dovuto a variabili ad alta varianza. Di quest'ultimo effetto tiene conto anche la distanza di Pearson.

Nota. Una **misura di distanza** tra due oggetti *s* e *t* può soddisfare una o più delle seguenti condizioni:

1. $d_{st} \geq 0$
2. $d_{ss} = 0$
3. $d_{st} = d_{ts}$
4. $d_{st} = 0 \quad se \quad s = t$
5. $d_{st} \leq d_{sz} + d_{zt}$

5*a*. $d_{st} = d_{sz} + d_{zt}$

6. $d_{st} + d_{uz} \leq max\ (d_{su} + d_{tz}), (d_{sz} + d_{tu})$

6*a*. $d_{st} + d_{uz} = max\ (d_{su} + d_{tz}), (d_{sz} + d_{tu})$

7. $d_{st} \leq max\ d_{sz}, d_{tz}$
8. $d_{st}^2 = (\mathbf{x}_s - \mathbf{x}_t)^{\mathrm{T}} (\mathbf{x}_s - \mathbf{x}_t)$

La proprietà 5 è detta *diseguaglianza triangolare*; la proprietà 6 è detta *diseguaglianza additiva* (o anche *four-point condition*), la proprietà 7 è detta *diseguaglianza ultrametrica*. Se vale la proprietà 7, allora valgono anche le proprietà 4, 5, 6 e 8; se vale la proprietà 6, allora vale anche la proprietà 5; se vale la proprietà 8, allora vale anche la proprietà 4. In base alle diverse proprietà soddisfatte, le distanze si possono classificare in base alle condizioni presentate nella Tab.4-1.

nome	*proprietà*							
misura di prossimità	1							
pseudo-distanza	1	2	3					
misura di dissimilarità	1	2	3	4				
distanza metrica	1	2	3	4	5			
distanza additiva	1	2	3	4		6		
distanza ultrametrica	1	2	3	4			7	
distanza centroide	1	2	3	4	5a	6a		
distanza euclidea	1	2	3	4				8

TAB. 4-1

Nel caso in cui i dati non siano espressi da variabili quantitative reali, ma da variabili dicotomiche (**variabili binarie**, 0 e 1), le precedenti formule per il calcolo della distanza non sono appropriate ed è quindi necessario ricorrere ad espressioni diverse.
Le distanze per dati binari vengono calcolate in accordo con la Tab. 4-2.

s / t	1	0	
1	*a*	*b*	*a* + *b*
0	*c*	*d*	*c* + *d*
	a + *c*	*b* + *d*	*p*

TAB. 4-2

dove le righe 1 e 0 rappresentano i valori delle variabili binarie del campione *s* e le colonne 1 e 0 rappresentano quelli del campione *t*. Gli elementi *a*, *b*, *c* e *d* sono le occorrenze delle quattro possibilità nel confronto tra i due campioni rappresentati da *p* variabili binarie ($p = a + b + c + d$).
Ad esempio, se il campione *s* è definito dai valori [1 1 0 1 0 0 0 0] di 8 variabili binarie ed il campione *t* è definito dai valori [1 0 0 1 0 1 0 1], allora i valori dei 4 parametri sono: $a = 2$, $b = 1$, $c = 2$, $d = 3$.

Le distanze più utilizzate per dati binari sono:

$d_{st} = b + c$ o $d_{st} = \sqrt{b + c}$ distanza di Hamming

$$d_{st} = \frac{b + c}{a + b + c + d} \quad \text{o} \quad d_{st} = \sqrt{\frac{b + c}{a + b + c + d}}$$ distanza di Tanimoto

La prima espressione della distanza di Hamming (o coefficiente di Hamming) è la versione binaria della distanza di Manhattan, mentre la seconda è la versione binaria della distanza euclidea. Le due distanze di Tanimoto (o coefficienti di Tanimoto) sono le rispettive distanze di Hamming normalizzate.
Di norma, una misura di distanza può essere trasformata in una **misura di similarità** in base a semplici trasformazioni, quali, ad esempio :

$$s_{st} = 1 - \frac{d_{st}}{d_{\max}} \quad \text{o} \quad s_{st} = \frac{1}{1 + d_{st}}$$

dove d_{st} è la distanza tra gli oggetti *s* e *t*, misurata sulla base di *p* variabili, secondo una metrica definita (ad esempio, distanza euclidea, distanza di Manhattan, distanza di Camberra, ecc.), e d_{max} è la distanza massima tra gli oggetti considerati o una distanza assunta come riferimento e maggiore di tutte le altre distanze.
Nel caso in cui una misura di distanza sia normalizzabile, cioè $0 \le d_{st} \le 1$, una misura naturale di similarità è $s_{st} = 1 - d_{st}$. E' questo il caso, ad esempio, della distanza media di Camberra, cioè quando la distanza di Camberra viene divisa per il numero di variabili *p*.
In generale, una misura di similarità tra due oggetti *s* e *t* ha le seguenti proprietà:

$$0 \le s_{st} \le 1$$
$$s_{ss} = 1$$
$$s_{st} = s_{ts}$$
$$s_{st} = 1 \quad se \quad s = t$$

In alcuni casi, è importante valutare la similarità globale tra due campioni mediante misure di similarità basate su diversi criteri, cioè gli stessi campioni sono descritti da gruppi diversi di variabili. Quindi, se per ciascun gruppo è possibile valutare la similarità tra due campioni, una misura globale di similarità può essere definita in analogia con quanto viene fatto per la definizione delle funzioni di desiderabilità (*I criteri multipli di decisione*, v. Cap. 11), cioè come media geometrica delle singole misure di similarità:

$$s_{st} = \sqrt[c]{s_{st,1} \cdot s_{st,2} \cdot \ldots \cdot s_{st,c}}$$

dove c è il numero di criteri (o gruppi di variabili), $s_{st,k}$ è la singola k-esima misura di similarità tra gli oggetti s e t con il criterio k. Si osservi che le proprietà richieste per una misura di similarità rimangono valide e che è sufficiente che una singola misura di similarità sia uguale a zero per rendere nulla la similarità globale.

Nello sviluppo del loro algoritmo di calcolo, molti metodi di *cluster analysis* utilizzano inizialmente la **matrice delle distanze**.
La *matrice delle distanze* è definita come la matrice, di dimensione $n \times n$ (dove n è il numero di oggetti), in cui in ogni riga (un oggetto) vi sono tutte le distanze degli altri oggetti da quello considerato (gli elementi diagonali di questa matrice sono uguali a zero, cioè la distanza di ogni oggetto da se stesso è nulla). Essa è, ovviamente, una matrice simmetrica. Una volta definita una corrispondenza tra distanze e similarità, l'algoritmo utilizzato da un particolare metodo può richiedere la trasformazione di una matrice delle distanze in una **matrice di similarità**. Anch'essa è una matrice quadrata simmetrica, i cui elementi diagonali sono uguali ad uno (cioè la similarità di ciascun oggetto con se stesso è massima).
Ovviamente, la misura della similarità (o della dissimilarità) è una misura relativa all'insieme degli oggetti che si prendono in considerazione e dipende in modo decisivo dai parametri con cui si confrontano gli oggetti stessi. Utilizzando parametri differenti si possono ottenere risultati molto diversi tra loro.
Inoltre, si deve tener ben presente che i risultati ottenuti con i metodi di *cluster analysis* che utilizzano le distanze sono *non invarianti alle scalature* dei dati che modificano la loro matrice di varianza/covarianza (autoscalatura, scalatura di intervallo, eccetera).

4.3 - I metodi di *cluster analysis*

Anche se i metodi *cluster analysis* sono molto numerosi, possiamo dividerli comunemente in due grandi categorie: i **metodi gerarchici** e i **metodi non-gerarchici**.
I primi includono metodi come *single linkage*, *average linkage*, *complete linkage*, mentre gli altri includono metodi le cui strategie sono molto più differenziate tra loro, quali, ad esempio, i metodi *Jarvis-Patrick* e *k-means*.
I clusters che ciascun metodo individua sono caratterizzati dalla loro *posizione* nello spazio *p*-dimensionale da un **centroide**, definito come il vettore delle medie delle variabili calcolate per gli oggetti assegnati al cluster, o da un **centrotipo**, definito come l'oggetto più rappresentativo tra gli oggetti assegnati al cluster (di norma, il più vicino al centroide). A differenza del centroide, il centrotipo è sempre un oggetto presente nei dati.

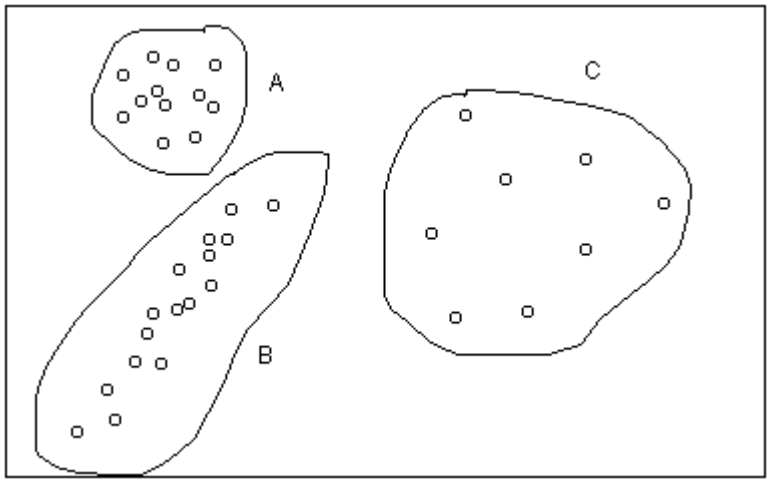


FIG. 4-2

Inoltre ciascun cluster è caratterizzato dalla propria *dimensionalità* (il numero di oggetti che contiene), dalla *compattezza* degli oggetti intorno al centroide (definita dalla deviazione standard rispetto al centroide), dalla propria forma nello spazio *p*-dimensionale.

In Fig. 4-2 sono mostrati tre clusters di diverso tipo. Il primo (A) è un cluster sferico e compatto; il secondo (B) è un cluster compatto dalla forma allungata ed il terzo (C) è un cluster costituito da alcuni oggetti sparsi.

In generale, la maggior parte dei metodi di *cluster analysis* si sviluppano secondo la procedura di Fig. 4-3. Dopo aver selezionato un tipo di distanza (ad esempio, la distanza euclidea), viene calcolata la matrice delle distanze e da questa, eventualmente, la matrice di similarità. Applicando l'algoritmo di *clustering*, si ottiene la partizione finale degli oggetti in clusters. La possibile interpretazione di ciascun *cluster* porta alla identificazione di classi.

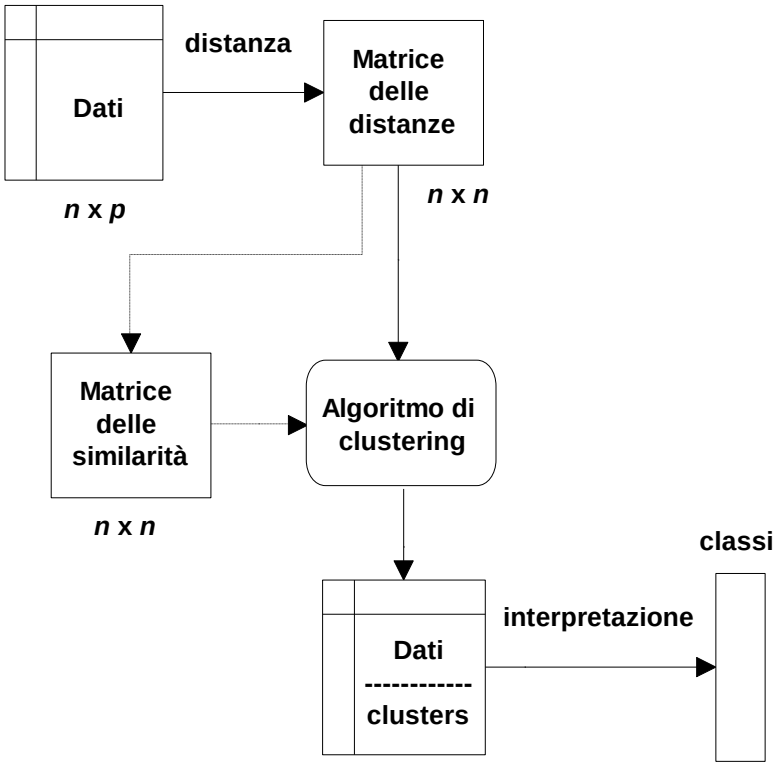


Fig. 4-3

4.4 - I metodi gerarchici

I metodi gerarchici si suddividono in due grandi categorie: i **metodi gerarchici divisivi** e i **metodi gerarchici agglomerativi**.
Il primo gruppo di metodi (poco usati in pratica per diversi motivi) si basa su strategie che partono da un insieme che comprende tutti i dati iniziali e separano via via i dati che differiscono maggiormente dagli altri. Al contrario, i metodi gerarchici agglomerativi sono i più utilizzati e partono da un numero di clusters pari al numero di oggetti, procedendo via via alla loro fusione in clusters di dimensione sempre maggiore.

🗀 I metodi gerarchici agglomerativi

I metodi agglomerativi richiedono di norma i seguenti passi preliminari:

a. definizione della metrica da utilizzare

b. calcolo della matrice delle distanze tra gli oggetti

c. calcolo della corrispondente matrice di similarità

Una volta calcolata la matrice di similarità, la matrice viene analizzata e ridotta secondo i seguenti criteri:

a. vengono individuati i due cluster più simili (nella prima fase, gli oggetti più simili);

b. i due cluster (o i due oggetti) vengono uniti in un unico nuovo cluster, ad un dato livello di similitudine;

c. per il nuovo cluster viene calcolato il suo livello di similitudine rispetto ai cluster (o oggetti) restanti, secondo criteri che differiscono da metodo a metodo. Questa operazione consiste nell'eliminare dalla matrice di similarità le righe e le colonne relative ai due cluster (o oggetti) che vengono posti insieme ed aggiungere una riga e una colonna relativa alle similarità del nuovo cluster con tutti i restanti cluster (o oggetti).

I metodi di cluster gerarchico agglomerativo sono noti con i seguenti nomi:

1. *Weighted Average Linkage*
2. *Unweighted Average Linkage*
3. *Single Linkage*
4. *Complete Linkage*
5. *Centroid Linkage*
6. *Median Linkage*
7. *Ward method*

Indicando con *f* il cluster ottenuto dalla fusione del cluster *s* con il cluster *t*, di dimensioni rispettivamente n_s , n_t e $n_f = n_s + n_t$, la similarità del cluster *f* con un altro generico cluster *k* viene calcolata con le seguenti regole:

1. $s_{kf} = 0.5 \cdot (s_{ks} + s_{kt})$

2. $s_{kf} = \frac{n_s s_{ks} + n_t s_{kt}}{n_f}$

3. $s_{kf} = max(s_{ks}, s_{kt})$

4. $s_{kf} = min(s_{ks}, s_{kt})$

5. $s_{kf} = \frac{n_s s_{ks}}{n_f} + \frac{n_t s_{kt}}{n_f} - \frac{n_s n_t s_{st}}{n_f^2}$

6. $s_{kf} = 0.5 \cdot (s_{ks} + s_{kt}) - 0.25 \cdot s_{st}$

7. $s_{kf} = \frac{(n_s + n_k) s_{ks}}{n_f + n_k} + \frac{(n_t + n_k) s_{kt}}{n_f + n_k} - \frac{n_k s_{st}}{n_f + n_k}$

Le operazioni definite dai 7 algoritmi precedenti possono essere eseguite utilizzando le distanze al posto delle similarità. In questo caso, per i metodi 3 e 4 (*single* e *complete linkage*) si deve scambiare il massimo con il minimo.
Nella Fig.4-4 si può osservare che il metodo *single linkage* fa coincidere la distanza tra i due cluster con la distanza minima tra gli oggetti; al contrario, con il metodo *complete linkage* si attribuisce ai due cluster una distanza eguale a quella massima tra gli oggetti; per il metodo *centroid linkage* si attribuisce ai due cluster una distanza eguale a quella dei loro centroidi.

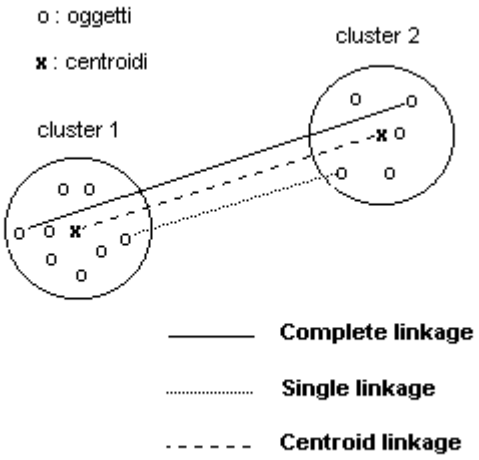


FIG. 4-4

Questo tipo di procedura parte quindi dall'analisi della matrice di similarità (simmetrica, di dimensione uguale al numero di dati *n*). Dopo ogni fusione di due cluster, vengono eliminate le colonne (e le righe) relative ai cluster *s* e *t* che si uniscono e viene aggiunta una colonna (e una riga) relativa alle similarità del nuovo cluster con tutti gli altri restanti. Così facendo, la dimensione della matrice di similarità si riduce di uno ad ogni passo.
Le formule 1 - 7 costituiscono le diverse procedure per aggiornare la matrice di similarità, calcolando la similarità tra il nuovo cluster e tutti i restanti.

Il risultato di questa procedura viene normalmente rappresentato mediante un grafico chiamato **dendrogramma**, che permette un'analisi visiva altamente informativa della gerarchia delle similarità tra gli oggetti considerati.
Un esempio grafico del risultato ottenuto applicando un metodo di *clustering* gerarchico agglomerativo è riportato in Fig.4-5.
Sull'asse delle ascisse sono riportati i diversi campioni (rappresentati dalle lettere maiuscole da A ad J), ciascuno descritto da un certo numero di variabili. Sull'asse delle ordinate sono riportati i valori di similarità (compresi tra 0, in alto, e 1, in basso).

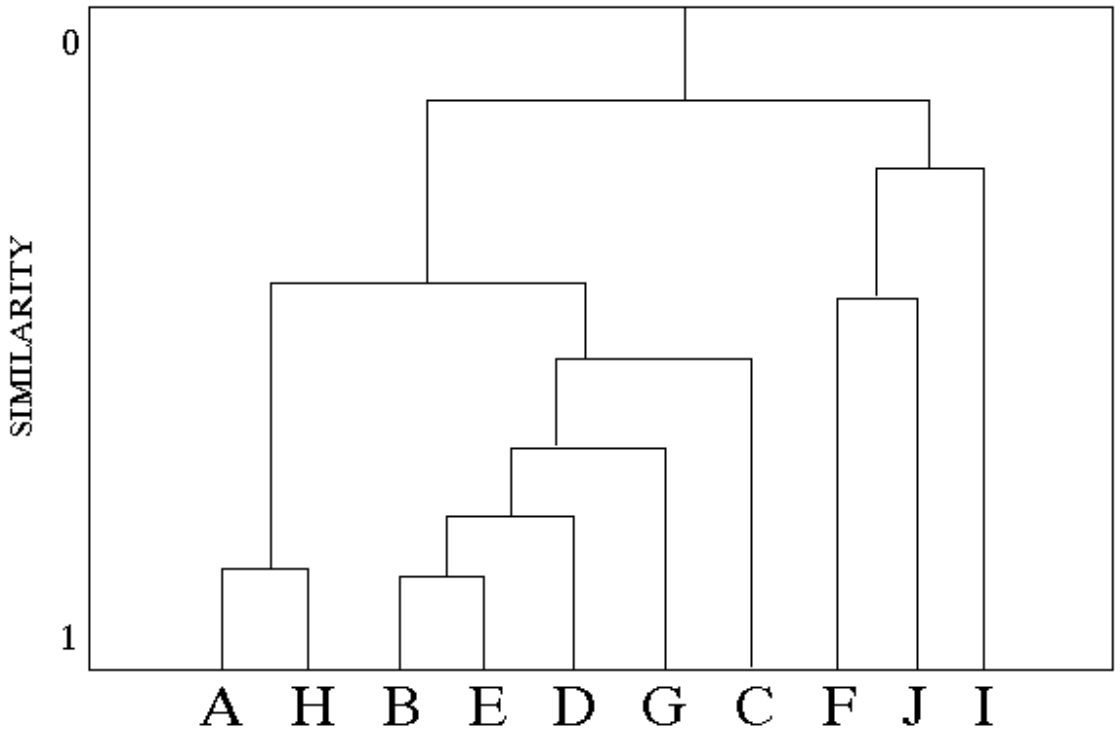


FIG. 4-5

Esaminando il grafico dal basso verso l'alto, si possono osservare le coppie di campioni che sono più simili tra loro:

1) i campioni B ed E sono i più simili tra loro perchè si uniscono per primi;
2) successivamente si uniscono i campioni A e H;
3) successivamente D si unisce col cluster costituito dai campioni B-E;
4) l'ultima fusione riguarda i due cluster [A,H,B,E,D,G,C] e [F,J,I].

Se selezioniamo un certo livello di similarità (ad esempio 0.5) possiamo distinguere i seguenti 5 cluster:

1. [A,H] ; 2. [B,E,D,G,C]; 3. [F]; 4. [J]; 5. [I]

I cluster 3, 4 e 5 sono detti *singletons*, contengono cioè un solo elemento.
Utilizzando metodi diversi, si ottengono raggruppamenti anche significativamente differenti tra loro. Chiaramente, la scelta di un certo livello di similarità conserva quel carattere di soggettività tipico di tutti i metodi di *cluster analysis*.

Esempio 1

Supponiamo di voler studiare il comportamento di 5 diversi metodi analitici (A,B,C,D,E) mediante misure su 4 diversi standard (1, 2, 3, 4). Vogliamo misurare la similarità tra i diversi metodi analitici, assumendo quindi che le caratteristiche di ciascun metodo siano ben rappresentate dalle 4 misure effettuate.
I risultati sperimentali sono i seguenti:

	1	2	3	4
A	100	80	70	60
B	80	60	50	40
C	80	70	40	50
D	40	20	20	10
E	50	10	20	10

Le distanze euclidee (al quadrato) tra i campioni sono:

	A	B	C	D	E
A	0				
B	40.0	0			
C	38.7	17.3	0		
D	110.4	70.7	78.1	0	
E	111.4	72.1	80.6	14.1	0

Utilizziamo il metodo gerarchico agglomerativo *Weighted Average Linkage*, le cui formule per il ricalcolo della similarità o della distanza sono le seguenti:

$$s_{kf} = 0.5 \times (s_{ks} + s_{kt}) \quad \text{ovvero} \quad d_{kf} = 0.5 \times (d_{ks} + d_{kt})$$

Dall'analisi della tabella precedente risulta che gli oggetti D ed E sono i più simili, avendo la minima distanza di 14.1. Eliminiamo quindi le colonne relative ai campioni D ed E e aggiungiamo una colonna relativa al cluster D^* (per convenzione, al cluster ottenuto si assegna l'etichetta inferiore):

	A	B	C	D^*
A	0			
B	40.0	0		
C	38.7	17.3	0	
D^*	110.9	71.4	79.3	0

Ad esempio, la distanza D^*-A viene calcolata dalla tabella precedente applicando la formula:

$$d_{D^*A} = 0.5 \times (110.4 + 111.4) = 110.9$$

Analogamente per tutte le altre distanze tra D^* e gli altri dati B e C.
Dall'analisi della nuova tabella risulta che gli oggetti B e C sono i più simili, con una distanza di 17.3. Il nuovo cluster *f* ottenuto unendo B e C si chiamerà B^*.

	A	B^*	D^*
A	0		
B^*	39.3	0	
D^*	110.9	75.3	0

Gli oggetti A e B^* sono i più simili.

	A^*	D^*
A^*	0	
D^*	93.1	0

Il dendrogramma finale è riportato in Fig.4-6.

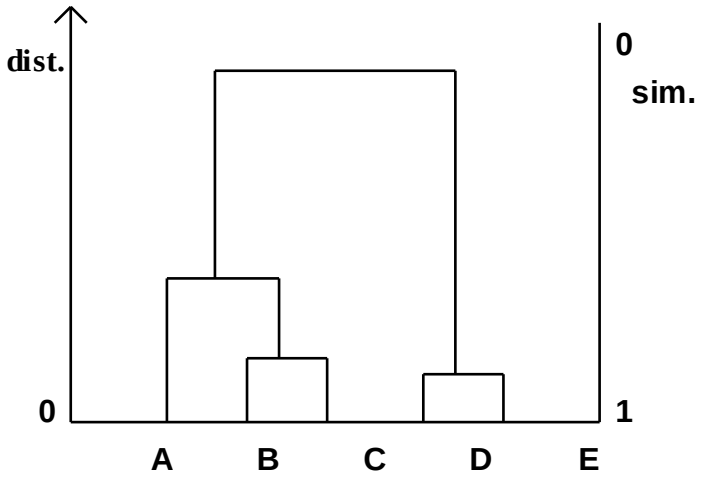


FIG. 4-6

Metodo gerarchico divisivo di Mc Naughton

In questo metodo divisivo, inizialmente l'insieme degli oggetti viene suddiviso in due sottoinsiemi, che a loro volta successivamente sono suddivisi in due sottoinsiemi, fino a che ciascun sottoinsieme contiene un solo oggetto.
Ad ogni passo, viene selezionato l'oggetto più dissimile da tutti gli altri e sarà quello la cui somma delle distanze da tutti gli altri è massima. Isolato il primo oggetto, si effettua il confronto tra la distanza media calcolata tra ciascuno degli oggetti dell'insieme rimasto e la distanza di ciascuno di questi oggetti con l'oggetto isolato. Viene così individuato l'oggetto dell'insieme ancora indiviso più simile all'oggetto (meno distante da) precedentemente isolato.
La procedura prosegue confrontando le distanze tra gli oggetti ancora indivisi e le loro distanze medie con gli oggetti già selezionati, fino all'esaurimento dell'insieme di partenza.
La procedura viene ripetuta separatamente su ciascuno dei due sottoinsiemi ottenuti, fino ad ottenere sottoinsiemi contenenti un singolo oggetto.
Nell'esempio precedente, la matrice iniziale delle distanze è la seguente:

	A	B	C	D	E
A	0				
B	40.0	0			
C	38.7	17.3	0		
D	110.4	70.7	78.1	0	
E	111.4	72.1	80.6	14.1	0

La somma delle distanze di ciascun oggetto da tutti gli altri è riportata nella seguente tabella:

	Somma	Totale
A	40.0 + 38.7 + 110.4 + 111.4	300.5
B	40.0 + 17.3 + 70.7 + 72.1	200.1
C	38.7 + 17.3 + 78.1 + 80.6	214.7
D	110.4 + 70.7 + 78.1 + 14.1	273.3
E	111.4 + 72.1 + 80.6 + 14.1	278.2

A è l'oggetto più dissimile dagli altri in quanto ha la distanza massima da tutti gli altri (300.5). L'oggetto A viene isolato dagli altri per primo.
Si confronta la distanza di ciascun oggetto da A col valor medio della distanza di ciascun restante oggetto dai restanti altri.

	A	altri
B	40.0	53.4
C	38.7	58.7
D	110.4	54.3
E	111.4	55.6

L'oggetto C viene assegnato nel cluster di A in quanto è il più simile ad A (distanza 38.7). La tabella precedente viene nuovamente aggiornata con le distanze di B, D ed E da A+C e tra loro:

	A + C	altri
B	28.6	71.4
D	94.3	42.4
E	96.0	43.1

L'oggetto B viene assegnato al cluster contenente A e C (distanza 28.6) e la tabella viene nuovamente aggiornata:

	A+B+C	altro
D	86.4	14.1
E	88.0	14.1

L'oggetto D viene assegnato ad un nuovo cluster (distanza 14.1) insieme all'oggetto E. Il metodo prosegue separatamente su ciascuna coppia di sottoinsiemi ottenuti (in questo caso sui sottoinsiemi [A,B,C] e [D,E]), fino ad una separazione totale degli oggetti. Il risultato finale è rappresentato in Fig.4-7.

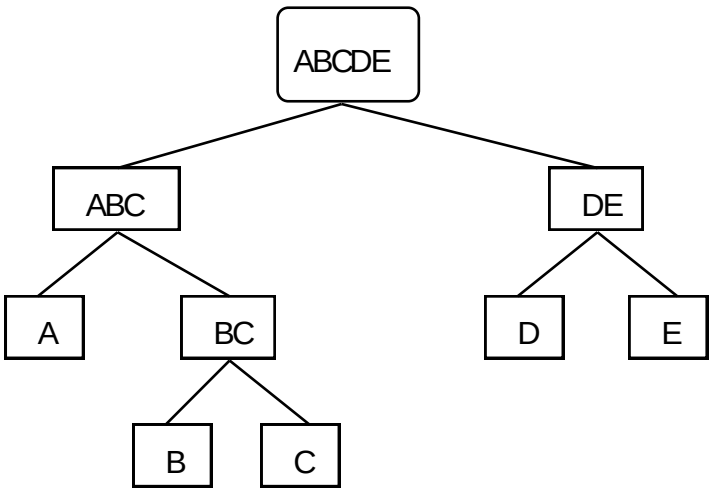


FIG. 4-7

4.5 - I metodi non-gerarchici

I metodi di cluster non-gerarchici si basano su tecniche molte diverse tra loro e quindi difficilmente schematizzabili. Un gruppo di metodi si basa su tecniche generalmente chiamate *tecniche di ricollocamento* (*relocation techniques*), secondo le quali, dopo una partizione iniziale dei dati, questi vengono spostati da un cluster all'altro fino a soddisfare un criterio prefissato. Tra questi metodi, il più noto è senz'altro il metodo *k-means.*

🗀 Il metodo *k-means*

Il **metodo k-means** (MacQueen, 1967) è un metodo il cui algoritmo di ricollocamento dei dati è basato sul confronto delle distanze di ogni oggetto dal centroide di ogni cluster. Il numero di clusters *G* è fissato a priori dall'utente insieme alla metrica da utilizzare.
L'algoritmo è il seguente:

a1 disponi una partizione (casuale) degli oggetti in G clusters

b1 determina il centroide di ogni cluster: $\bar{c}_g = \bar{x}_{g1}\,\bar{x}_{g2}\ldots\bar{x}_{gp}$

b2 calcola la distanza tra ciascun oggetto e ciascun centroide

b3 assegna ciascun oggetto al cluster più vicino

b4 se almeno un oggetto viene spostato in un altro cluster, vai a b1

c1 termina

I nuovi centroidi si calcolano *dopo* aver definito la ricollocazione di *tutti* gli oggetti.
Una variante del metodo prevede il ricalcolo dei centroidi ogniqualvolta un oggetto deve essere ricollocato in un altro cluster. In questo caso si utilizzano formule che consentono l'aggiornamento immediato dei centroidi. I nuovi centroidi ottenuti dopo aver spostato l'oggetto *i* dal cluster **a** al cluster **b** sono dati dalle due espressioni:

$$\overline{x}_{a'j} = \frac{1}{n_a - 1}\left(n_a \overline{x}_{aj} - x_{ij}\right)$$

$$\overline{x}_{b'j} = \frac{1}{n_b + 1}\left(n_b \overline{x}_{bj} + x_{ij}\right)$$

I risultati ottenuti col metodo *k-means* dipendono quindi dalla partizione iniziale, dalla scelta della metrica con cui misurare le distanze oggetti-centroidi, dal numero di cluster selezionato. Nella pratica, tuttavia, la dipendenza dalla partizione iniziale è da considerarsi trascurabile.

🗀 Il metodo di Jarvis-Patrick

Il **metodo di Jarvis-Patrick** è uno dei metodi non-gerarchici più potenti ed è basato sul calcolo di tutte le distanze tra gli oggetti e sull'analisi di una **matrice di intorni** derivata dalla matrice delle distanze, di dimensione (n,L), dove L è un parametro, generalmente fissato tra 20 e 30; per ogni riga, gli elementi di questa matrice rappresentano gli L oggetti più vicini (numeri interi tra 1 e n) a ciascun oggetto della riga.
L'algoritmo è il seguente:

1. selezionare la distanza da utilizzare

2. definire la lunghezza della lista di intorni più vicini L
3. definire il numero k di intorni comuni ($k < L$)
4. calcolo della matrice delle distanze
5. costruzione di una *matrice di intorni*, di dimensione (n,L), in cui ad ogni oggetto viene associata una lista degli L intorni più vicini.

L'algoritmo di costruzione dei differenti *clusters* si basa sull'analisi della matrice degli intorni: due oggetti s e t vengono posti nello stesso cluster se vengono soddisfatte le seguenti condizioni:

a - l'oggetto s compare nella lista degli L più vicini all'oggetto t

b - l'oggetto t compare nella lista degli L più vicini all'oggetto s

c - k oggetti (diversi da s e da t) sono comuni ad entrambe le liste.

Un aumento di k significa richiedere una condizione più restrittiva per porre gli oggetti nello stesso cluster, condizione che comporta un aumento del numero finale di cluster. Una buona scelta dei valori dei parametri L e k può essere effettuata secondo la regola: $L = n / 3$ e $k = n / 4$, dove n è il numero dei dati.
La caratteristica principale di questo metodo è il fatto che il numero di cluster viene calcolato dal metodo stesso e che la loro dimensionalità può essere molto variabile.

Nota. Proprio per la grande variabilità che i risultati ottenuti applicando i metodi di *cluster analysis* possono avere e per la inevitabile soggettività nella valutazione dei risultati, è fondamentale adottare alcuni criteri generali di comportamento. In particolare, se il problema è quello di ristrutturare i nostri dati assegnando a ciascun oggetto anche una classe di appartenenza (*non* nota in precedenza), l'aspetto più rilevante è quello di poter interpretare i gruppi risultanti, cioè riuscire a dare loro un significato coerente col problema in esame. Non importa, quindi, che si siano trovati pochi o tanti gruppi: ciò che importa è che sia possibile dar loro un significato.
Un utilizzo meno problematico dei metodi di *cluster analysis* riguarda invece l'esigenza di ridurre l'eventuale ridondanza di informazione presente nei dati.

Ciò accade quando abbiamo un elevato numero di oggetti, molti dei quali rappresentano situazioni molto simili tra loro: in questo caso le diverse situazioni sono rappresentate in modo sbilanciato e l'informazione legata alle situazioni meno rappresentate potrebbe essere mascherata dalle altre. In situazioni di questo tipo, non è tanto importante interpretare i gruppi che risultano dall'applicazione di una tecnica di *cluster analysis*, quanto considerare un numero di gruppi sufficientemente grande da campionare tutto lo spazio nel modo più rappresentativo ed esaustivo possibile.

Esempio 2

Per i dati WINES, 38 campioni di vini rappresentati da 17 variabili analitiche, è stato utilizzato il metodo di cluster gerarchico *complete linkage* (Fig.4-8). Si è richiesto di evidenziare 3 clusters.

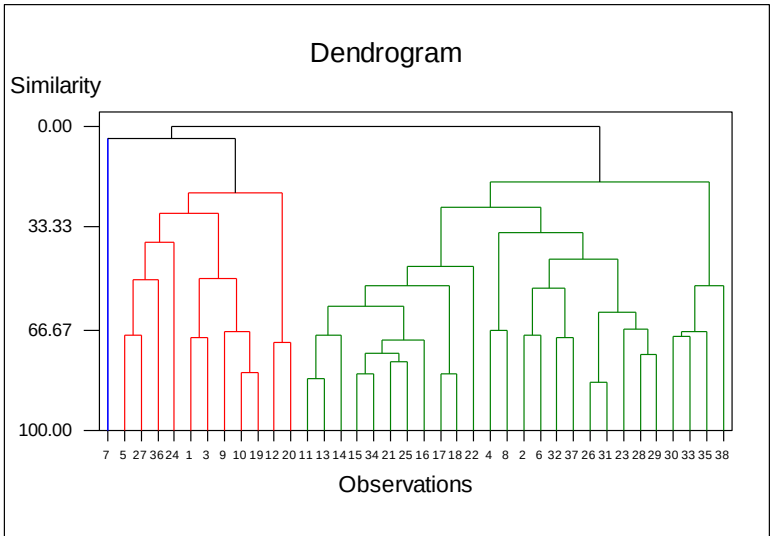


FIG. 4-8

Come si può osservare dal dendrogramma di Fig.4-8, l'oggetto 7 è un *singleton* (come appariva anche dall'analisi delle componenti principali). Tuttavia in questo caso, la similarità viene misurata conservando tutta l'informazione

originale (le 17 variabili originali). Questa informazione può contenere anche il rumore sperimentale e quindi due oggetti che hanno lo stesso andamento del rumore sperimentale possono apparire più simili di quello che sono in realtà o viceversa.
Un dendogramma più aderente al grafico degli *scores* ottenuto dall'analisi delle componenti principali (Cap.3 - Fig.3-9) con le prime due componenti principali si ottiene utilizzando come variabili gli scores delle 2 componenti. In questo caso la similarità misurata è legata al significato delle prime due componenti principali, come si può vedere dalla Fig.4-9. Si osservi, ad esempio, il cluster costituito dai campioni 12, 20 e 24 che anche nel grafico degli scores appaiono isolati (in alto a sinistra). Ritornando al caso precedente, con tutte le variabili originali, la Fig.4-10 mostra i profili dei centroidi di ciascuno dei 3 clusters, cioè i valori delle 17 variabili originali mediati sugli oggetti assegnati a ciascun cluster; ovviamente i valori delle variabili relative al terzo cluster coincidono con i valori delle variabili dell'oggetto 7. Questo tipo di grafico rappresenta l'andamento medio delle variabili di ciascun cluster ed è particolarmente utile nell'interpretazione dei clusters.

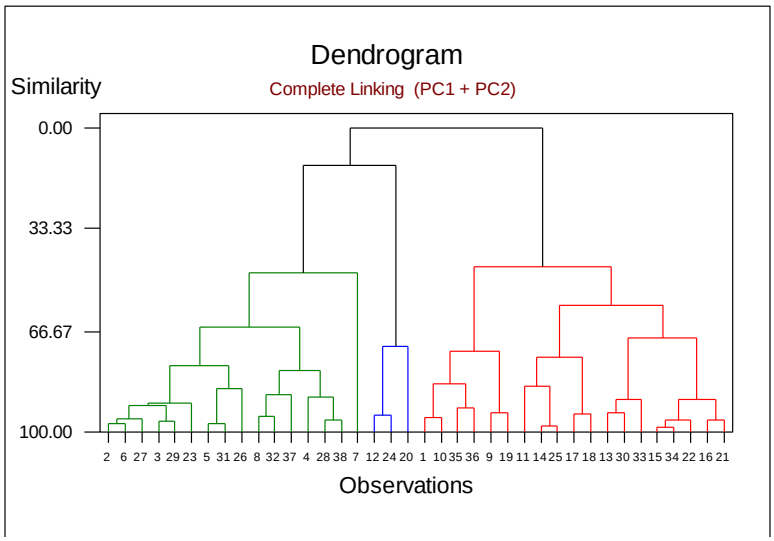


FIG. 4-9

La Fig.4-11 rappresenta le distanze degli oggetti da ciascuna coppia di cluster: gli assi rappresentano le distanze dai centroidi dei cluster corrispondenti.

Gli oggetti sulla diagonale principale sono oggetti che distano in egual modo da entrambi i clusters; gli oggetti vicini all'asse delle ascisse (in basso) sono i più vicini al cluster descritto dalle ordinate, mentre gli oggetti più vicini all'asse delle ordinate (a sinistra) sono i più vicini al cluster descritto dalle ascisse. Quindi, oggetti vicini all'origine degli assi sono oggetti vicini ad entrambi i cluster; oggetti in basso all'estrema destra sono oggetti che appartengono decisamente al cluster rappresentato dalle ordinate (quindi anche lontani dal cluster rappresentato dalle ascisse); viceversa per gli oggetti in alto all'estrema sinistra.

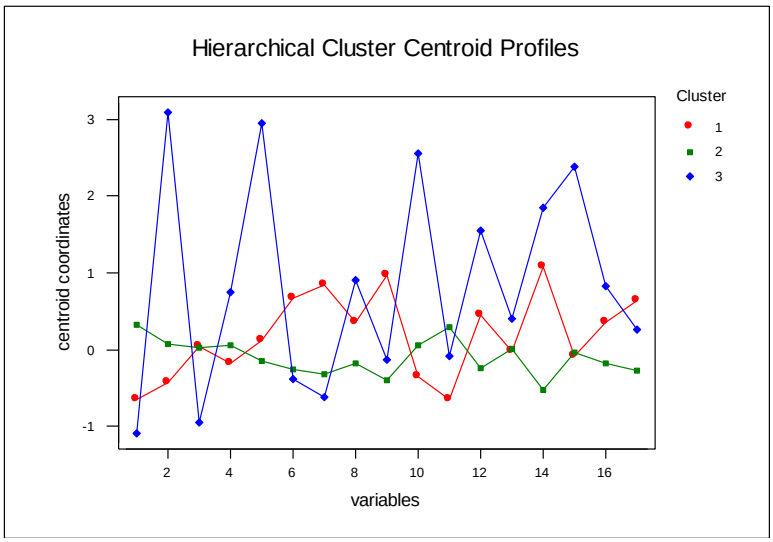


FIG. 4-10

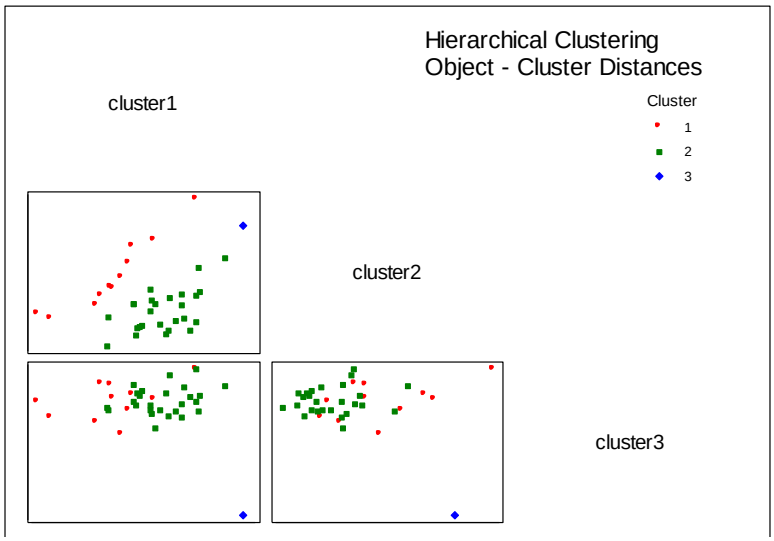


FIG. 4-11

4.6 - La similarità mediante componenti principali.

Le componenti principali possono essere utilizzate vantaggiosamente nell'analisi dei cluster, sostituendo le variabili originali $\boldsymbol{x_j}$ (j = 1, p) con le prime M componenti significative $\boldsymbol{t_m}$ (m = 1, M).
Poichè le componenti principali preservano la distanza euclidea tra gli oggetti, l'utilizzo di tutte le componenti principali equivale ad utilizzare una distanza euclidea calcolata su tutte le variabili originali. Il reale vantaggio nell'utilizzo delle componenti principali nei metodi di *cluster analysis* si ottiene invece ricercando il numero $M < p$ di componenti significative e misurando la distanza euclidea tra gli oggetti nel sottospazio M-dimensionale. In questo caso, il vantaggio è duplice in quanto nella misura della distanza si elimina sia il contributo dovuto alla ridondanza delle variabili (tutte le variabili che portano la stessa informazione sono rappresentate dalla stessa componente) sia il contributo dovuto ad informazione spuria e rumore contenuto nelle componenti a bassa varianza. Inoltre, poichè l'informazione viene ridistribuita in componenti principali tra loro ortogonali è possibile modulare la misura di similarità includendo soltanto quelle componenti che portano l'informazione che più interessa.

4.7 - La similarità rispetto ad un riferimento.

Mentre l'analisi dei clusters ha come obiettivo il confronto globale tra tutti i campioni presenti, vi è spesso l'esigenza di un confronto mirato tra un particolare campione assunto come riferimento e tutti gli altri campioni. Per quanto questa informazione sia presente implicitamente nei risultati dell'analisi dei cluster (il cluster a cui appartiene il campione di riferimento ci informa su quali sono gli altri campioni più simili), in molti casi si richiede un vero e proprio ordinamento di tutti i campioni rispetto al campione di riferimento.
In questi casi, è sufficiente definire il campione di riferimento e la metrica da utilizzare: i campioni vengono poi ordinati in modo decrescente in funzione della loro similarità rispetto al riferimento, oppure, direttamente, in modo crescente in funzione della loro distanza dal riferimento.
Ad esempio, questa metodologia risulta particolarmente interessante quando si assume come riferimento il composto biologicamente più attivo e si ricercano i composti ad esso più simili, in base all'assunzione implicita che a composti simili corrispondano simili proprietà chimico-fisiche e biologiche (v. Cap.12).

🕮 Bibliografia

D.L. MASSART E L.KAUFMAN (1983). *The interpretation of analytical chemical data by the use of cluster analysis.* Wiley, N.Y.

P.WILLETT (1987). *Similarity and clustering in chemical information systems.* Research Studies Press, Letchworth, UK

J. ZUPAN (1982). *Clustering of Large Data Sets.* Research Studies Press, Letchworth, UK

I.E.FRANK E R.TODESCHINI (1994). *The data analysis handbook*. Elsevier, Amsterdam, The Netherlands

5

MODELLI, *BIAS* E VALIDAZIONE

5.1 - Il concetto di modello

Prima di affrontare lo studio dei metodi di classificazione e di regressione, presentati nei capitoli successivi, è necessario definire alcuni concetti fondamentali di carattere generale. In particolare è necessario chiarire che cosa si intende per **modello matematico**, definire il concetto di ***bias*** e di **complessità di un modello**, ed infine descrivere i diversi metodi di validazione di un modello.

I modelli hanno un ruolo essenziale nello *sviluppo della conoscenza* in quanto sintetizzano lo stato della conoscenza rispetto ad un problema e possono essere utilizzati per "prevedere" eventi futuri descrivibili mediante i parametri che costituiscono il modello stesso.

La prima ovvia considerazione che si può fare nell'atto di studiare un *sistema* è che in esso vi sia qualcosa da conoscere o, in altri termini, che esso contenga una certa quantità di informazione. La *quantità di informazione* contenuta nel sistema dipende dai modi con cui descriviamo il sistema stesso: la mediazione teorica e linguistica svolge un ruolo determinante sulle potenzialità a nostra disposizione per conoscere il sistema. La *descrizione del sistema* definisce i parametri e le variabili con cui esso viene rappresentato e il contenuto di informazione disponibile è intrinsecamente legato alla *variabilità* dei parametri che lo rappresentano: se tutti i parametri che descrivono il sistema fossero delle costanti non si avrebbe alcun contenuto di informazione. Conoscere il sistema significa quindi trasferire il contenuto di informazione presente nel sistema al soggetto conoscente: i modelli rappresentano uno dei modi con cui avviene il trasferimento di informazione dal sistema al soggetto conoscente (Fig.5-1).

Quanto detto è importante perchè i modelli che possiamo costruire dipendono dalle variabili stesse che descrivono il sistema.

In un'accezione più comune e più restrittiva, un modello matematico è definito da una relazione causale, descritta da un'equazione matematica, tra diverse variabili indipendenti (predittori) e una (e talvolta, più di una) variabile risposta.

In un modello, anche la risposta, cioè la variabile dipendente, può essere quantitativa o qualitativa. Da questo punto di vista, un modello di regressione, un modello di analisi fattoriale o un modello di analisi della varianza (ANOVA) sono esempi di modelli quantitativi, mentre i modelli di classificazione sono un esempio di modelli qualitativi.

FIG. 5-1

Nota. In molti casi si può ancora parlare di modello anche quando non vi sia un'esplicita relazione matematica causale tra le variabili, ma sia definibile la realizzazione di un algoritmo (cioè di un algoritmo ove tutti i passi sono esplicitamente definiti e parametrizzati). Ad esempio, in questo senso si può parlare di un modello in componenti principali ove siano definite le seguenti grandezze, calcolate dai dati del *training set*:

a. i valori dei parametri per l'eventuale scalatura dei dati
b. il numero *M* di componenti principali significative
c. la matrice **L** degli *M* autovettori (matrice dei *loadings*)

In questo caso, un nuovo campione, noti i corrispondenti valori dellle variabili con cui si è definito il modello in componenti principali, può essere proiettato nello *spazio-modello* utilizzando le quantità precedentemente definite.

Il processo di ricerca di un modello si basa su 4 fasi fondamentali: l'identificazione, la realizzazione, la validazione e l'applicazione.

🗀 L'identificazione.

L'identificazione è il processo di ricerca e di scelta di un tipo di modello appropriato per la situazione data e per gli obiettivi prefissati. Non esistono procedure rigorose per identificare il tipo di modello più appropriato; si può dire che esistono due modi opposti: da un lato l'identificazione del modello viene effettuata su basi puramente concettuali e teoriche; dall'altro l'identificazione del modello avviene su basi puramente empiriche. Nella realtà, l'identificazione del modello avviene attraverso un approccio eclettico che oscilla tra i due estremi e che si avvale di conoscenze teoriche ed esperienze pratiche che guidano la scelta verso la tipologia di modelli più appropriata.
Possiamo distinguere i modelli in due grandi classi: i **modelli deterministici** (o strettamente causali) ed i **modelli stocastici** (o statistici). I primi vengono costruiti ricorrendo a ipotesi ed idee a priori sulle connessioni funzionali (fisiche, chimiche, biologiche, eccetera) espresse in forma di equazioni matematiche o sistemi di equazioni simultanee tra le variabili considerate. Modelli di questo tipo contengono soltanto variabili deterministiche, al contrario dei secondi che contengono almeno una variabile stocastica. Per questi ultimi - i modelli stocastici - le eventuali ipotesi a priori sulle connessioni funzionali tra le variabili non si esplicitano in equazioni matematiche o sistemi di equazioni predefinite, ma riguardano al più la forma generale e la tipologia del modello da scegliere. Lo sviluppo del modello statistico viene determinato dal suo adattamento ai dati sperimentali disponibili.

🗀 La realizzazione

Una volta scelto il tipo di modello, la fase della realizzazione consiste nel passare da una forma generale del modello ad una forma numerica specifica (*fitting*) e nello stimare i parametri del modello (*estimation*). In questa fase, così come nella fase successiva di validazione, si cerca anche di individuare quali sono le *fonti di incertezza*, cioè le cause che apportano rumore nei dati, le fonti di errore, le approssimazioni che inevitabilmente dobbiamo introdurre nella forma del modello (linearità, semplicità, arrotondamenti e troncamenti, eccetera).

🗀 La validazione

In senso lato, la validazione è la fase di sviluppo del modello, del controllo (*testing*) e dell'uso controllato del modello (*monitoring*). In questa fase il

modello viene via via modificato col fine di migliorarne le sue capacità predittive, di renderlo stabile a piccole alterazioni dei dati e delle condizioni iniziali, di riadattarlo tenendo conto di ulteriori aspetti che emergono dall'analisi dei dati e del problema stesso.

🗀 L'applicazione

L'applicazione è l'ultimo stadio di una procedura di modellamento. Una volta trovato un modello che ha superato anche la fase di validazione, il modello viene utilizzato per le applicazioni per cui è stato previsto, in modo da predire gli eventi incogniti, una volta noti i parametri, le variabili predittrici e le eventuali condizioni al contorno.

Le fasi, così come sono proposte, sono fasi sequenziali. E' tuttavia chiaro che nella realtà vi è un continuo riflusso anche all'indietro di uno o due stadi, o, se si vuole, il processo è in realtà un processo iterativo nel quale il modello non rimane mai uguale a se stesso, ma si sviluppa utilizzando informazione sempre nuova, via via che questa risulta disponibile: il modello non è mai IL MODELLO, conclusivo e inalterabile, ma è sempre uno strumento interpretativo e applicativo che utilizziamo fino a che questo non sia in qualche aspetto migliorabile.
In generale, cioè senza fare riferimento esplicito a modelli deterministici o statistici, si definisce **ordine di un modello** il valore della potenza più elevata di una variabile indipendente (o predittrice). Un *modello del primo ordine* è quindi un modello in cui gli esponenti di tutte le variabili sono uguali ad uno.
Un modello (di regressione) è un **modello lineare** nella risposta se i coefficienti b_j della combinazione lineare non sono funzione della risposta stessa (cioè, $\partial y / \partial b_j = 0$). Un **modello non lineare** intrinsecamente è un modello per il quale non esiste alcuna trasformazione matematica che lo renda lineare.
Il numero di parti indipendenti di informazione necessario per stimare i parametri di un modello è chiamato **gradi di libertà del modello**.
Un modello è un **modello additivo** se l'insieme dei predittori ha un effetto additivo sulla risposta. Un modello statistico è un **modello *biased*** (v. oltre) se i parametri sono calcolati utilizzando *stimatori biased*. Scopo dei modelli *biased* è quello di trovare una complessità ottimale del modello mediante un compromesso tra bias e varianza.
Un modello è detto **modello annidato** (***nested***) se esso è un caso particolare di un modello più generale che lo contiene.

5.2 - Il concetto di bias

Il concetto di ***bias*** è un concetto statistico di grande rilevanza, la cui conoscenza consente una comprensione più profonda di molti metodi statistici e delle modalità con cui è possibile costruire modelli matematici.
In primo luogo definiamo uno *stimatore*.
La quantità **b** espressa come *regola per calcolare* una stima della grandezza β viene chiamata **stimatore** di β.
Uno stimatore **b** di β viene detto ***unbiased*** se

$$E(\mathbf{b}) = \beta$$

cioè se il *valore medio atteso E* di **b** coincide col valore vero β. Ciò significa che osservazioni ripetute di **b** tratte da campioni diversi hanno un valore medio uguale ad $E(\mathbf{b})$.
Se uno stimatore non è *unbiased*, possiamo definire il ***bias*** dello stimatore come:

$$B(\mathbf{b}) = \beta - E(\mathbf{b})$$

cioè $B(\mathbf{b})$ rappresenta lo scostamento sistematico del valore medio atteso $E(\mathbf{b})$ dello stimatore dal valore vero β.
Il bias è quindi l'**errore sistematico** dello stimatore **b**.
La *qualità* di uno stimatore **b** risulta dal valore minimo del suo **errore medio quadratico** (***Mean Squared Error, MSE***), cioè può essere vista come la combinazione di due effetti:

1. un effetto dovuto alla *varianza* di **b**
2. un effetto dovuto al *bias* di **b**

Una regola di fondamentale importanza afferma che la *varianza* aumenta linearmente con la **complessità del modello** (ad esempio, col numero di variabili considerate), mentre il *bias* diminuisce con l'aumentare della complessità. Possiamo quindi dire che l'obiettivo dei metodi *biased* è quello di ricercare una soluzione (cioè, un modello) che rappresenti un compromesso tra la complessità del modello stesso e la sua variabilità (*bias-variance trade-off*), in modo tale da minimizzare **MSE** (Fig. 5-2). L'errore quadratico medio di uno *stimatore biased* **b** viene definito come:

$$\mathbf{MSE}(\mathbf{b}) = E\ (\mathbf{b}-\beta)^2\ = E\left[\left((\mathbf{b}-E(\mathbf{b}))-(\beta-E(\mathbf{b}))\right)^2\right] =$$

$$= E\ (\mathbf{b}-E(\mathbf{b}))^2\ +(\beta-E(\mathbf{b}))^2 = V(\mathbf{b})+B^2(\mathbf{b})$$

dove il termine di varianza $V(\mathbf{b})$ indica la **precisione** della stima e il termine $B(\mathbf{b})$ indica l'**accuratezza** della stima. Come si può osservare dalla definizione, $V(\mathbf{b})$ è il valore atteso degli scarti quadratici tra i singoli valori stimati **b** ed il loro valore atteso $E(\mathbf{b})$, cioè la varianza dello stimatore **b**.

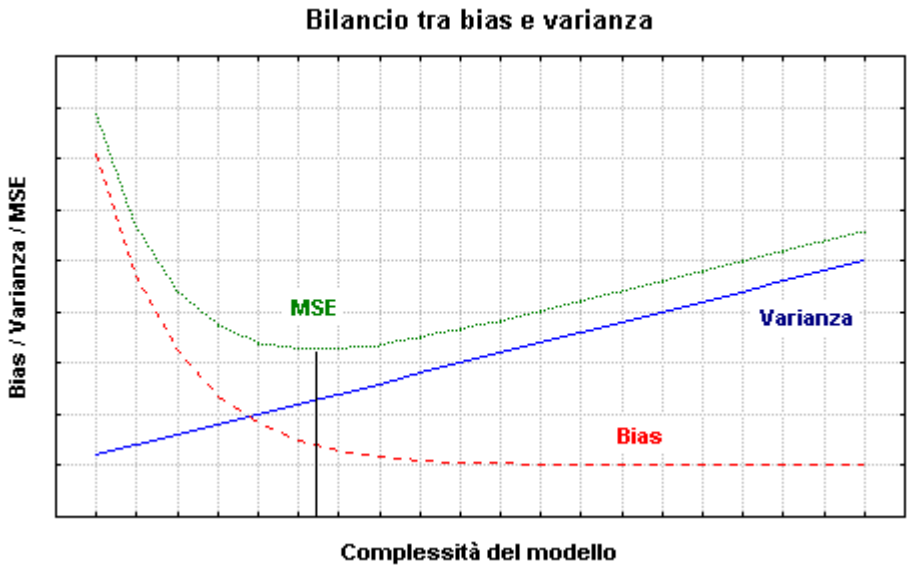


FIG. 5-2

Viene chiamato **errore medio quadratico** la quantità

$$E\ (\mathbf{b}-\beta)^2$$

Se **b** è uno stimatore *unbiased* di β, la vicinanza di **b** a β viene misurata dalla sola varianza di **b**:

$$V(\mathbf{b}) = E\ (\mathbf{b}-\beta)^2\ = E\ \left(\mathbf{b}-E(\mathbf{b})\right)^2$$

essendo $E(\mathbf{b}) = \beta$.
Uno stimatore **b** viene chiamato **unbiased** se $B(\mathbf{b}) = 0$.

Nella Fig.5-3, viene schematicamente trattata l'idea di *bias*. Nella parte sinistra della figura sono presentati le stime ottenute con un metodo *unbiased* (il valore atteso dei parametri $E(\mathbf{b})$ coincide con β). Si può osservare che tuttavia in questo caso le singole stime presentano una notevole variabilità. L'introduzione di un *bias* (parte destra della figura) porta ad un valore atteso dei parametri diverso dal valor vero β, ma le singole stime sono molto più omogenee e stabili (bassa varianza).

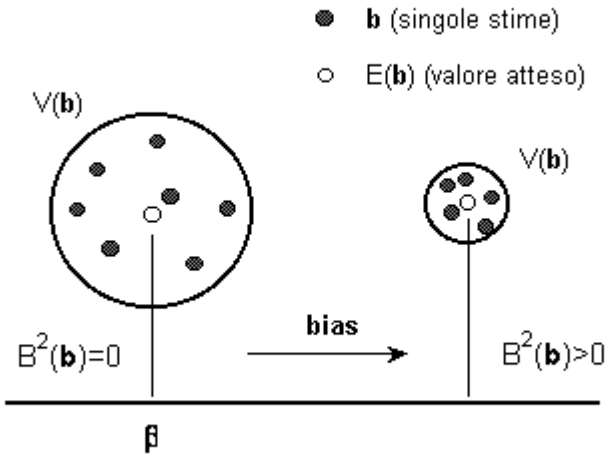


FIG. 5-3

Completiamo questa introduzione al concetto di *bias* con alcune altre importanti definizioni.

Uno stimatore **b** viene detto **asintoticamente unbiased** quando al crescere del numero di osservazioni *n* il suo *bias* tende a zero:

$$B(\mathbf{b}) \to 0 \quad se \quad n \to \infty$$

Uno stimatore **b** viene detto **consistente** se esso tende al valore vero β al crescere del numero di osservazioni *n*:

$$\mathbf{b} \to \beta \quad se \quad n \to \infty$$

Si può dimostrare che all'aumentare del numero di osservazioni *n*, ogni stimatore *unbiased* o *asintoticamente unbiased* tende al valore vero β e la sua varianza tende a zero:

$$Se \;\; n \to \infty \;\; e \;\; B(\mathbf{b}) = 0 \;\; o \;\; B(\mathbf{b}) \to 0$$

$$allora \;\; \mathbf{b} = \beta \;\; e \;\; V(\mathbf{b}) \to 0$$

5.3 - La validazione dei modelli

La validazione di un modello consiste, in generale, nel ricercare quella **struttura del modello** (la sua complessità ottimale) che massimizza la sua capacità predittiva. Nello stesso tempo il modello trovato deve avere delle caratteristiche di **stabilità** che lo rendano sufficientemente indipendente dai dati particolari utilizzati per costruirlo (il campione, il *training set*).
Come si è in precedenza messo in evidenza, la varianza dei parametri stimati che caratterizzano un modello è direttamente proporzionale alla complessità del modello stesso (Fig. 5-2). Questo fatto ha una portata di straordinaria rilevanza nella valutazione delle possibili prestazioni del modello nella fase predittiva: infatti, mentre un aumento della complessità del modello accresce sempre (o non fa diminuire mai) la qualità descrittiva del modello stesso (*fitting*), al contrario, un non controllato aumento della complessità del modello ne deteriora le sue prestazioni in predizione (*overfitting*). In Fig. 5-4, viene mostrato l'andamento della varianza spiegata in *fitting* da un modello di regressione in componenti principali (PCR, v. Cap. 8) rispetto alla varianza spiegata in predizione. Come si può osservare dalla figura, un aumento del numero di componenti principali considerate significative nel modello di

regressione finale comporta un aumento della varianza totale spiegata in *fitting* (R^2). Al contrario, un aumento delle componenti principali considerate comporta un aumento della varianza spiegata in predizione R^2_{cv} solo fino alla quinta componente: questo è il massimo potere predittivo raggiungibile dal modello. Ulteriori aggiunte di componenti comportano un deterioramento del potere predittivo del modello.
Come si può osservare da un certo punto in poi le prestazioni del modello in *fitting* e in predizione divergono nettamente: mentre da un lato riusciamo a descrivere sempre più accuratamente il comportamento dei dati campionari considerati, dall'altro il modello perde in generalità e nella sua capacità di predire il comportamento di campioni nuovi.

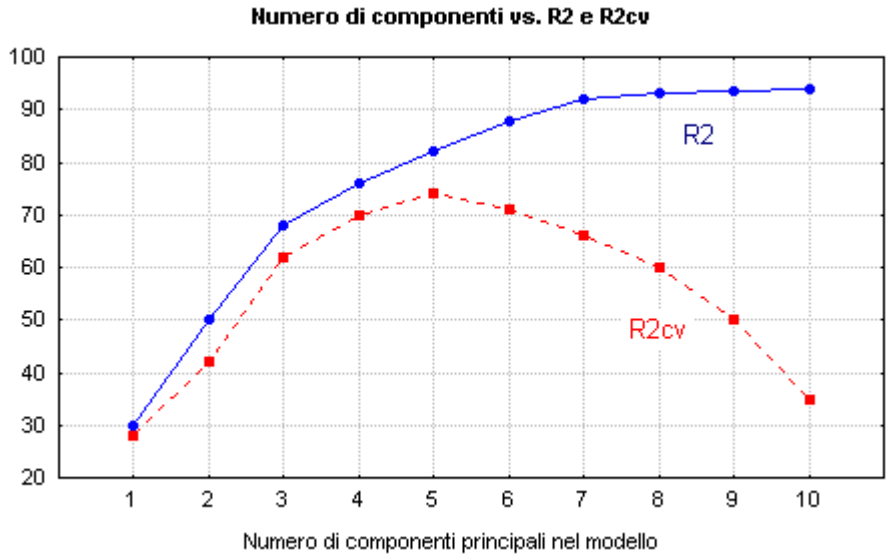


FIG. 5-4

Ciò significa che la struttura del modello (il numero di variabili, il numero di componenti principali, il numero di classi, ecc.) deve essere *sempre* controllata mediante tecniche di validazione che consentano di valutare la presenza di *overfitting*, dovuta a correlazione, a rumore, alle caratteristiche del metodo utilizzato, alla specificità del campione, ad una ingiustificata complessità del

modello, e di ottimizzare il modello per quanto riguarda le sue prestazioni in predizione.
Le procedure di validazione sono fondamentali non solo perchè in molti casi in cui il rapporto oggetti/variabili è basso è possibile rilevare forti differenze tra la capacità del modello di descrivere la risposta considerata e la sua capacità di predirla (si può passare dal 100% in *fitting* allo 0% in predizione!), ma anche e soprattutto dal punto di vista metodologico in quanto la qualità di un modello che viene utilizzato, in momenti successivi, per predire proprietà di campioni nuovi deve essere rappresentata da parametri che siano in qualche modo legati al problema della predizione. Tutti i parametri calcolati con procedure che non tengono conto degli aspetti predittivi non hanno - ovviamente! - nulla a che fare con le prestazioni dello stesso modello in predizione.
Si è già incontrata la tecnica di validazione nota col nome di *leave-one-out*, secondo la quale le caratteristiche predittive del modello vengono valutate globalmente ricostruendo la stessa procedura di calcolo con l'esclusione di un oggetto alla volta dal calcolo (e quindi dal modello) e valutando la capacità del modello di predire la proprietà considerata dell'oggetto escluso. Ad esempio, una tipica grandezza di questo tipo è *PRESS* (definita tra i parametri per la valutazione dei metodi di regressione) e la grandezza da essa derivata, R^2_{cv}%, interpretabile come la percentuale di varianza spiegata dal modello in predizione.

Nota. A tutt'oggi il numero di modelli di regressione multivariata, pubblicati anche da autorevoli riviste internazionali, ottenuti da un esiguo numero di dati (da 8-10 a 15-20) e per i quali viene fornito solo il valore R^2 (in *fitting*), è straordinariamente alto, cioè costituisce ancora la quasi totalità. In molti casi, il ricalcolo del modello in predizione mediante tecniche di validazione (R^2_{cv}) rivela un abbassamento importante delle prestazioni del modello ($R^2_{cv} << R^2$): modelli che vengono quindi considerati ottimi o buoni, in quanto valutati solo nelle loro prestazioni descrittive, sono in realtà modelli mediocri o totalmente inaffidabili, quando valutati nelle loro potenzialità predittive. Questa confusione tra due aspetti modellistici così diversi - quello descrittivo e quello predittivo - viene ulteriormente accentuata dal fatto che le risposte ottenute da modelli soltanto descrittivi vengono molto spesso chiamate "risposte predette": si tratta invece di "risposte calcolate", mentre la qualifica di "risposte predette" va esclusivamente riservata alle risposte ottenute nella fase di validazione o nei

casi in cui il modello viene applicato per predire risposte relative a casi in cui queste ultime sono effettivamente incognite.

In generale, la validazione di un modello, cioè la valutazione della sua capacità predittiva, avviene secondo lo schema di Fig. 5-5. Una parte dei dati complessivi viene utilizzata per costruire il *training set* (insieme di apprendimento), mentre la parte restante viene utilizzata per costruire l'*evaluation set* (insieme di valutazione). Il primo insieme di dati viene utilizzato per costruire un modello ridotto (diverso dal modello finale) che viene successivamente utilizzato per predire la risposta degli oggetti che fanno parte del secondo insieme di dati: nessuna parte dell'informazione contenuta in questi dati deve essere utilizzata nella costruzione del modello ridotto.
L'insieme delle predizioni effettuate sui dati dell'*evaluation set* viene utilizzato per calcolare uno o più parametri che consentono di valutare la capacità predittiva del modello finale, modello che, di norma, viene comunque costruito utilizzando tutti i dati disponibili. I parametri di valutazione direttamente calcolati per il modello finale non contengono quindi informazione sulla sua capacità predittiva, ma solamente sulla sua capacità di descrivere il complesso dei dati. La linea tratteggiata indica che al modello finale si perviene utilizzando tutti i dati e che la procedura di validazione ha solo lo scopo di fornire dei parametri di valutazione della capacità predittiva del modello.

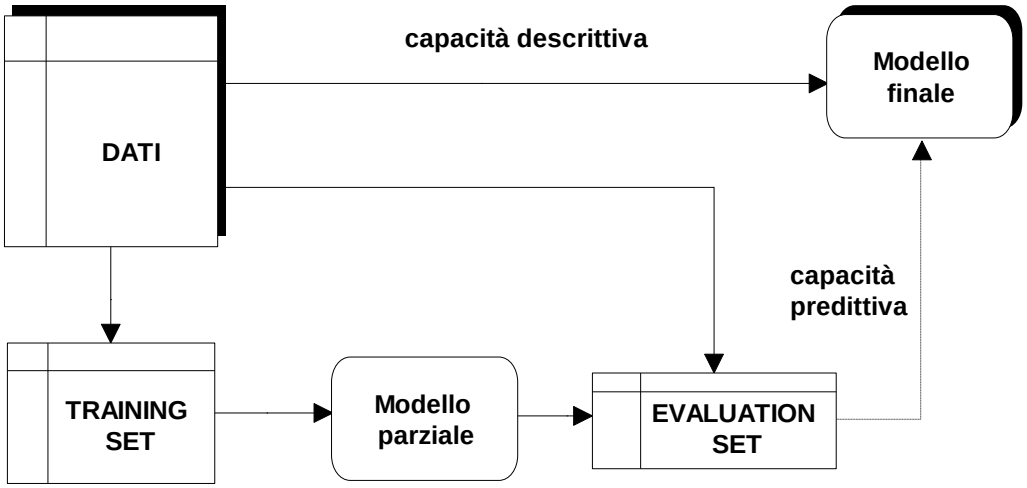


FIG. 5-5

5.4 - Le tecniche di validazione

Le tecniche di validazione chemiometriche differiscono tra loro sostanzialmente per le modalità con cui vengono ripartiti gli oggetti tra *training set* ed *evaluation set* e per il numero di *training / evaluation set* che vengono generati. Le tecniche più comuni sono:

- metodo *leave-one-out*
- metodo *leave-more-out*
- metodo *training/evaluation splitting*
- metodo *bootstrap*

🗀 Metodo *leave-one-out*

Secondo questo metodo, dati *n* oggetti, si calcolano *n* modelli in ciascuno dei quali viene escluso un oggetto alla volta: ciascun modello viene calcolato con gli *n*-1 oggetti restanti e viene utilizzato per predire la risposta (sia essa una riposta quantitativa di un modello di regressione o una risposta qualitativa di un

modello di classificazione): al fine di stimare un parametro che sia effettivamente in relazione col potere predittivo del modello, lo scarto tra la risposta sperimentale e quella predetta viene accumulato per tutti gli *n* oggetti che, a turno, vengono esclusi dal modello. Di norma, il modello finale effettivo sarà sempre calcolato con tutti gli *n* dati disponibili, in modo da poter sfruttare *tutta* l'informazione disponibile in *tutti* i dati. Questo metodo di validazione è quello che comporta la perturbazione minima in quanto ogni modello viene sempre calcolato utilizzando *n*-1 dati ed è quello che fornisce l'unico modo per un confronto univoco in predizione tra modelli ottenuti con metodi diversi e descrittori diversi. Il limite di questo metodo di validazione è che esso tende al valore di R^2 in *fitting* al crescere del numero *n* di campioni; per questo motivo il metodo *leave-one-out* può fornire valori predittivi ancora troppo ottimistici se il numero di campioni è abbastanza elevato.

🗀 Metodo *leave-more-out*

Questo metodo è una generalizzazione del metodo *leave-one-out*. I dati vengono suddivisi in *G gruppi di cancellazione* in modo tale che *k* oggetti pari a *n*/*G* vengono assegnati all'insieme di valutazione e quindi esclusi dal *training set*: il modello, stimato da un *training set* costituto da *n* - *k* oggetti, viene utilizzato per predire la risposta dei *k* oggetti esclusi dal modello. I *k* oggetti vengono esclusi una sola volta ciascuno (*cross-validation*); una volta esclusi, questi vengono reinseriti nel modello e ne vengono esclusi altri *k*. Ogni *i*-mo oggetto viene quindi assegnato ad un gruppo di cancellazione *g*, secondo l'espressione:

$$g = \mathrm{mod}\,(\,i - 1, G\,) + 1$$

dove mod(a,b) è l'operazione *modulo*, cioè *a modulo b.*

Nota. L'operazione modulo fornisce il resto della divisione tra il primo ed il secondo argomento, dove gli argomenti sono numeri interi. Ad esempio:

$5 \bmod 2 = 1 \qquad 16 \bmod 4 = 0 \qquad 18 \bmod 5 = 3 \qquad 27 \bmod 7 = 6$

Secondo questa modalità la collocazione di ciascun oggetto nell'insieme di valutazione avviene in modo sistematico: il risultato dipende ovviamente dall'ordine in cui si presentano gli oggetti nelle matrice complessiva dei dati.
Un'alternativa alla selezione sistematica degli oggetti è quella in cui i *k* oggetti che di volta in volta costituiscono l'insieme di valutazione sono selezionati casualmente, ma una sola volta, senza seguire un ordine sistematico.
Nel contesto del metodo *leave-more-out*, il metodo *leave-one-out* è un metodo *leave-more-out* con $G = n$, cioè con un numero di gruppi di cancellazione pari al numero di oggetti e quindi $k = 1$.
Rispetto al metodo *leave-one-out*, questo metodo di validazione è più perturbativo, cioè valuta in modo più severo le prestazioni predittive del modello. Per evitare che, ricercando modelli che abbiano il massimo potere predittivo col metodo *leave-one-out*, questa valutazione permanga ancora troppo ottimistica, sarebbe opportuno effettuare alcuni controlli sulla stabilità del modello in predizione anche col metodo *leave-more-out*. In ogni caso, il modello finale è comunque calcolato utilizzando tutti i dati disponibili.

🗀 Metodo *training/evaluation splitting*

Secondo questo metodo, dati *n* oggetti, l'insieme dei dati viene suddiviso in due parti: il *training set*, i dati utilizzati per costruire il modello, e l'*evaluation set*, i dati utilizzati per valutare il potere predittivo del modello calcolato dal *training set*. Per stabilire l'entità della suddivisione nei due sottoinsiemi, viene fissata una percentuale di oggetti da collocare nell'*evaluation set*, normalmente compresa tra il 10% e il 50%. La suddivisione avviene di norma secondo una procedura casuale. Questo metodo viene chiamato *insieme di valutazione singolo* (*Single Evaluation Set, SES*). E' stato dimostrato che questa procedura di norma produce modelli molto instabili, in quanto esiste una forte dipendenza dei risultati sia dalla dimensione dell'insieme di valutazione, soprattutto quando il numero di oggetti nell'insieme di valutazione è piccolo, sia dagli oggetti che sono stati casualmente inseriti in questo insieme. Questa tecnica potrebbe venire utilizzata per ottenere una prima rapida stima delle caratteristiche predittive di un modello, ma la variabilità dei risultati è tale da renderla comunque poco attendibile.
Una tecnica migliore è quelle di ripetere la procedura molte volte al fine di ottenere un valor medio attendibile del parametro per valutare il potere predittivo del modello. In questo caso la tecnica si chiama *insieme ripetuto di valutazione* (*Repeated Evaluation Set, RES*). Il limite più evidente di questa tecnica che richiede molte centinaia (o migliaia) di ripetizioni è il cospicuo

tempo di calcolo, ragione per cui non può venire considerata come una comune strategia di validazione.

🗀 Metodo *bootstrap*

Secondo questo metodo, dati *n* oggetti, vengono generati *k* insiemi, tutti *n* dimensionali, costruiti estraendo *a caso con ripetizione n* oggetti dall'insieme originale. Ciò significa che ciascun insieme dei dati è sempre *n*-dimensionale, ma ogni volta alcuni oggetti compariranno più volte nello stesso insieme, mentre altri non compariranno affatto. Il modello ottenuto dai primi viene utilizzato per predire le risposte degli oggetti esclusi. Questo metodo viene utilizzato solo se si dispone di calcolatori sufficientemente potenti in quanto è necessario procedere all'estrazione di molti campioni *n*-dimensionali (almeno alcune migliaia), per avere una stima attendibile del valor medio del parametro predittivo selezionato.

Nota. In generale si ritiene che un buon modello debba conservare un buon potere predittivo anche se sottoposto ad una grande perturbazione: ad esempio, costruendo il modello col 50% dei dati disponibili e utilizzando l'altro 50% per valutarne il potere predittivo (2 gruppi di *cross-validazione* nel metodo *leave-more-out*).
Tuttavia, è necessario tener presente che l'eliminazione dal *training set* di un numero eccessivo di oggetti nella costruzione del modello può comportare che venga completamente a mancare una parte di informazione fondamentale nella costruzione del modello stesso e che quindi questo - inevitabilmente - non possa prevedere correttamente comportamenti caratteristici di oggetti presenti solo nell'*evaluation set*. In questo caso, la stima della capacità predittiva del modello sarà inevitabilmente sottostimata rispetto alla disponibilità effettiva di informazione contenuta nei dati.

🕮 BIBLIOGRAFIA

M. FORINA, G. DRAVA, R. BOGGIA, S. LANTERI E P. CONTI (1994). Validation procedures in near-infrared spectrometry. *Analytica Chimica Acta*, **295**, 109-118.

W.J. KRZANOWSKY (1988). *Principles of Multivariate Analysis. A User's Perspective.* Oxford University Press, pp. 564.

B. EFRON (1982): *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied mathematics, Philadelphia, PA (USA).

6

I METODI DI CLASSIFICAZIONE

6.1 - Introduzione

I metodi di classificazione sono metodi che hanno l'obiettivo di costruire, sulla base di un certo numero di variabili indipendenti (descrittori), un modello capace di individuare la classe cui appartiene ciascun oggetto. Così come i *metodi di regressione* hanno lo scopo di identificare una relazione funzionale tra variabili indipendenti e una risposta quantitativa, i metodi di classificazione cercano una relazione tra le variabili indipendenti ed una *risposta qualitativa* (una classe).
Condizione necessaria perchè i metodi di classificazione siano applicabili è che le classi siano preventivamente definite e che ciascun oggetto del *training set* sia assegnato ad una delle classi definite.

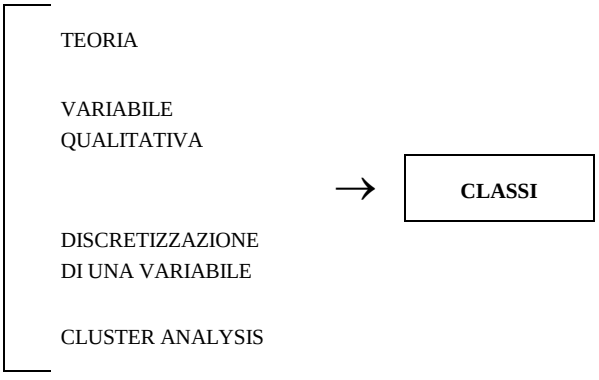


FIG. 6-1

La definizione preventiva delle classi può avvenire in accordo ad uno dei seguenti criteri (Fig. 6-1):

- le classi sono note *a priori* per via teorica.
- le classi sono definite da una variabile categorica.
- le classi vengono ricercate mediante i metodi di *cluster analysis*.
- le classi sono definite mediante la categorizzazione di una variabile quantitativa.

In particolare, la categorizzazione di una variabile quantitativa dovrebbe essere utilizzata tutte le volte che la risposta quantitativa è affetta da una quantità di rumore o di incertezza tali da non consentire il calcolo di modelli di regressione attendibili. Come indicato nello schema seguente, la variabile x_1 (continua, con valori compresi, ad esempio, tra 1.2 e 4.7) viene categorizzata in tre classsi.

classe 1	classe 2	classe 3	x_1 (variabile categorica)
< 2.20	2.20 - 3.30	> 3.30	x_1 (variabile continua)

Ad esempio, ogni oggetto il cui valore di x_1 è minore di 2.20 viene assegnato alla classe 1. Una volta utilizzata la variabile quantitativa per costruire le classi, essa viene eliminata definitivamente dai modelli di classificazione.
In questi casi, le classi definite risultano ordinate e l'applicazione dei metodi di classificazione risulta una strategia di modellamento più *robusta* dei metodi di regressione, cioè meno sensibile all'incertezza dovuta agli errori sperimentali nella valutazione delle risposte.
Classificare significa quindi assegnare un nuovo oggetto la cui classe è incognita ad una delle classi per le quali è stato costruito un modello di classificazione o definito un algoritmo in grado di classificare, in base ai dati forniti dal *training set*, ove le classi sono invece note a priori. In ogni caso, nella ricerca di un modello di classificazione, le variabili utilizzate per costruire il modello devono essere indipendenti dalle classi.
Nella Fig. 6-2 viene mostrato il criterio più naturale di classificazione: l'oggetto incognito viene assegnato alla classe in base alla minima distanza dell'oggetto dal centroide della classe, cioè alla classe C.
Tra i metodi di classificazione possiamo distinguere i **metodi di classificazione modellanti** e i **metodi di classificazione non-modellanti**: diversamente da questi ultimi, i primi producono un modello con il quale sono definibili anche i

confini di ciascuna classe, cioè le dimensioni di uno spazio che racchiude gli oggetti di ciascuna classe.

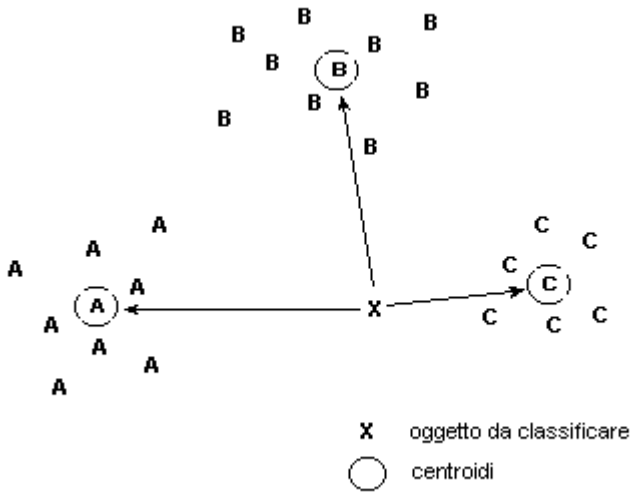


FIG. 6-2

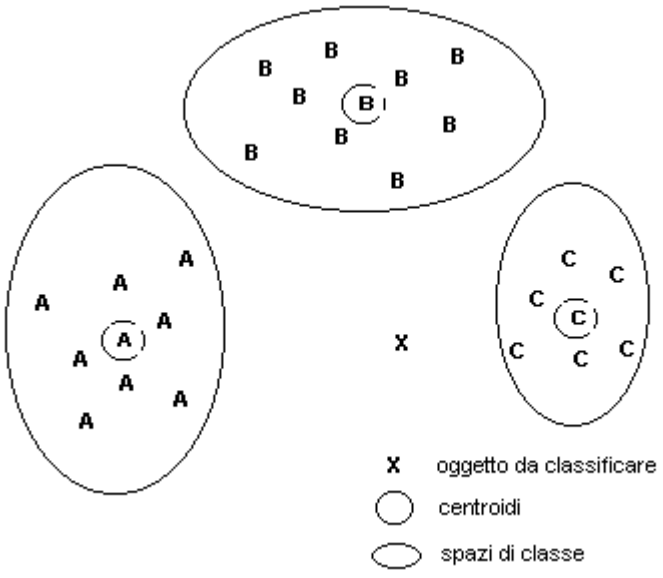


FIG. 6-3

Ciò significa che un oggetto può anche non essere classificato in una classe se non rientra nello spazio che delimita una classe: in questo caso l'oggetto viene considerato un *outlier* (se le ellissi di Fig. 6-3 rappresentano i confini delle tre classi, l'oggetto considerato **x** non sarebbe classificato in nessuna di esse).
Esiste una vasta gamma di metodi modellanti basati sulla definizione di funzioni empiriche di densità di probabilità, capaci quindi di assegnare su basi probabilistiche un oggetto ad una classe. Metodi di questo tipo fanno parte dei metodi di classificazione noti anche come **metodi di classificazione confusi** (***fuzzy classification methods***), per cui un oggetto può appartenere con probabilità assegnate a più di una classe. Un altro gruppo importante di metodi di classificazione è quello dei **metodi di classificazione ad albero** (***classification tree methods*)**, in cui la classificazione procede gerarchicamente attraverso separazioni binarie degli oggetti e produce modelli grafici di facile interpretabilità.

6.2 - Parametri di valutazione della classificazione

I risultati di una classificazione possono essere sintetizzati in una matrice detta **matrice di confusione C**. In questa matrice le righe rappresentano le *classi vere* ovvero le classi cui gli oggetti del *training set* appartengono, mentre le colonne rappresentano le classi assegnate agli oggetti dopo aver applicato una tecnica di classificazione, ovvero le *classi calcolate* oppure le *classi predette* mediante una procedura di validazione.
Ad esempio, supponiamo che 30 oggetti siano distribuiti in tre classi A, B e C (10 appartengono alla classe A, 12 alla classe B e 8 alla classe C).

classi assegnate

classi vere	**classi**	**A'**	**B'**	**C'**	n_g
	A	**9**	**1**	**0**	**10**
	B	**2**	**8**	**2**	**12**
	C	**1**	**2**	**5**	**8**
	$n_{g'}$	**12**	**11**	**7**	n **= 30**

TAB. 6-1

La Tab. 6-1 rappresenta una matrice di confusione; sono riportati i risultati, sotto forma di matrice di confusione, ottenuti da un ipotetico metodo di classificazione.
Sulla diagonale principale sono riportati gli oggetti classificati correttamente dal metodo: nel nostro caso, 9 della classe A, 8 della classe B e 5 della classe C.
I numeri nelle celle fuori dalla diagonale principale riportano gli oggetti che, pur appartenendo ad una certa classe, vengono *erroneamente* assegnati ad un'altra classe: nel nostro caso, 1 oggetto di A viene assegnato alla classe B, 2 oggetti di B vengono assegnati alla classe A e 2 alla classe C, 1 oggetto di C viene assegnato alla classe A e 2 alla classe B.
L'ultima colonna della tabella riporta i totali di riga (n_g), corrispondenti al numero di oggetti distribuiti nelle classi originali; l'ultima riga della tabella riporta i totali degli oggetti ridistribuiti nella 3 classi in base al modello calcolato ($n_{g'}$).

Il parametro che più semplicemente sintetizza il risultato di una procedura di classificazione è la **percentuale di classificazioni corrette** (***non-error rate*, NER%**), definito come:

$$NER\% = \frac{\sum_g NER\%_g}{G} \times 100$$

dove NER%$_g$ sono i *non-error rate* di ciascuna classe.
Nell'esempio proposto, NER% = 73.1% . Inoltre, è possibile calcolare lo stesso parametro per ognuna delle tre classi:

NER% (A) = 90.0% NER% (B) = 66.7% NER% (C) = 62.5%

Parametro complementare a NER% è la **percentuale di errore** (***error rate***, **ER%**), definito come 100 – NER%.

Nel nostro caso, ER% = 26.7% (8/30).
Un altro parametro importante per valutare la qualità di un modello di classificazione è il **rischio di errore di classificazione** (***misclassification risk*, MR%**).
La definizione di questo parametro richiede la definizione della **matrice delle perdite** o **matrice dei costi** (***loss matrix, L***). Questa matrice ha la stessa struttura della matrice di confusione e contiene negli elementi fuori dalla diagonale la quantificazione delle penalità dei diversi tipi di errore che si possono commettere nella classificazione. In pratica, la matrice ***L*** è una matrice dei pesi per i diversi tipi di errori di classificazione. Ad esempio, una matrice delle perdite per il caso precedente può essere definita così:

		classi assegnate		
		A'	**B'**	**C'**
classi vere	**A**	**0**	**1**	**2**
	B	**1**	**0**	**1**
	C	**2**	**1**	**0**

TAB. 6-2

Questo significa che gli errori di classificazione che confondono A con C sono ritenuti più gravi (valore 2 di penalità) di quelli che confondono A con B e C con B (valore 1 di penalità). Ovviamente, sulla diagonale principale, i valori sono uguali a zero, essendo le classificazioni corrette.
Questa matrice viene definita arbitrariamente, in funzione del problema considerato. Nel caso più frequente, non essendo esplicitata, si assume che tutti gli errori abbiano lo stesso peso: in questo caso tutti gli elementi fuori dalla diagonale principale sono uguali ad 1.
Prima di definire il rischio di errore, è necessario definire ancora un'altra grandezza importante nei problemi di classificazione: la **probabilità a priori di una classe**. In molti casi viene infatti richiesto di attribuire ad ogni classe una probabilità a priori, tipica della classe. Se non vi sono precedenti indicazioni

specifiche, due sono le modalità più tipiche per attribuire la probabilità a priori delle classi, secondo una delle due seguenti espressioni:

$$P_g = \frac{1}{G} \quad \text{oppure} \quad P_g = \frac{n_g}{n}$$

Nel primo caso, ad ogni classe viene assegnata la stessa probabilità, indipendentemente dal numero di oggetti che rappresenta ciascuna classe nel campione studiato. Nel secondo caso, la probabilità a priori di ogni classe viene definita dal rapporto tra il numero degli oggetti della classe e il numero di oggetti totali (probabilità proporzionale). In quest'ultimo caso, la classificazione di un oggetto in una classe poco rappresentata ha una probabilità a priori più piccola.
Il **rischio di errore di classificazione** tiene conto sia del numero di errori, sia della loro importanza (matrice delle perdite), sia della probabilità a priori di ciascuna classe ed è così definito:

$$MR\% = \sum_g \frac{\left(\sum_{g'} L_{gg'} c_{gg'}\right) P_g}{n_g} \times 100$$

dove g e g' indicano la classe vera e la classe assegnata, rispettivamente, P_g la probabilità a priori della classe g, $c_{gg'}$ gli elementi della matrice di confusione e $L_{gg'}$ gli elementi della matrice delle perdite.
In questo caso, tenendo conto della matrice delle perdite sopra definita, i valori di MR% sono:

MR% = 31.1% MR% (A) = 3.3% MR% (B) = 11.1% MR% (C) = 16.7%

Nel caso in cui la matrice delle perdite fosse tutta definita con valori fuori dalla diagonale uguali a 1, otterremmo:

MR% = 26.9% MR% (A) = 3.3% MR% (B) = 11.1% MR% (C) = 12.5%

Nel caso più semplice, in cui tuttte le classi hanno la stessa probabilità a priori, lo stesso numero di oggetti e gli elementi non diagonali della matrice delle perdite hanno valori tutti uguali a 1, la percentuale di errori ed il rischio di

errore coincidono (ER% = MR%). Queste due quantità coincidono anche nel caso in cui le classi abbiano una probabilità a priori proporzionale $P_g = n_g / n$.

Quando l'attenzione viene posta su ciascuna singola classe separatamente, è possibile definire due altri parametri particolarmente interessanti: la **sensibilità** (***sensitivity***, ***Sn***) e la **specificità** (***specificity***, ***Sp***).

La sensibilità di una classe viene definità come il rapporto percentuale tra gli oggetti assegnati a quella classe c_{gg} ed il numero totale di oggetti appartenenti alla stessa classe n_g:

$$Sn_g = \frac{c_{gg}}{n_g} \times 100$$

La sensibilità definisce quindi la capacità di una classe di rappresentare gli oggetti di quella classe.

La specificità di una classe viene definita come il rapporto percentuale tra gli oggetti della classe considerata assegnati alla classe g' e gli oggetti totali assegnati a quella classe $n_{g'}$:

$$Sp_g = \frac{c_{gg}}{n_{g'}} \times 100$$

La specificità definisce la capacità di una classe di isolare dalle altre classi gli oggetti di quella classe, cioè il suo grado di purezza.

Nell'esempio discusso:

Classi	*Sn%*	*Sp%*
A	90.0	75.0
B	66.7	72.7
C	62.5	71.4

TAB. 6-3

Nota. Quando il metodo di classificazione utilizzato produce un solo insieme per ogni classe definita (come avviene nella maggior parte dei casi, ma non

necessariamente nei metodi di classificazione ad albero), la sensibilità *Sn* coincide con il *NER%* della classe.

Per i metodi di classificazione viene assunta come situazione di riferimento quella secondo la quale, in assenza di un modello, tutti gli oggetti - qualsiasi sia la loro classe vera - sono assegnati alla classe più grande. Questo riferimento viene chiamato ***No-Model*** (*Nessun Modello*). Valori dei parametri di classificazione vicini ai valori assunti nella situazione *No-Model* indicano esiti scadenti della procedura di classificazione. Ovviamente, il valore di *No-Model* è unico e indipendente dai metodi di classificazione.
Ad esempio, l'**errore percentuale senza modello** (***NO-Model Error Rate***, **NOMER%)** viene definito come:

$$NOMER\% = \frac{n - n_M}{n} \times 100$$

dove n_M è il numero di oggetti della classe più numerosa. Nel caso considerato, dovremmo assegnare i 10 oggetti della classe A e gli 8 oggetti della classe C alla classe B che ne contiene 12, commettendo cioè 18 errori su 30: *NOMER* % = 60.0 %.
Il corrispondente valore calcolato per il rischio di errore nella situazione *No-Model* e per una probabilità a priori di classe pari a 1/G è definito come (G-1)/G ed è, in questo caso, 66.7% (2 classi su 3 sono malclassificate).

6.3 - Informazione ed entropia

Introduciamo a questo punto un tipo di indici, noti col nome di **indici di diversità**, o indici di informazione o entropia. Questi indici, il cui utilizzo è del tutto generale e non ristretto ai problemi di classificazione, sono una misura della quantità di informazione contenuta nei dati e il più noto tra questi prende il nome di **entropia di Shannon**, ***H***. Questo indice è definito come

$$H = -\sum_k p_k \log_2 p_k = -\sum_k \frac{n_k}{n} \log_2 \frac{n_k}{n}$$

dove p_k è la probabilità del *k*-mo evento e viene calcolata come

$$p_k = \frac{n_k}{n}$$

dove n_k sono gli eventi di tipo *k* (o della *k*-esima classe di equivalenza) ovvero il numero di oggetti appartenenti alla classe *k* (cioè aventi in comune la caratteristica *k*) ed *n* è il numero di eventi totali.
L'uso del logaritmo in base 2 consente di ottenere dei valori la cui unità di misura è espressa in ***bit***.
Il valore massimo della funzione *H* corrisponde alla massima diversità ed si ha quando tutti gli oggetti sono traloro diversi; in questo caso si hanno *n* contributi in cui la probabilità è 1/*n*:

$$H^{max} = -n \cdot \left(\frac{1}{n} log_2 \frac{1}{n} \right) = -log_2 \frac{1}{n} = log_2 n$$

Quindi, poichè il valore massimo della funzione entropia è dato da $log_2 n$, si può definire anche un'entropia normalizzata come:

$$H' = \frac{H}{log_2 n} \qquad 0 \le H' \le 1$$

Un altro indice di diversità utilizzato è **l'indice di diversità di Gini, *G.I.***, definito come

$$G.I. = \sum_{k \neq k'} p_k \cdot p_{k'}$$

dove p_k è la probabilità del *k*-mo evento e la sommatoria scorre sui prodotti tra tutte le coppie di eventi diversi. Questo indice rende conto della quantità di impurezza di un insieme.

Nota. I fattori di conversione tra logaritmi decimali (log), logaritmi naturali (ln) e logaritmi in base 2 (log_2) sono:

Da	A	dividi per
log	$\log_2$	0.30103
log	ln	0.43429
ln	$\log_2$	0.69315

TAB. 6-4

Ad esempio:

$$log\ 2 = 0.30103$$
$$log_2\ 2 = 0.30103/0.30103 = 1$$

Questi indici possono essere utilizzati nei modi più diversi, a secondo di come vengono definite le classi di equivalenza. Ad esempio, è possibile valutare la degenerazione di una variabile, cioè se i suoi valori sono tutti uguali (cioè, una costante) la degenerazione è massima; se, al contrario, i valori sono tutti diversi, la degenerazione è minima, o meglio nulla. In questo caso, l'**indice di degenerazione** *D%* può essere scritto come funzione dell'entropia di Shannon normalizzata:

$$D\% = \left(\frac{log_2 n - \sum_x \frac{n_x}{n} log_2 \frac{n_x}{n}}{log_2 n} \right) \times 100 = \left(1 - H'\right) \times 100$$

dove n_x è il numero di oggetti che hanno lo stesso valore della variabile *x*. La massima degenerazione coincide con la minima informazione nei dati, come è il caso di una costante. Ad esempio, la quantità massima di informazione contenuta in 50 dati relativi ad una variabile è

$$log_2 50 = 5.644\ bit$$

E' questo il caso in cui tutti i 50 valori sono diversi tra loro, cioè *D%* = 0.
Se 10 dati sono tra loro uguali e i restanti 40 sono invece tutti diversi tra loro, la quantità di informazione complessivamente contenuta nei dati è:

$$H = -\left[40 \cdot \left(\frac{1}{50} log_2 \frac{1}{50}\right) + \frac{10}{50} log_2 \frac{10}{50}\right] = 4.979\ bit$$

La quantità di informazione complessiva è scesa a 4.979 bit per il fatto che 10 dei 50 dati sono rappresentati dallo stesso valore numerico. In questo caso la degenerazione è:

$$D\% = \left(\frac{5.644 - 4.979}{5.644}\right) \times 100 = 11.8\%$$

Molti descrittori molecolari topologici si basano sull'entropia di Shannon, definendo opportunamente i criteri con cui costruire le classi di equivalenza. Ad esempio, è possibile definire due semplici indici molecolari topologici sulla base dell'equivalenza degli atomi e dei legami tra gli atomi.
Nell'applicazione degli indici di entropia a problemi di classificazione, le classi di equivalenza vengono costruite dalle classi calcolate, ove la probabilità di ciascuna classe è definita dal rapporto tra gli oggetti di una classe presenti nella classe calcolata e gli oggetti totali della classe calcolata: se una classe calcolata contiene solo oggetti di quella classe, l'entropia è zero.
Sia l'indice di Shannon che l'indice di Gini possono essere minimizzati per la ricerca del miglior modello di classificazione.
I valori di entropia di Shannon per le classi calcolate nell'esempio sono:

$$H(\text{A}') = -\left(\frac{9}{12} log_2 \frac{9}{12} + \frac{2}{12} log_2 \frac{2}{12} + \frac{1}{12} log_2 \frac{1}{12}\right) = 1.041$$

$$H(\text{B}') = -\left(\frac{1}{11} log_2 \frac{1}{11} + \frac{8}{11} log_2 \frac{8}{11} + \frac{2}{11} log_2 \frac{2}{11}\right) = 1.096$$

$$H(\text{C}') = -\left(0 + \frac{2}{7} log_2 \frac{2}{7} + \frac{5}{7} log_2 \frac{5}{7}\right) = 0.863$$

Le entropie relative possono essere calcolate tenendo conto della frazione dei dati totali di ciascuna delle classi calcolate. Le frazioni sono rispettivamente: 12/30, 11/30 e 7/30.

Le entropie relative sono quindi:

$$H_R(\mathrm{A}') = 1.041 \times \frac{12}{30} = 0.416$$
$$H_R(\mathrm{B}') = 1.096 \times \frac{11}{30} = 0.402$$
$$H_R(\mathrm{C}') = 0.863 \times \frac{7}{30} = 0.201$$

L'entropia totale residua è data dalla somma delle entropie relative ed è quindi $H_R = 1.019$.
Il modello ottenuto ha ancora una quantità di informazione non risolta (cioè che permane incognita) di 1.019 bit.
L'entropia iniziale, cioè relativa ai dati indivisi, viene calcolata da:

$$H_0 = -\left(\frac{10}{30} log_2 \frac{10}{30} + \frac{12}{30} log_2 \frac{12}{30} + \frac{8}{30} log_2 \frac{8}{30}\right) = 1.566$$

Ciò significa che il modello di classificazione ottenuto rende conto di una quantità percentuale di informazione pari a

$$\frac{1.566 - 1.019}{1.566} \times 100 = 34.9\%$$

I corrispondenti valori dell'indice di diversità di Gini sono:

G.I.(A') = 0.210 *G.I.*(B') = 0.215 *G.I.*(C') = 0.204

Gli indici relativi, calcolati utilizzando le stesse frazioni definite in precedenza, sono:

$G.I._R$ (A') = 0.080 $G.I._R$ (B') = 0.079 $G.I._R$ (C') = 0.048

Il valore dell'indice di Gini prima della classificazione è pari a 0.329, mentre la somma dei valori relativi, dopo la classificazione, è 0.207.
Con questo indice il modello di classificazione consente di diminuire la quantità di impurezza iniziale del 37.1%.

6.4 - Il metodo k-nearest neighbours (K-NN)

Il **metodo K-NN** è un metodo di classificazione non-parametrico (cioè, che prescinde dalla conoscenza della distribuzione statistica delle variabili) che utilizza per la classificazione il concetto di *analogia*. Il metodo si basa sulla scelta di una distanza (generalmente la distanza euclidea) e sulla selezione di numero intero di *k* intorni (gli oggetti più *vicini* ad ogni oggetto da classificare) ai quali si estende la valutazione delle classi cui essi appartengono al fine di classificare l'oggetto considerato.

L'algoritmo su cui si basa il metodo K-NN è il seguente:

a. scalatura dei dati
b. selezione della distanza da utilizzare
c. scelta del numero di intorni *k* utili per la classificazione
d. calcolo della matrice delle distanze
e1. per ogni oggetto si considerano i *k* oggetti più vicini
e2. l'oggetto viene assegnato alla classe più rappresentata nei *k* vicini

In pratica, è necessario provare diversi valori di *k* per ricercare il valore ottimale, cioè quello in cui si commettono meno errori di classificazione sui dati relativi al *training set*.
Di norma, i valori utilizzati di *k* sono 1,3,4,5,6,7..10.
Il criterio di assegnazione (e2) della classe di appartenenza non è l'unico possibile, ma è senz'altro il più utilizzato. Si tratta di un criterio maggioritario, cioè la classe cui assegnare l'oggetto considerato è quella più presente nei *k* oggetti più vicini. In caso di parità (ad esempio, $k = 4$, 2 oggetti appartenenti alla classe A e due oggetti alla classe B), l'oggetto viene assegnato alla classe da cui è minima la somma delle distanze calcolate separatamente per i due oggetti più vicini di ciascuna classe. Ciò spiega anche perchè normalmente non si utilizza $k = 2$, poichè il risultato sarebbe del tutto identico a quello ottenuto utlizzando $k = 1$.

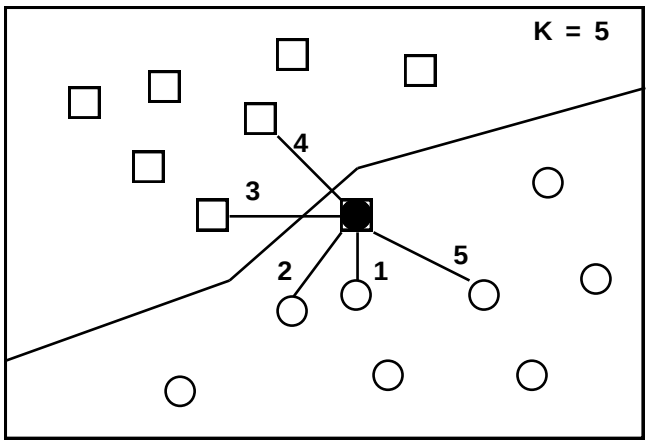


FIG. 6-4

Il metodo K-NN non fornisce un vero e proprio modello matematico, cioè una funzione analitica applicabile successivamente agli oggetti incogniti da classificare. Il "modello K-NN" è costituito dal miglior valore di *k* trovato, dalla metrica utilizzata e da tutti i dati appartenenti al *training set*. La predizione della classe per un nuovo oggetto viene effettuata aggiungendo l'oggetto al *training set* e valutando a quale classe esso viene assegnato in funzione dei *k* oggetti più vicini appartenenti al *training set*. Questo metodo fornisce normalmente dei buoni risultati ed è particolarmente efficace quando le superfici di separazione tra le classi sono non-lineari e particolarmente complesse. Tuttavia, la mancanza di un vero e proprio modello matematico gli attribuisce un carattere empirico che è sovente poco gradito.

6.5 - L'analisi discriminante

L'**Analisi Discriminante Lineare** (**LDA**) e l'**Analisi Discriminante Quadratica** (**QDA**) sono due metodi di classificazione largamente utilizzati ed hanno un fondamento statistico completamente sviluppato e studiato. Si tratta di **metodi Bayesiani** e quindi classificano gli oggetti in base alla seguente regola:

un oggetto $\mathbf{x}_i$ viene classificato nella classe g se

$$p(g|\mathbf{x}_i) > p(k|\mathbf{x}_i) \qquad \text{per tutti i } k \neq g \text{ e } k = 1,...,G \qquad (1)$$

dove $p(g|\mathbf{x}_i)$ è la *probabilità a posteriori* che l'oggetto $\mathbf{x}_i$ appartenga alla classe g e G è il numero totale di classi. Questa probabilità viene calcolata dalla **regola di Bayes**:

$$p(g|\mathbf{x}_i) = \frac{P_g f(\mathbf{x}_i|g)}{\sum_k P_k f(\mathbf{x}_i|k)} \qquad (2)$$

dove P_g è la *probabilità a priori* della classe g e $f(\mathbf{x}_i|g)$ è la *densità di probabilità* che una classe g contenga l'oggetto $\mathbf{x}_i$. La densità di probabilità è generalmente ignota e deve essere stimata dagli oggetti del *training set.*
Utilizzando la regola di Bayes, l'equazione (1) può essere espressa come

$$f(\mathbf{x}_i|g)P_g > f(\mathbf{x}_i|k)P_k \qquad \text{per tutti i } k \neq g \text{ e } k = 1,...,G \qquad (3)$$

Di fatto, un oggetto $\mathbf{x}_i$ viene classificato nella classe g se è minima l'espressione

$$D_g(\mathbf{x}_i) = (\mathbf{x}_i - \bar{\mathbf{x}}_g)^T \mathbf{S}_g^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_g) + ln|\mathbf{S}_g| - 2\,ln\,P_g \qquad (4)$$

dove $D_g(\mathbf{x}_i)$ è detto ***discriminant score,*** $\mathbf{S}_g^{-1}$ è l'inversa della **matrice di covarianza della classe** g e $\bar{\mathbf{x}}_g$ è il corrispondente **centroide di classe**.

La quantità $(\mathrm{x}_i - \bar{\mathrm{x}}_g)^{\mathbf{T}} \mathrm{S}_g^{-1}(\mathrm{x}_i - \bar{\mathrm{x}}_g)$ è il quadrato di una distanza chiamata **distanza di Mahalanobis.**
Gli elementi della matrice di covarianza di classe sono così definiti:

$$s^2_{jkg} = \frac{\sum_i (x_{ijg} - \bar{x}_{jg})(x_{ikg} - \bar{x}_{kg})}{n_g - 1} \qquad i = 1,..., n_g$$

dove s^2_{jkg} indica l'elemento della matrice di covarianza relativo alle variabili *j* e *k*, calcolato da tutti gli n_g oggetti appartenenti alla classe *g*.

L'ipotesi fondamentale che sottende questo approccio alla classificazione è che le variabili siano distribuite *normalmente*. Infatti, l'equazione (4) viene ottenuta dall'equazione (3) sostituendo l'equazione di una **distribuzione normale multivariata** (5) al posto delle densità di probabilità $f(\mathbf{x}_i|g)$, facendone il logaritmo naturale, eliminando un termine costante e moltiplicando per -2.

$$f(\mathbf{x}_i|g) = \frac{1}{(2\pi)^{p/2}|\mathbf{S}_g|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x}_i - \bar{\mathbf{x}}_g)^{\mathrm{T}}\mathbf{S}_g^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_g)\right] \qquad (5)$$

Nonostante questa richiesta di normalità delle variabili, l'analisi discriminante fornisce buoni risultati anche in presenza di scostamenti non piccoli dalle condizioni di normalità.
Se si può assumere che le matrici di covarianza di classe siano uguali (o simili) per tutte le classi (cioè, provengono dalla stessa popolazione), si calcola un'unica matrice di covarianza pesata (***pooled covariance matrix*, $\mathbf{S}_p$**):

$$\mathbf{S}_p = \frac{\sum_g (n_g - 1)\mathbf{S}_g}{n - G}$$

In questo caso, i valori dei *discriminant scores* seguono un andamento lineare: ciò significa che la separazione tra le classi è lineare, cioè è definita da una retta, da un piano o da un iperpiano (**analisi discriminante lineare**).
Se le matrici di covarianza sono tra loro diverse, vengono utilizzate tutte le matrici di covarianza di classe $\mathbf{S}_g$, $g = 1,\ldots G$. In questo secondo caso, le ipersuperfici che separano le classi sono quadratiche (**analisi discriminante quadratica**).

Nota. Diversamente dal metodo K-NN, entrambi i metodi di analisi discriminante (LDA e QDA) sono invarianti alla scalatura dei dati. Ciò

significa che i risultati non cambiano adottando tipi di scalature diverse o utilizzando direttamente i dati originali.

Un'estensione di questo approccio all'analisi discriminante è l'**Analisi Discriminante Regolarizzata** (**RDA**), che fa uso di una doppia procedura di regolarizzazione della matrice di covarianza di classe $\mathbf{S}_g$.
Con la prima procedura di regolarizzazione, basata sul parametro λ, la matrice di covarianza $\mathbf{S}_g$ è rimpiazzata da una combinazione lineare della matrice di covarianza di classe e della matrice di covarianza *pooled*, secondo lo schema:

$$S_g(\lambda) = (1-\lambda)S_g + \lambda S_p \qquad (6)$$

La seconda procedura di regolarizzazione si basa sulla convenienza di 'spingere' la matrice di covarianza di classe verso un multiplo della matrice identità, dove il multiplo è l'autovalore medio della matrice di covarianza.

$$\mathbf{S}_g(\lambda,\gamma) = (1-\gamma)\mathbf{S}_g(\lambda) + \frac{\gamma}{p} tr\ \mathbf{S}_g(\lambda)\ \mathbf{I} \qquad (7)$$

Quando il parametro γ è uguale ad 1, la distanza di Mahalanobis, compresa nella (5), diviene una **distanza euclidea** moltiplicata per un fattore costante se la matrice di covarianza è unica ($\lambda = 1$), diverso per ogni classe se $\lambda < 1$.
In questo caso ($\gamma = 1$), se $\lambda = 1$, la classsificazione avviene assegnando un oggetto alla classe il cui centroide è più vicino (**Nearest Mean Classifier, NMC**), come indicano le 3 distanze dai centroidi di Fig.6-2; se $\lambda = 0$, la classificazione avviene allo stesso modo, ma la distanza euclidea è pesata dall'inverso del valor medio della somma degli elementi diagonali (la traccia) della matrice di covarianza di classe (**Weighted Nearest Mean Classifier, WNMC**). Quest'ultimo caso, maggiore è la varianza totale di una classe, minore è la distanza: vale a dire che, a parità delle altre condizioni, un oggetto viene più facilmente assegnato ad una classe ad alta varianza che ad una classe i cui oggetti sono più concentrati vicino al baricentro della classe (minore varianza).
In Fig.6-5 sono schematizzati i diversi casi caratteristici del'analisi discriminante regolarizzata.

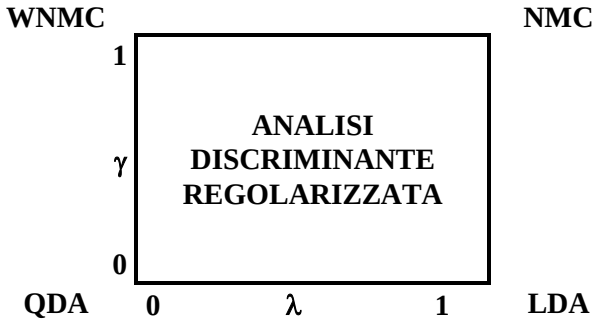


FIG. 6-5

Nell'analisi discriminante regolarizzata, l'assegnazione di un oggetto $\mathbf{x}_i$ avviene ancora in base all'espressione (4), in cui la matrice di covarianza $\mathbf{S}_g$ viene sostituita dalla matrice $\mathbf{S}_g(\lambda,\gamma)$.
Nella pratica, il metodo esplora una griglia di valori dei parametri λ e γ e sceglie la coppia di valori per i quali si ottiene il miglior risultato, cioè il minimo *rischio di errore* in predizione.

6.6 - Il metodo SIMCA

SIMCA (***Soft Independent Models of Class Analogy***) è stata la prima tecnica di modellamento di classe utilizzata in chemiometria. Proposto da Svante Wold, il metodo SIMCA è basato sull'idea di modellare separatamente ciascuna classe mediante una rappresentazione sintetica fornita dalle componenti principali significative di ogni classe.

Le principali caratteristiche del metodo sono:

a) non vi sono ipotesi sulla distribuzione delle variabili (metodo *non-parametrico*)
b) ogni classe è rappresentata da un modello costituito dalle sue componenti principali significative
c) il centro della classe è definito dall'origine delle componenti principali che rappresentano la classe
d) ogni classe viene (generalmente) autoscalata indipendentemente dalle altre

e) esistono una serie di descrittori delle prestazioni di ciascun modello di classe, utili a descrivere la classe e le sue relazioni con le altre.

L'algoritmo è il seguente:

1. *autoscaling* separato per ogni classe
2. analisi delle componenti principali separatamente sui dati di ogni classe
3. stima del numero di componenti principali significative per ogni classe
4. calcolo del *normal range* di ogni componente per ogni classe
5. calcolo dell'*extended range* o del *reduced range* di ogni componente per ogni classe

I dati autoscalati sono quindi rappresentati dal modello come

$$x_{ijg} = \sum_m t_{img} \ell_{jmg} + r_{ijg} \qquad \text{con } j = 1,..., p \qquad m = 1,..., M_g$$

essendo l'indice *i* è relativo agli oggetti della classe *g*-esima, *p* il numero di variabili e M_g il numero di componenti significative di ciascuna classe. Le quantità *t*, ℓ e *r* sono rispettivamente gli *scores*, i *loadings* e i residui.
Il **modello matematico** di ciascuna classe è quindi definito come:

$$\hat{x}_{ijg} = \sum_m t_{img} \ell_{jmg}$$

I residui sono quindi definiti come:

$$r_{ijg}^2 = \left(\hat{x}_{ijg} - x_{ijg}\right)^2$$

La deviazione standard residua di una classe rsd_g è una misura della dispersione degli oggetti del *training set* della classe *g* intorno al suo centro

$$rsd_g^2 = \frac{\sum_i \sum_j r_{ijg}^2}{\left(p - M_g\right)\left(n_g - M_g - 1\right)}$$

Questa grandezza rappresenta la *distanza tipica* degli oggetti della classe dal modello matematico della classe.

Il singolo oggetto del *training set* ha una *distanza SIMCA dal modello* definita dalla somma dei suoi residui estesa a tutte le variabili:

$$rsd_{ig}^2 = \frac{n_g \sum_j r_{ijg}^2}{\left(p - M_g\right)\left(n_g - M_g - 1\right)}$$

Questa espressione vale solo per gli oggetti del training set utilizzati per calcolare il modello di classe. Per gli oggetti delle altre classi o dell'*evaluation set*, la distanza SIMCA è calcolata da:

$$rsd_{ig}^2 = \frac{\sum_j r_{ijg}^2}{\left(p - M_g\right)}$$

I residui sono quindi le differenze tra le coordinate degli oggetti e le coordinate delle loro proiezioni nello spazio interno della classe rappresentata dal modello matematico. Quest'ultimo descrive, per ogni classe, le correlazioni significative che caratterizzano la classe.
La regola di classificazione SIMCA minimizza il rapporto tra la varianza residua individuale (di un oggetto) e la varianza residua media della classe:

$$D_{g^*}\left(\mathbf{x}_i\right) = \min_g \; rsd_{ig}^2 / rsd_g^2$$

Il rapporto

$$F = \frac{rsd_{ig}^2}{rsd_g^2}$$

è il rapporto tra due varianze e segue quindi una statistica *F*, con un numero di gradi di libertà uguale a (p - M_g) e (p-M_g)(n_g-M_g-*1*).
Un valore di probabilità selezionato di F viene utilizzato per calcolare il confine dello spazio della classe intorno al centro del modello matematico: lo spazio racchiuso da questo confine si chiama *SIMCA box.*
Gli oggetti del training set che hanno un valore calcolato di *F* superiore al confine critico F_C sono considerati **outliers** (* in Fig. 6-6).
Il *normal range* di ogni classe è definito per ogni componente *m* dai valori minimi e massimi $t_{mg}^{\min}$ e $t_{mg}^{\max}$.

Quando la classe non è sufficientemente ben rappresentata (pochi oggetti del *training set*) la distanza SIMCA viene aumentata (lungo ogni componente dove sia necessario) per ottenere una stima più realistica dello spazio di classe.
L'*extended range* non è una caratterisitca fissa di SIMCA. Nel caso si abbia un grande numero di oggetti si può ottenere una stima migliore dello spazio di classe utilizzando un range più ridotto (*reduced range*). Infatti, in questi casi il *normal range* è spesso sovrastimato a causa dell'errore sperimentale.

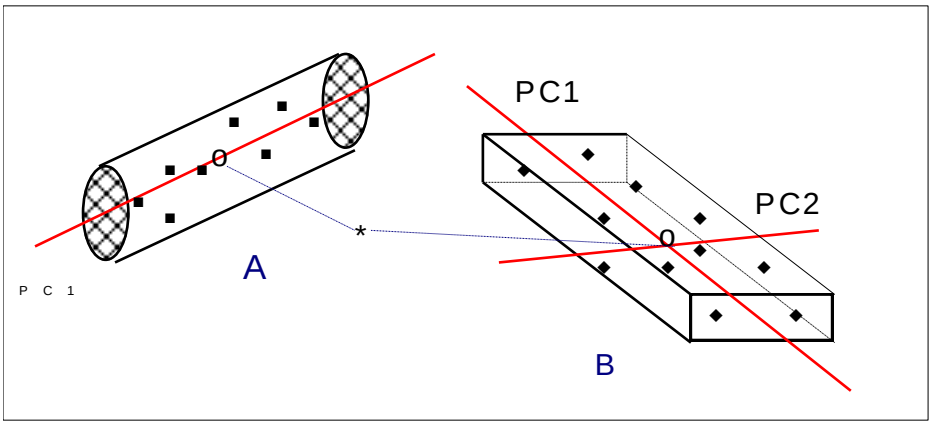


FIG. 6-6

Nota. La regola di classificazione di SIMCA può essere inserita nel contesto dell'analisi discriminante mediante la sua rappresentazione in termini di autovalori-autovettori.
La generica matrice di covarianza per la classe *g* può essere scritta come

$$\mathbf{S}_g = \mathbf{V}\Lambda\mathbf{V}^{\mathbf{T}} = \sum_j \lambda_{jg} \cdot \mathbf{v}_{jg} \cdot \mathbf{v}_{jg}^{\mathbf{T}}$$

dove λ_{jg} è il *j*-mo autovalore della classe *g* e $\mathbf{v}_{jg}$ il corrispondente autovettore.
L'inverso della matrice di covarianza della classe *g* diviene allora:

$$\mathbf{S}_g^{-1} = \sum_j \mathbf{v}_{jg} \cdot \mathbf{v}_{jg}^{\mathbf{T}} / \lambda_{jg}$$

e lo *score* discriminante diviene:

$$D_g(\mathbf{x}_i) = \left(\mathbf{x}_i - \bar{\mathbf{x}}_g\right)^{\mathrm{T}} \sum_j \frac{\mathbf{v}_{jg} \cdot \mathbf{v}_{jg}^{\mathrm{T}}}{\lambda_{jg}} \left(\mathbf{x}_i - \bar{\mathbf{x}}_g\right) + \sum_j ln\, \lambda_{jg} - 2\, ln\, P_g$$

La regola di classificazione di SIMCA minimizza un rapporto tra una varianza residua individuale e una varianza residua media, dato dal valore di *F*.
Poichè le due quantità che costituiscono il rapporto *F* sono dei residui, queste si possono ottenere in modo equivalente considerando per ogni classe le $p - M_g$ componenti residue (non-significative).
La regola di classificazione SIMCA diviene:

$$D_g(\mathbf{x}_i) = \left(\mathbf{x}_i - \bar{\mathbf{x}}_g\right)^{\mathrm{T}} \mathbf{S}_g^{-1}\left(M_g\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}_g\right)$$

dove

$$\mathbf{S}_g^{-1}\left(M_g\right) = \frac{\sum_{j=M_g+1}^{p} \mathbf{v}_{jg}\mathbf{v}_{jg}^{\mathbf{T}}}{\sum_{j=M_g+1}^{p} \lambda_{jg}}$$

Il valore di M_g , il numero di componenti significative di ogni classe, viene calcolato minimizzando una stima validata di:

$$\left\| \mathbf{S}_g - \sum_{m=1}^{M_g} \lambda_{mg}\mathbf{v}_{mg}\mathbf{v}_{mg}^{\mathbf{T}} \right\|$$

Si può inoltre osservare come nella regola di classificazione di SIMCA vengano a mancare gli utlimi due termini della classica regola di classificazione bayesiana. Questo significa che SIMCA presuppone una probabilità a priori uguale per tutte le classi e che, in ogni caso, trascura il contributo $\sum_j \ln \lambda_{jg}$ legato alla variabilità totale della classe.

6.7 - I metodi di classificazione ad albero

In questi ultimi anni sono stati sviluppati dei metodi di classificazione che si basano sulla costruzione di una sequenza di partizioni binarie dei dati (*binary split*) in grado di creare un *albero decisionale* come regola di classificazione (*binary decision tree*).
Gli alberi decisionali (Fig. 6-7) sono costituiti da una **radice** (*root*, il nodo superiore), ove tutti gli oggetti sono insieme; da **nodi** (*knots*, i punti intermedi dell'albero), ove gli oggetti sono provvisoriamente collocati durante la procedura di classificazione e da **foglie** (*leaves*, i nodi terminali dell'albero), a ciascuna delle quali è associata una classe e ove gli oggetti sono collocati al termine della sequenza decisionale.

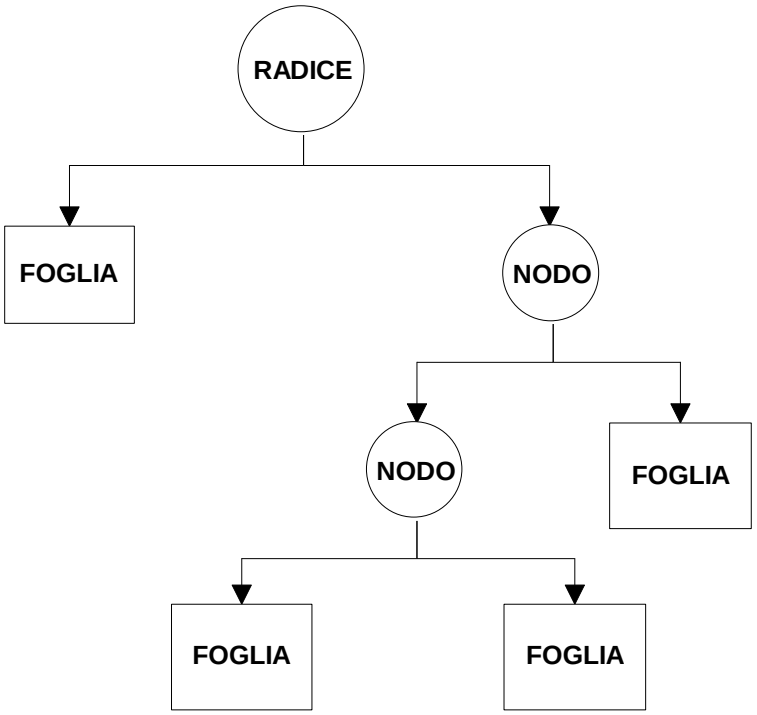


FIG. 6-7

6.8 - Il metodo CART

Il metodo di classificazione ad albero più noto prende il nome di ***CART*** (*Classification And Regression Tree*). Secondo questo metodo, ciascun nodo *k* dell'albero (il luogo ove avviene la partizione binaria) è caratterizzato da una singola variabile $j(k)$ e da un valore di soglia per questa variabile $t_j(k)$.
Per ogni *i*-mo oggetto, questo valore consente di rispondere alla domanda:

$$x_{ij(k)} \leq t_j(k) \,?$$

Se la risposta è *si*, il campione viene classificato nel ramo di sinistra dell'albero; in caso contrario, viene classificato nel ramo di destra. Partendo dalla radice, ogni oggetto viene classificato percorrendo sequenzialmente l'albero decisionale e confrontando i valori di soglia con i corrispondenti valori che le variabili assumono per l'oggetto considerato.
Per ogni nodo, viene selezionata la variabile che fornisce la migliore separazione dei dati, normalmente minimizzando l'indice di diversità di Gini, che misura l'impurezza di ciascun nodo (cioè la presenza di oggetti appartenenti a classi diverse).
La struttura ottimale dell'albero viene determinata mediante una procedura di validazione. L'albero decisionale così costruito ha sovente il vantaggio della semplicità, visto che ogni passo è percorribile esaminando semplicemente i valori di una sola variabile alla volta. Inoltre il motodo *CART* è invariante alla scalatura dei dati, è robusto alla presenza di *outliers* e il risultato fornisce direttamente una selezione delle variabili migliori per la separazione delle classi.

6.9 - Il metodo LDCT

Un altro metodo di classificazione ad albero è il metodo ***LDCT*** (*Linear Discriminant Classification Tree*). In questo caso ogni separazione binaria avviene in modo multivariato utilizzando un vettore discriminante ottenuto dall'analisi discriminante lineare (LDA). Per effettuare un'analisi discriminante in ogni nodo è necessario separare di volta in volta gli oggetti delle diverse classi ancora rappresentate nel nodo in due gruppi. Su questi due gruppi viene calcolato il vettore discriminante.
Sia la separazione delle classi nei due gruppi sia la dimensione complessiva dell'albero viene decisa dall'utente. Viene selezionato l'albero decisionale (e la

rispettiva sequenza di separazione delle classi) che permette di ottenere il miglior risultato (minimo *errore percentuale*) mediante la procedura di validazione *leave-one-out*.
Nel dividere di volta in volta le classi in due gruppi è anche possibile omettere dal calcolo del vettore discriminante una o più classi: in questo caso il vettore discriminante viene calcolato sugli oggetti delle classi definite nei due gruppi; gli oggetti restanti vengono assegnati successivamente al nodo di sinistra o di destra in base al vettore discriminante calcolato.
Il metodo produce in generale alberi decisionali con una struttura più semplice del metodo CART. Tuttavia, ad ogni passo la procedura di classificazione di un oggetto richiede il calcolo del corrispondente vettore discriminante e il migliore albero decisionale deve essere ricercato manualmente tra i tanti possibili. Evidentemente, questo ultimo problema non si pone quando esistono solo due classi.

Esempio

Diversi metodi di classificazione sono stati applicati sui dati OLITOS. Si tratta di 120 campioni di olio d'oliva provenienti dalla Toscana, divisi in 4 classi (Tab.6-5).

ID	*Nome della classe*	*n. campioni*	*prob. a priori*
1	Arno	50	0.25
2	Valdichiana	25	0.25
3	Costa e Sud	34	0.25
4	Altri olii toscani	11	0.25

TAB. 6-5

Per ogni campioni sono state misurate 25 variabili chimiche. Si osservi che la quarta classe in realtà raccoglie tutti i campioni non appartenenti alle prime tre classi e quindi provenienti da siti non omogenei.
A questi dati sono stati applicati diversi metodi di classificazione, assumendo che ciascuna classe abbia la stessa probabilità a priori ($P_g = 1/4$).

Il metodo K-NN è stato provato cercando la migliore soluzione tra valori di *k* compresi tra 1 e 7, cioè minimizzando il rischio di errore in predizione (MRcv %).

I dati sono stati preventivamente autoscalati. Il valore ottimale trovato corrisponde a k = 5. In Tab. 6-6 sono riportati i valori ottenuti di MRcv %, mentre in Fig.6-8 viene riportato il corrispondente grafico.

K	*MRcv %*
1	32.06
2	30.40
3	29.87
4	28.87
5	**28.13**
6	29.67
7	30.90

TAB. 6-6

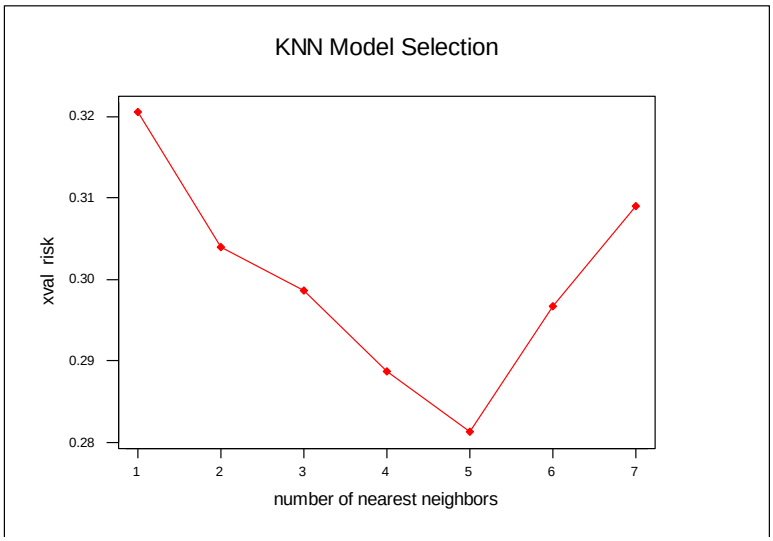


FIG. 6-8

La Fig. 6-9 è la rappresentazione grafica delle due matrici di confusione ottenute in *fitting* e in predizione.

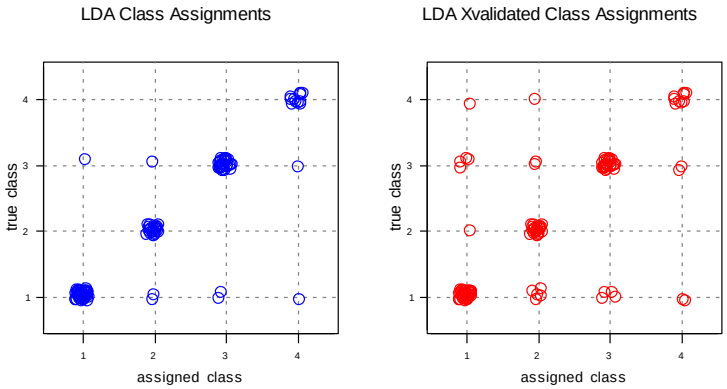


FIG. 6-9

Nella Fig. 6-10 sono rappresentate, per ogni coppia di classi, le distanze degli oggetti dal centroide di ciascuna classe, in modo del tutto analogo a quanto visto per i metodi di *cluster analysis*.

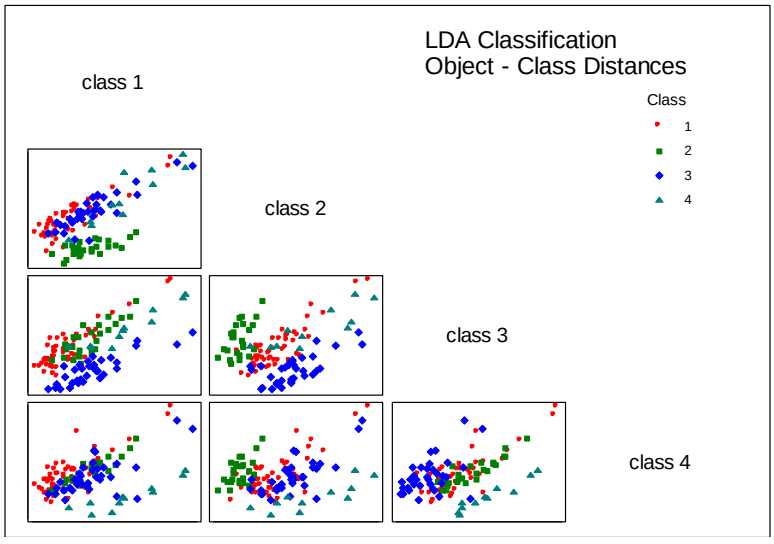


FIG. 6-10

Nella Fig. 6-11 sono rappresentate, per ogni coppia di classi, le probabilità di appartenenza di ciascun oggetto alle classi. La posizione degli oggetti relativamente alle coppie di assi coordinati è quindi inversa a quella del grafico precedente per le distanze: infatti, a piccole distanze dal centroide di una classe corrisponde un'alta probabilità di appartenenza alla stessa.

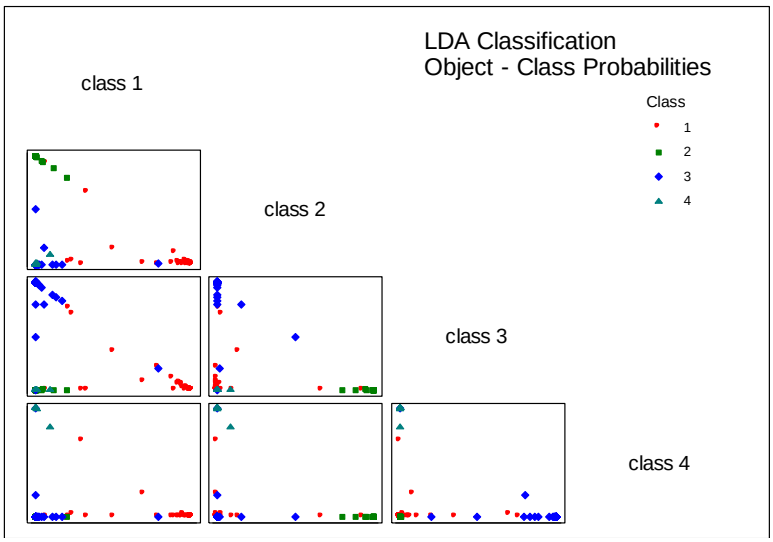


FIG. 6-11

Il metodo RDA viene utilizzato ottimizzando i due parametri λ e γ, minimizzando il rischio di errore in predizione. L'ottimizzazione è stata eseguita su una griglia di 5 valori per ciascuno dei 2 parametri, compresi tra 0 e 1, per un totale di 25 punti della griglia. I valori ottenuti sono riportati in Tab. 6-7.
I risultati relativi al caso $\lambda = 0, \gamma = 0$, corrispondente all'analisi discriminante quadratica (QDA) non sono disponibili in quanto la classe 4 ha un numero di oggetti inferiore al numero di variabili. Il risultato migliore corrisponde alla coppia di valori $\lambda = 0.75, \gamma = 0$, cioè ad un'analisi discriminante *unbiased* ($\gamma = 0$) e *quasi lineare* (per LDA, $\lambda = 1$).

ID	*lambda*	*gamma*	*MR %*	*MRcv %*	
1	**0.00**	**0.00**	-	-	QDA
2	0.00	0.25	29.15	42.21	
3	0.00	0.50	33.12	43.91	
4	0.00	0.75	38.60	46.65	
5	**0.00**	**1.00**	47.82	55.33	WNMC
6	0.25	0.00	1.00	20.13	
7	0.25	0.25	25.68	38.78	
8	0.25	0.50	27.92	37.75	
9	0.25	0.75	34.67	38.18	
10	0.25	1.00	43.08	50.83	
11	0.50	0.00	1.74	15.39	
12	0.50	0.25	23.92	34.71	
13	0.50	0.50	29.16	36.98	
14	0.50	0.75	32.41	36.91	
15	0.50	1.00	41.28	49.80	
16	**0.75**	**0.00**	**3.97**	**14.16**	MIGLIORE
17	0.75	0.25	25.66	31.21	
18	0.75	0.50	26.16	37.75	
19	0.75	0.75	32.64	36.41	
20	0.75	1.00	43.55	50.53	
21	**1.00**	**0.00**	4.71	16.93	LDA
22	1.00	0.25	25.44	31.48	
23	1.00	0.50	24.66	32.48	
24	1.00	0.75	32.88	35.65	
25	**1.00**	**1.00**	44.29	49.53	NMC

TAB. 6-7

La Fig. 6-12 riporta graficamente i valori del rischio di errore calcolati in *fitting* e in predizione per i 25 punti di griglia, disposti sequenzialmente, ove, per ogni valore di λ, sono riportati i corrispondenti valori di γ. La freccia indica il valore ottimale in predizione.

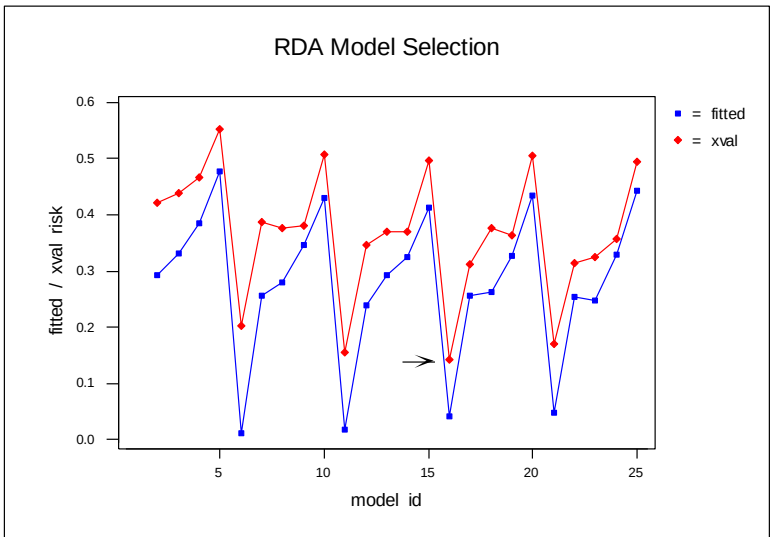


FIG. 6-12

Il metodo SIMCA è stato utilizzato secondo la procedura tradizionale e cioè effettuando l'autoscalatura dei dati indipendentemente per ogni classe e ricercando in validazione il numero di componenti significative di ogni classe. La tabbella seguente riporta il numero di componenti significative e la varianza totale spiegata da queste ultime per ogni classe.

Classe	*Numero di PC*	*Dev. Std. Cum. %*
1	7	66.3
2	1	93.7
3	3	84.8
4	6	62.7

TAB. 6-8

Le distanze tra le classi SIMCA sono riportate nella Tab. 6-9.

classi	*1*	*2*	*3*	*4*
1	0	1.741	1.412	1.251
2	1.004	0	1.404	1.178
3	1.045	1.881	0	1.278
4	1.371	2.453	1.550	0

TAB. 6-9

Nonostante si tratti formalmente di una matrice delle distanze (Tab.6-9), in questo caso la matrice non è simmetrica. Infatti, ad esempio, per calcolare la distanza degli oggetti della classe 1 dal modello della classe 2, gli oggetti vengono proiettati nello spazio della classe in base al modello della classe 2. Viceversa, per calcolare la distanza degli oggetti della classe 2 dal modello della classe 1, gli oggetti vengono proiettati nello spazio della classe 1 in base al modello della classe 1. Ciò significa che essendo la distanza 1 - 2 (1.741) maggiore della distanza 2 - 1 (1.004), è più facile che oggetti della classe 2 vengano erroneamente classificati nella classe 1 che viceversa. Analogamente sarà più facile classificare erroneamente gli oggetti della classe 2 in 4 (1.178) rispetto agli oggetti della classe 4 in 2 (2.453).
La Tab. 6-10 riporta il potere modellante ed il potere discriminante di ciascuna variabile originale. Sono evidenziate le prime 4 variabili più importanti.

Variabile	*Potere modellante*	*Potere discriminante*
x1	88.4	**7.453**
x2	87.9	4.125
x3	88.1	3.064
x4	87.4	2.554
x5	89.6	**6.233**
x6	84.6	**6.442**
x7	81.9	5.876
x8	82.8	3.270
x9	91.2	5.054
x10	85.2	3.897
x11	83.4	2.664
x12	87.4	3.465
x13	87.4	2.706
x14	88.2	4.427
x15	83.9	4.058
x16	92.1	4.267
x17	84.7	5.096
x18	91.6	5.888
x19	90.0	**6.075**
x20	87.0	4.807
x21	85.1	2.320
x22	88.2	4.646
x23	82.2	1.879
x24	84.3	4.572
x25	84.2	3.153

Tab. 6-10

Il metodo CART fornisce il diagramma di Fig. 6-13 che riporta i valori in *fitting* e in predizione associati a ciascuno nodo. I nodi sono numerati in senso inverso, cioè il nodo numero 12 è la radice dell'albero a cui è associato il valore del rischio di errore per il caso *no-model* (75.0 %); dopo la prima divisione binaria in questo nodo, i valori del rischio di errore in *fitting* e in validazione diminuiscono (nodo 11). Come si osserva, il numero ottimale di nodi in predizione è 4 (nodo 9), corrispondente a 3 divisioni binarie.

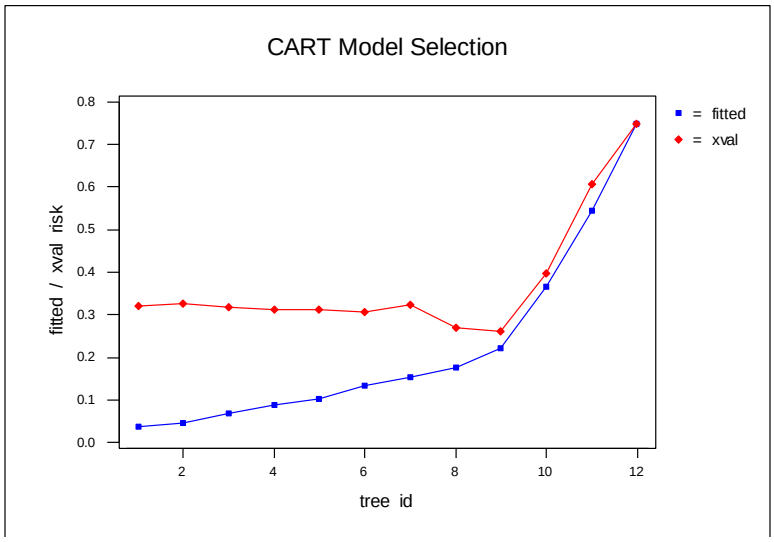


FIG. 6-13

Come si può notare dalla Fig. 6-14, le variabili utilizzate nel metodo CART, dopo il processo di validazione, sono solo 3, una per ogni divisione binaria, e precisamente: x1, x15 e x5. In particolare il primo passo ci dice che gli oggetti con valori maggiori di 1.31 della variabile x1 appartengono alla classe 4 (ramo di destra). Per gli altri oggetti si considera il secondo passo: valori maggiori di 91.04 della variabile x15 identificano gli oggetti che vengono assegnati alla classe 2. Gli oggetti restanti vengono classificati in base all'ultima divisione binaria: oggetti con valori maggiori di 12.87 della variabile x5 definiscono oggetti appartenenti alla classe 3, mentre gli oggetti restanti appartengono alla classe 1.
Le variabili x1 e x5 compaiono anche tra le variabili più importanti ottenute col metodo SIMCA. Le variabili x6 e x19, che compaiono anch'esse tra le variabili importanti in SIMCA, pur non essendo state selezionate dal metodo CART, risultano tra le più importanti dopo quelle selezionate (dati non riportati).
Anche il metodo LDCT fornisce un albero decisonale semplice nella struttura e ricavato mediante una procedura di validazione. La prima separazione avviene calcolando il vettore discriminante lineare su due gruppi costituiti il primo dagli oggetti delle classi 1 e 3, il secondo dagli oggetti delle classi 2 e 4 (Fig. 6-15).

Dopo la prima separazione, su ciascun nodo viene effettuata la seconda separazione: in un caso dividendo le classi 1 e 4 dalla 3; nell'altro caso dividendo le classi 2 e 4.
Il risultato ottenuto (errore percentuale e rischio di errore) è migliore di quello ottenuto con CART, grazie al fatto che ogni separazione avviene in modo multivariato (il vettore discriminante).

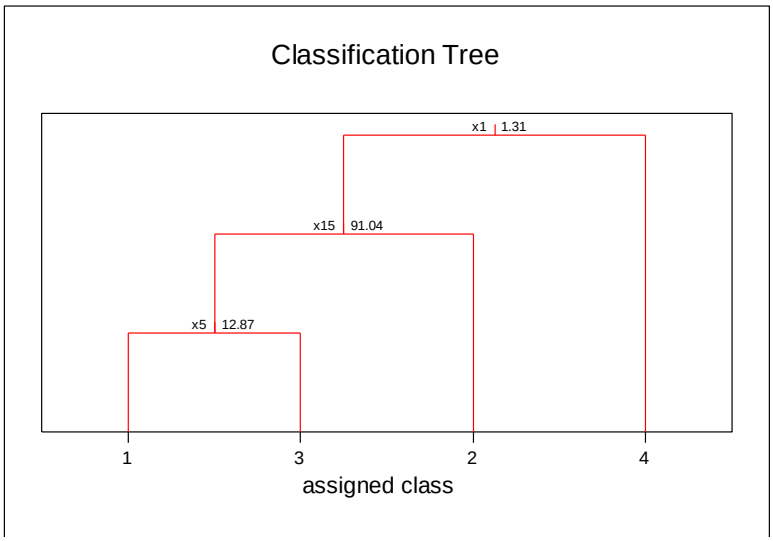


FIG. 6-14

Le Tab. 6-11 riassume i risultati ottenuti con i diversi metodi, in predizione (classi predette) e in *fitting* (classi calcolate), riportando i rispettivi valori dell'errore di classificazione e del rischio di errore.
Come si può osservare, il risultato migliore è quello ottenuto col metodo RDA (minimo rischio di errore). La notevole differenza tra l'errore percentuale ed il rischio di errore nel metodo k-nn è dovuto alla notevole penalizzazione per l'alto numero di errori commessi nel classificare oggetti di una classe dimensionalmente piccola (classe 4, 11 oggetti). Il metodo CART risulta inferiore agli altri dal punto di vista dei risultati globali, ma rimane apprezzabile per la semplicità del modello costituito da 3 passi monovariati che comportano quindi l'utilizzo di sole 3 variabili.

Naturalmente i risultati qui riportati sono solo indicativi in quanto per ogni metodo sarebbe opportuno procedere ad un raffinamento del modello, valutando, con considerazioni specifiche, la presenza di eventuali *outliers* e l'adeguatezza del metodo prescelto in funzione dei reali obiettivi.

	Errore %	*Rischio Er.%*	*Errore %*	*Rischio Er %*
Metodo	*predetto*	*predetto*	*calcolato*	*calcolato*
No-model	58.3	75.0	58.3	75.0
K-NN	17.5	28.1	-	-
LDA	18.3	16.9	6.7	4.7
RDA	**16.7**	**14.2**	5.8	4.0
SIMCA	**16.7**	19.2	5.0	4.9
CART	26.7	26.2	22.5	17.6
LDCT	**16.7**	17.0	6.7	6.5

TAB. 6-11

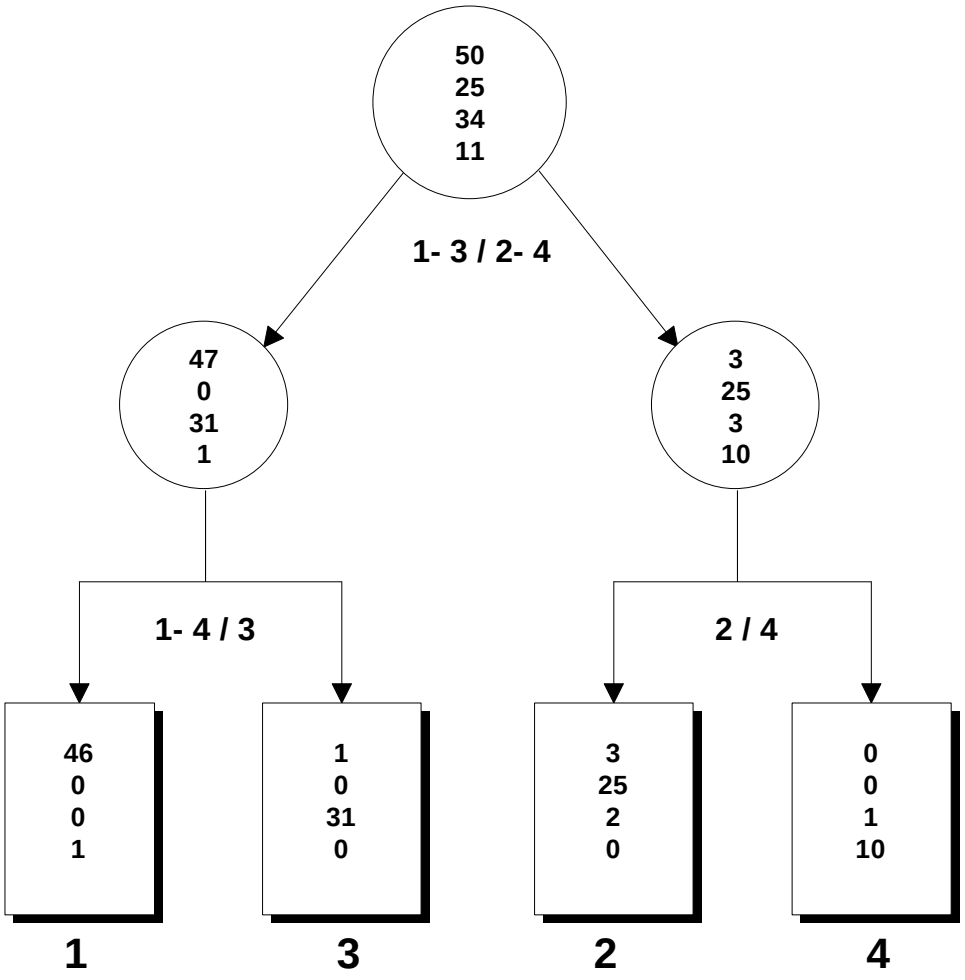


FIG. 6 - 15

Le due Tabelle 6-12 e 6-13 riportano le matrici di confusione ottenute rispettivamente in predizione e in fitting. Il metodo K-NN è un metodo che opera implicitamente in predizione e quindi non sono riportati i valori ottenuti per le classi calcolate.

Classi					
vere	*predette*	*1*	*2*	*3*	*4*
1	K-NN	44	3	3	0
	LDA	39	5	4	2
	RDA	40	4	4	2
	SIMCA	45	2	0	3
	CART	36	7	5	2
	LDCT	40	5	5	0
2	K-NN	1	24	0	0
	LDA	1	24	0	0
	RDA	1	24	0	0
	SIMCA	1	21	2	1
	CART	4	20	1	0
	LDCT	0	25	0	0
3	K-NN	2	2	29	1
	LDA	4	2	26	2
	RDA	4	2	26	2
	SIMCA	5	2	26	1
	CART	6	1	24	3
	LDCT	3	3	27	1
4	K-NN	5	2	2	2
	LDA	1	1	0	9
	RDA	0	1	0	10
	SIMCA	1	0	2	8
	CART	2	0	1	8
	LDCT	2	0	1	8

TAB. 6 - 12

	Classi				
vere	*calcolate*	*1*	*2*	*3*	*4*
1	K-NN	-	-	-	-
	LDA	45	2	2	1
	RDA	45	2	2	1
	SIMCA	50	0	0	0
	CART	36	7	3	4
	LDCT	46	3	1	0
2	K-NN	-	-	-	-
	LDA	0	25	0	0
	RDA	0	25	0	0
	SIMCA	0	23	1	1
	CART	3	21	0	1
	LDCT	0	25	0	0
3	K-NN	-	-	-	-
	LDA	1	1	31	1
	RDA	1	0	32	1
	SIMCA	3	1	30	0
	CART	5	1	25	3
	LDCT	0	2	31	0
4	K-NN	-	-	-	-
	LDA	0	0	0	11
	RDA	0	0	0	11
	SIMCA	0	0	0	11
	CART	0	0	0	11
	LDCT	1	0	0	10

TAB. 6-13

🕮 Bibliografia

P.A. LACHENBRUCH (1975). *Discriminant Analysis.* Hafner Press, New York, NY.

D.J. HAND (1981). *Discrimination and Classification.* Wiley, Chichester.

I.E. FRANK E J.H. FRIEDMAN (1989). *Classification: oldtimers and newcomers.* J. Chemometrics, **3**, 463.

L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN E C.J. STONE (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

J.H. FRIEDMAN (1989). *Regularized Discriminant Analysis.* J. Am. Statist. Assoc., **84**, 165

R. TODESCHINI E E. MARENGO (1992). *Linear Discriminant Classification Tree (LDCT): a user-driven multicriteria classification method.* Chemometrics and Intelligent Laboratory Systems, **16**, 25.

7

I METODI DI REGRESSIONE: I FONDAMENTI

7.1 - Introduzione

I metodi di regressione sono tra i metodi matematici più utilizzati in molte discipline scientifiche e come tali sono anche largamente utilizzati in chemiometria in quanto capaci di dare una risposta ad uno dei problemi più comuni: cercare la migliore relazione tra un insieme di variabili che descrive gli oggetti studiati e un insieme di riposte misurate per gli stessi oggetti. Da una parte, la forma della relazione ci descrive la modalità con cui la descrizione del sistema si raccorda con la misura sperimentale (*fitting*), e, dall'altra, il modello ottenuto, una volta verificata la sua qualità (*validazione*), ci consente di *predire le future risposte* di oggetti per i quali conosciamo soltanto le variabili che li descrivono ma non le misure sperimentali.
In modo più rigoroso, i metodi di regressione sono metodi matematici che forniscono informazioni circa relazioni funzionali quantitative tra una risposta **y** e un certo numero *p* di descrittori indipendenti $\mathbf{x}_1$, ..., $\mathbf{x}_p$:

$$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$$

Il problema generale della regressione si riconduce quindi a:

- stabilire il **tipo** di modello (la relazione funzionale *f*)
- determinare i **parametri** del modello
- valutare l'**attendibilità** del modello

Sia **X** la matrice dei dati con *n* righe (le osservazioni) e *p* colonne (le variabili), **y** il vettore delle *n* risposte sperimentali, **b** il vettore dei coefficienti del modello, di dimensione *p*', dove *p*' è il numero dei parametri del modello.
Ad esempio, un modello lineare in *p* variabili lineari con intercetta b_0 è definito nel seguente modo:

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij}$$

Un modello in *p* variabili è *lineare* se la risposta è una combinazione lineare delle variabili del modello, cioè se i coefficienti **b** sono dei fattori moltiplicativi delle variabili.
Un modello lineare in *p* variabili è quindi anche:

$$y_i = b_0 + \sum_{j=1}^{p} b_j f_j(x_{ij})$$

In generale, dalla matrice originale dei dati **X** si ottiene la **matrice del modello** $\mathbf{X_M}$. Questa matrice può contenere, oltre alle colonne della matrice **X**, anche delle colonne aggiuntive che risultano da trasformazioni delle colonne originali, quali, ad esempio, termini quadratici $(\mathbf{x}_1^2, \mathbf{x}_2^2)$ e termini misti $(\mathbf{x}_1 \times \mathbf{x}_2)$. In particolare, quando il modello prevede il termine b_0 (l'intercetta), la matrice del modello viene costruita dalla matrice dei dati aggiungendo una colonna di 1, che indica la presenza nel modello di un termine costante.

Nota. Per semplicità nella scrittura delle espressioni, la matrice del modello si scriverà con lo stesso simbolo **X** della matrice originale dei dati, con *p*' parametri.
Si tenga quindi presente che la matrice del modello non coincide, di norma, con la matrice iniziale dei dati.

7.2 - Il metodo dei minimi quadrati ordinari

In termini matriciali, il problema delle regressione lineare col metodo dei **minimi quadrati ordinari** (***Ordinary Least Squares, OLS***) è rappresentato dal seguente modello:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

dove β è il vettore dei coefficienti veri da stimare, **X** è la matrice del modello ed **e** è il vettore degli errori; l'analisi dimensionale è la seguente:

$$(n, 1) = (n, p') (p', 1) + (n, 1)$$

con $p' = p + 1$
In termini non matriciali, possiamo scrivere esplicitamente la relazione tra la risposta *y* dell'*i*-esimo oggetto e la sua descrizione con le variabili indipendenti con la combinazione lineare:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

La soluzione consiste nel determinare il vettore dei coefficienti **b**. I seguenti passaggi algebrici portano alla soluzione cercata:

$$\mathbf{y} = \mathbf{X}\,\mathbf{b}$$

$$\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b}$$

$$\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b}$$

Poichè $\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X} = \mathbf{I}$, la soluzione **OLS** risulta:

$$\mathbf{b}_{ols} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

L'analisi dimensionale è la seguente:

$$(p', 1) = (p', p') (p', n) (n, 1)$$

con $p' = p + 1$
La varianza del vettore dei coefficienti è una misura della stabilità del modello di regressione ed è data dall'espressione:

$$V(\mathbf{b}_{\mathrm{OLS}}) = \sigma^2 \cdot tr\left(\mathbf{X^T X}\right)^{-1} = \sigma^2 \cdot \sum_{j=1}^{p}\left(1/\lambda_p\right)$$

dove σ^2 è la varianza dovuta all'errore sperimentale, **X** è la matrice del modello, λ sono gli autovalori della matrice $\mathbf{X^T X}$, la matrice di dispersione del modello. Si osservi che la varianza dei coefficienti dipende dal prodotto di due contributi separati, il primo (σ^2) che dipende solo dalla risposta ed il secondo che dipende solo dallo spazio dei predittori. In particolare, gli elementi della diagonale principale di $(\mathbf{X^T X})^{-1}$ sono proporzionali all'incertezza con cui vengono determinati i parametri del modello. Dalla rappresentazione della varianza dei coefficienti nello spazio ortogonale di cui λ sono gli autovalori (ultimo termine dell'espressione), si può osservare che l'eventuale presenza di collinearità tra le variabili indipendenti rende l'ultimo autovalore (o gli ultimi) molto piccolo o addirittura nullo: in questi casi la varianza dei coefficienti tende ad infinito o, equivalentemente, il modello ottenuto è completamente instabile.
Una volta calcolati i coefficienti del modello, è possibile determinare il vettore delle risposte calcolate $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

Quest'ultima espressione costituisce il **modello di regressione**, costituito dunque dai descrittori definiti nella matrice del modello e dai **coefficienti di regressione b** ottenuti. Per quanto riguarda i coefficienti di regressione, di particolare interesse sono i **coefficienti di regressione standardizzati b'**, ottenuti dall'espressione

$$b'_j = b_j \cdot \frac{s_j}{s_y}$$

dove s_y e s_j sono le deviazioni standard della risposta e del *j*-esimo predittore. Questi coefficienti corrispondono ai coefficienti di regressione calcolati per variabili autoscalate e rappresentano i *pesi* di ciascuna variabile nel modello di regressione. Questi coefficienti danno quindi una misura dell'importanza di ciascuna variabile nel modello e la somma dei loro quadrati può essere considerata una misura della complessità del modello.

Sostituendo al vettore dei coefficienti **b** l'espressione ricavata in precedenza, si ottiene un'importante relazione:

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\,\mathbf{y}$$

dove la matrice **H**, di dimensione *n* x *n*, mette in relazione le risposte calcolate con quelle sperimentali e prende il nome di **matrice d'influenza** (*matrice dei leverages* o ***hat matrix*)**. La matrice **H** è definita come:

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

Gli elementi interessanti di questa matrice sono gli elementi diagonali h_{ii}, detti anche *funzioni di varianza* dell'oggetto *i*-esimo, per i quali valgono le seguenti proprietà:

$$h_{\min} = 1/n \quad \sum_i h_{ii} = p' \qquad \bar{h} = p'/n \qquad h^* > 3p'/n$$

dove h^* è un valore di controllo oltre il quale il dato può essere considerato *influente* nel determinare i parametri del modello. Il singolo valore di *h* per l'*i*-esimo oggetto è dato da:

$$h_{ii} = x_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}x_i$$

Il valore minimo del *leverage* si trova in corrispondenza del centroide dello spazio del modello, mentre i valori massimi sono relativi ai punti più distanti dal centro del modello. Questa proprietà del *leverage* è di grande importanza in quanto consente di utilizzare i valori di *h* per valutare la "distanza" di ciascun oggetto dal centro dello spazio del modello.
Per gli *n* oggetti utilizzati nella costruzione del modello i valori di *h* sono sempre compresi tra 1/*n* e 1, ma, nell'applicazione del modello a nuovi oggetti, il valore di *h* può essere anche molto maggiore di 1: si ha così a disposizione un indice che misura il *grado di estrapolazione* dal modello. E' evidente quindi che nell'applicare a fini predittivi il modello ottenuto a nuovi campioni, un alto valore del *leverage* (ad esempio, maggiore di h^*) deve renderci molto più cauti

nell'accettare il valore predetto della risposta. In generale, è preferibile avere oggetti che abbiano circa la stessa influenza nel determinare il modello di regressione. Questo si può ottenere soltanto con oggetti i cui valori delle variabili che li descrivono sono ottenuti mediante un disegno sperimentale, ma non accade con oggetti definiti in spazi sperimentali non controllati direttamente dallo sperimentatore.

7.3 - I parametri di valutazione dei modelli di regressione

Per ogni modello di regressione si assume, come situazione di riferimento o *modello di ordine zero*, la quantità riferita al valor medio della risposta, detta **somma totale dei quadrati** (***TSS***, ***Total Sum of Squares***):

$$TSS = \sum_i \left(y_i - \bar{y}\right)^2$$

Un modello di regressione è tanto migliore quanto più piccola è la **somma dei quadrati dei residui** (***RSS***, ***Residual Sum of Squares***):

$$RSS = \sum_i \left(y_i - \hat{y}_i\right)^2$$

ottenuta dalla differenza tra ciascun valore sperimentale della risposta e dalla risposta calcolata.
Nello stesso tempo, un modello di regressione è tanto migliore quanto più grande è la **somma dei quadrati del modello** (***MSS, Model Sum of Squares***):

$$MSS = \sum_i \left(\hat{y}_i - \bar{y}\right)^2$$

Valgono pertanto le seguenti relazioni:

$$TSS = MSS + RSS$$

$$\frac{MSS}{TSS} \equiv R^2 = 1 - \frac{RSS}{TSS}$$

dove R^2 viene chiamato **coefficiente di determinazione** e, moltiplicato per 100, rappresenta la **varianza percentuale spiegata** dal modello.
La quantità R (o comunemente anche r) è nota come **coefficiente di correlazione multipla** e rappresenta la correlazione tra la risposta sperimentale **y** e i predittori $\mathbf{x}_j$. E' definita come

$$R = \sqrt{1 - \frac{RSS}{TSS}}$$

Nota. Definiamo la correlazione tra la risposta **y** e p predittori come $R_{y(1,\ldots,p)}$.
Ad esempio, $R_{y(1)}$ e $R_{y(1,2)}$ rappresentano le correlazioni tra la risposta **y** e, rispettivamente, una variabile predittrice e due variabili predittrici (non le variabili numero 1 e numero 2). Le *proprietà fondamentali* di R sono le seguenti:

1. $0 \le R_{y(1,\ldots,p)} \le 1$

2. Se $R_{y(1,\ldots,p)} = 0 \;\Rightarrow\; R_{y(j)} = 0 \quad j = 1,\ldots,p,$ cioè se la correlazione è nulla, allora tutte le correlazioni a coppia tra la risposta **y** e ciascun singolo predittore sono nulle.

3. Se $R_{y(1,\ldots,p)} = 1$, allora la risposta **y** è una perfetta combinazione lineare dei p predittori.

4. Nel caso di un singolo predittore **x**, $R_{y(1)} = |r_{yx}|$, cioè il coefficiente di correlazione di **y** con il singolo predittore coincide con il valore assoluto del coefficiente di correlazione tra **y** e **x**.

5. $R^2_{y(1)} \le R^2_{y(1,2)} \le R^2_{y(1,2,3)} \le \ldots \le R^2_{y(1,\ldots,p)}$, cioè per qualsiasi aumento del numero di predittori, il coefficiente di correlazione multipla non può diminuire mai.

In aggiunta al parametro precedente, è possibile definire un parametro, chiamato ***R*² *adjusted***, che tiene conto anche dei gradi di libertà con cui è calcolato R^2:

$$R^2_{adj} = 1 - \frac{RSS_{p'}/(n-p')}{TSS/(n-1)} = 1 - (1-R^2)\cdot\left(\frac{n-1}{n-p'}\right)$$

Il parametro R^2 *adjusted* è stato ideato per valutare la convenienza nell'aggiungere una variabile al modello, visto che, per la proprietà 5 della nota precedente, questo non può essere fatto con il parametro *R*. Infatti, il parametro R^2_{adj} presenta un massimo per la complessità ottimale del modello e ridiscende quando l'aggiunta di una variabile al modello non è adeguatamente compensata da un significativo aumento di *R*. Entrambi i parametri precedenti misurano esclusivamente la capacità del modello di raccordarsi contemporaneamente con tutti gli *n* oggetti utilizzati nella costruzione del modello; essi misurano ciò che viene chiamato *fitting*. Per ottenere dei parametri che misurino la capacità predittiva del modello ottenuto è necessario utilizzare le tecniche di validazione. In questo caso, la somma dei quadrati dei residui viene effettuata utilizzando i valori delle risposte predette al posto delle risposte calcolate. Questa grandezza, formalmente identica a *RSS*, viene chiamata ***PRESS*** (*PRedictive Error Sum of Squares*) e viene calcolata da:

$$PRESS = \sum_i \left(y_i - \hat{y}_{i/i}\right)^2$$

dove $\hat{y}_{i/i}$ indica il valore predetto per l'*i*-esimo campione da un modello in cui il campione non è stato preso in considerazione nella costruzione del modello stesso.
Utilizzando *PRESS* al posto di *RSS*, otteniamo la percentuale di varianza spiegata dal modello in predizione:

$$Q^2 \equiv R^2_{cv} = 1 - \frac{PRESS}{TSS}$$

Diversamente da R^2 e in modo simile a R^2_{adj}, quest'ultimo parametro presenta un massimo per la complessità ottimale del modello e ridiscende ogni volta che aggiungiamo al modello variabili non predittive. Tuttavia, diversamente dal parametro R^2_{adj}, R^2_{cv} viene esplicitamente valutato rispetto al potere predittivo del modello e non al livello di *fitting* del modello.

Due utili parametri associati a *RSS* e *PRESS* sono, rispettivamente, ***SDEC*** (*Standard Deviation Error in Calculation*, o anche ***SEC***) e ***SDEP*** (*Standard Deviation Error in Prediction*, o anche ***SEP***), definiti come segue:

$$SDEC = \sqrt{\frac{RSS}{n}} \quad \text{e} \quad SDEP = \sqrt{\frac{PRESS}{n}}$$

Questi parametri hanno il vantaggio di essere dimensionalmente confrontabili con la risposta studiata.
Un test di carattere generale utilizzato per valutare la significatività globale di un modello di regressione è il ***test F di Fisher***, che consente di analizzare la significatività della differenza tra due varianze attraverso il valore del loro rapporto (v. App.B).
Dopo aver effettuato un'analisi di regressione, è possibile verificare se la relazione funzionale che è stata trovata tra la variabile dipendente *y* e le variabili indipenenti x_j è significativa. A questo scopo, si può effettuare un'**analisi della varianza** (***ANOVA***) al fine di ottenere una statistica *F di Fisher*, dove l'ipotesi nulla e l'ipotesi alternativa sono così definite:

H_0: tutti i coefficienti di regressione b_j (j = 1, p) sono uguali a zero

H_1: almeno un coefficiente di regressione è diverso da zero, cioè esiste un modello.

Questo significa che se l'ipotesi H_0 è vera (cioè è falsa l'ipotesi alternativa H_1), il modello di regressione non esiste in quanto tutti i coefficienti del modello non sono significativamente diversi da zero. Quindi, per un buon modello di regressione il valore calcolato di *F* deve essere grande.
La variazione totale dovuta alla regressione è definita da *TSS*, la variazione spiegata dalla regressione da *MSS* e la variazione residua - non spiegata dal modello di regressione - da *RSS*.

Nota. Secondo la simbologia utilizzata comunemente nell'analisi della varianza, i termini precedenti sono definiti rispettivamente come: SS_T , SS_M *e* SS_r.

I gradi di libertà per la variazione totale, quella dovuta al modello e quella residua sono rispettivamente:

$$DF_T = n-1\,, \quad DF_M = p'-1 \quad e \quad DF_r = n-p'$$

dove *n* è il numero di osservazioni e *p'* il numero totale di parametri del modello di regressione. Le varianze vengono quindi calcolate dalle somme dei quadrati diviso i rispettivi gradi di libertà.
Il valore *F* per il test dell'ipotesi nulla è:

$$F = \frac{MSS/(p'-1)}{RSS/(n-p')} = \frac{MS_M}{MS_r}$$

e deve essere confrontato con il valore critico di tabella:

$$F_c\ \alpha(1);(p'-1),(n-p')$$

Il valore di *F* calcolato viene confrontato col valore critico *F* di tabella, fissato il livello di fiducia (comunemente 5% o 1%), relativo ai gradi di libertà al numeratore e denominatore. Se il valore calcolato di *F* è almeno uguale (o superiore) al valore critico (test a 1 coda), l'ipotesi nulla è da rifiutare, ossia è statisticamente probabile che, al livello di significatività prescelto ♋, esista una relazione di dipendenza *f* tra le variabili considerate nella regressione.
In pratica, se la varianza dovuta al modello è significativamente più grande della varianza legata alla parte non spiegata dal modello e al rumore sperimentale, si può affermare che esiste dell'informazione non casuale dovuta alla relazione di regressione che lega la risposta con i predittori.
Questo test viene anche utilizzato per ricercare il miglior sottoinsieme di variabili predittrici nei metodi di regressione *stepwise* (v. Cap. 8).
La radice quadrata della varianza media residua MS_r , quest'ultima denotata anche con $\mathbf{s^2}$, si chiama **errore standard della stima** s_y (o semplicemente **s**)

$$s_y = \sqrt{\frac{RSS}{n-p'}}$$

e fornisce un'indicazione globale dell'accuratezza con cui la funzione usata per modellare i dati descrive la dipendenza di *y* da *x*. Questa grandezza viene

utilizzata come stima dell'errore sperimentale σ^2, che contribuisce a determinare l'errore complessivo nella stima dei coefficienti di regressione.
Questa grandezza viene utilizzata per calcolare gli **intervalli di confidenza** dei parametri della regressione e le **bande di confidenza** entro cui sono compresi, ad un certo livello di confidenza, i valori stimati della variabile dipendente *y*.
Gli *intervalli di confidenza* di ciascun coefficiente di regressione sono così definiti:

$$b_j \pm t_{(\alpha;\, n-p')} \cdot V(b_j)^{1/2} = b_j \pm t_{(\alpha;\, n-p')} \cdot s_y \cdot \sqrt{d^{jj}}$$

dove $V(b_j)$ è la varianza del *j*-esimo coefficiente di regressione, $t_{(\alpha;\, n-p')}$ è il valore critico di tabella della variabile *t* di Student, al livello di significatività α, per $n - p'$ gradi di libertà; s_y è l'errore standard della stima e d^{jj} è l'elemento diagonale *j*-esimo ottenuto dall'inversa della matrice di dispersione del modello, cioè da $(\mathbf{X^T X})^{-1}$.

Esempio

Consideriamo i 17 dati relativi all'insieme REGTEST, con 5 variabili indipendenti e una risposta, riportati nell'Appendice D.
La matrice di correlazione tra le 5 variabili predittrici è riportata in Tab. 7-1.

	x_1	x_2	x_3	x_4	x_5
x_1	1	0.9074	0.9999	0.9357	0.6712
x_2	0.9074	1	0.9071	0.9105	0.4466
x_3	0.9999	0.9071	1	0.9332	0.6711
x_4	0.9357	0.9105	0.9332	1	0.4629
x_5	0.6712	0.4466	0.6711	0.4629	1

TAB. 7-1

Come si può osservare, esiste una forte correlazione tra molte delle variabili considerate. Il metodo di regressione *OLS* fornisce i seguenti risultati:

n.param.	R^2	R^2_{adj}	R^2_{cv}	*F*	*SDEC*	*s*	*SDEP*
5 + 1	99.08	98.67	93.49	31.61	516.5	642.1	1376.2

I valori di tabella del test *F* al livello di significatività di 0.05% e 0.01% per 5 e 11 gradi di libertà sono **3.20** e **5.32**: l'ipotesi nulla deve essere rifiutata.

	b_0	b_1	b_2	b_3	b_4	b_5
coefficienti	1962.94	-15.85	.06	1.59	-4.22	-394.31
dev. stand.	1071.36	97.65	.02	3.09	7.18	209.64
coeff.stand.	-	-0.459	0.214	1.403	-0.082	-0.112
dev. stand.	-	2.828	0.081	2.728	0.139	0.060

TAB. 7-2

I valori dei coefficienti di regressione calcolati e le loro incertezze sono riportate in Tab. 7-2.

7.4 - I metodi diagnostici per la regressione

L'attendibilità di un modello di regressione deve essere sempre controllata mediante un insieme di test, metodi e parametri in grado di fornire un'informazione completa e dettagliata sul modello calcolato.
In particolare, vengono utilizzati i residui e i valori di *leverage* per ottenere una serie di grafici e di test molto efficaci nel valutare il comportamento dei singoli punti nel modello di regressione costruito.
Il grafico più comune utilizzato nella diagnostica di regressione è il grafico che consente il confronto tra la risposta sperimentale e la risposta calcolata (e/o la risposta in predizione, Fig. 7-1).
Un perfetto allineamento dei punti sulla retta ottimale è un indice di un modello perfetto, cioè privo di errore.

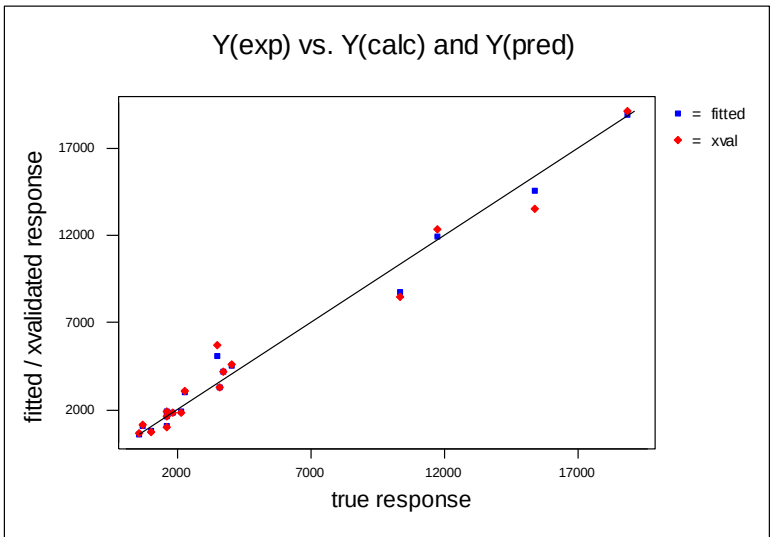


FIG. 7-1

I **residui ordinari** di un modello di regressione (gli errori) sono definiti come la differenza tra il valore della risposta sperimentale e quella calcolata dal modello:

$$r_i = y_i - \hat{y}_i$$

Un aspetto importante della diagnostica riguarda il controllo delle assunzioni sulla distribuzione deglli errori, che devono avere un valore atteso della media uguale a zero ed una distribuzione normale intorno al valor medio. Deviazioni da queste assunzioni sono indice di problemi di varia natura nel modello calcolato.
In generale, i residui vengono riportati rispetto alla risposta sperimentale o rispetto alla sequenza degli oggetti. I residui possono essere tutti quelli calcolati nelle diverse forme (dal modello calcolato o in predizione, standardizzati o no, etc.).
Il grafico **a** della Fig. 7-2 mostra un caso di andamento ideale. I grafici **b**, **c** e **d** della Fig. 7-2 mostrano tre tipiche situazioni patologiche:

- **b**: i residui crescono con l'aumentare della risposta, cioè la varianza non è distribuita omogeneamente (**eteroscedasticità**, contrapposta a *omoscedasticità*);

- **c**: la scelta del modello non è adeguata, cioè si rilevano degli andamenti non casuali (ad esempio, mancano nel modello dei termini non-lineari);

- **d**: l'andamento dei residui rivela un andamento costante non casuale, dovuto generalmente alla presenza di un errore sistematico.

I residui ordinari non hanno, in generale, la stessa varianza. La loro varianza è infatti definita dalla seguente espressione:

$$V(r) = \sigma^2 \cdot (\mathbf{I} - \mathbf{H})$$

e, in particolare, per ciascun singolo oggetto,

$$V(r_i) = \sigma^2 \cdot (1 - h_{ii})$$

dove **H** è la matrice d'influenza e h_{ii} i suoi elementi diagonali.
E' quindi evidente che un punto con alto *leverage* ha una piccola varianza rispetto a punti a basso *leverage*.

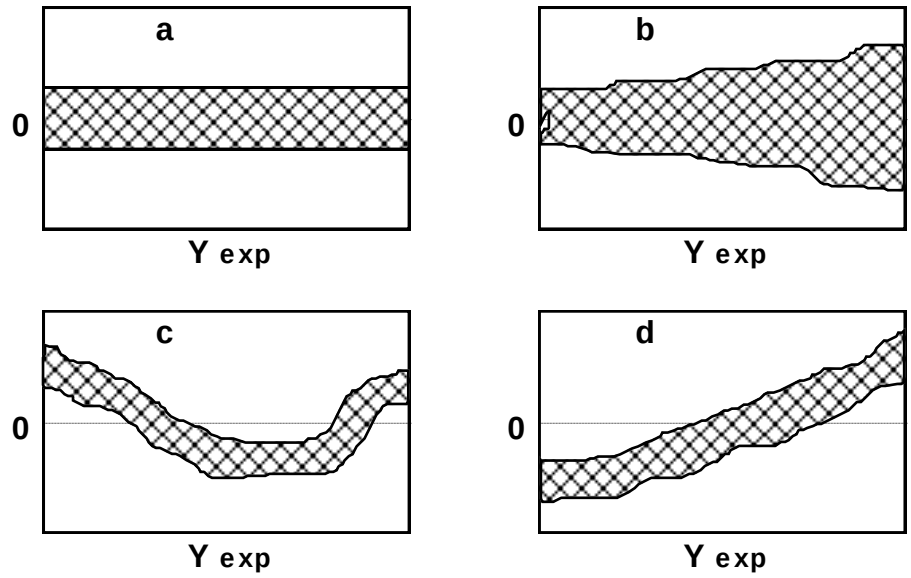


FIG. 7-2

Nota. I valori della diagonale principale della matrice dei *leverages* rappresentano l'influenza che ciascun oggetto ha nel definire i parametri del modello. Nel caso in cui l'*i*-esimo oggetto non sia presente nel modello, il suo *leverage* viene calcolato dall'espressione:

$$h_{i/i} = \mathbf{x}_i^{\mathrm{T}} (\mathbf{X}_{(i)}{}^{\mathrm{T}} \mathbf{X}_{(i)})^{-1} \mathbf{x}_i = \frac{h_{ii}}{1 - h_{ii}}$$

dove $\mathbf{x}_i$ rappresenta l'*i*-esimo campione non presente nel modello.

Per ottenere residui che abbiano una varianza comune è necessario dividere i residui ordinari per la loro deviazione standard:

$$r_i' = \frac{r_i}{s \cdot \sqrt{1 - h_{ii}}}$$

dove s è la radice quadrata di s^2 , la varianza stimata da RSS / (n - p').
Questi residui sono chiamati **residui standardizzati** e sono particolarmente indicati per il controllo della normalità dei residui (mediante *normal probability plots*) e per individuare eventuali residui anomali confrontadoli con i valori di 1, 2, 3 unità di deviazine standard σ.
Oltre ai residui ordinari e standardizzati definiti in precedenza, si possono calcolare altri tipi di residui di particolare interesse.
I **residui in predizione** sono calcolati come differenza tra il valore sperimentale della risposta e il valore predetto, cioè con un modello ottenuto in assenza del campione considerato:

$$r_{i/i} = y_i - y_{i/i}$$

dove il simbolo i/i indica un valore calcolato per il campione i-esimo da un modello ottenuto senza di esso.
La varianza dei residui in predizione è definita come:

$$V(r_{i/i}) = \sigma^2 \cdot \left[1 + \mathbf{x}_i^{\mathrm{T}} \left(\mathbf{X}_{(i)}^{\mathrm{T}} \mathbf{X}_{(i)} \right)^{-1} \mathbf{x}_i \right]$$

In questo caso, anche la stima della varianza dell'errore quando viene eliminato l'i-esimo punto può essere direttamente ricavata dall'espressione:

$$s_{(i)}^2 = \frac{(n - p')s^2 - \dfrac{r_i^2}{1 - h_{ii}}}{n - p' - 1}$$

dove il simbolo (i) indica il calcolo del parametro in assenza dell'i-esimo campione.
I residui standardizzati in predizione, detti **residui studentizzati** (*studentized residuals* o *jackknifed residuals*) si ottengono quindi dalla seguente espressione:

$$r'_{i/i} = \frac{r_i}{s_{(i)} \cdot \sqrt{1 - h_{ii}}} = \frac{s \cdot r'_i}{s_{(i)}} = \frac{r'_i}{\left(\dfrac{n - p' - r_i'^2}{n - p' - 1} \right)^{1/2}}$$

Come si può osservare, nel caso che il valore predetto della risposta sia ottenuto mediante il metodo *OLS* (ed anche *RIDGE* o *PCR*) con la procedura di validazione *leave-one-out*, è possibile calcolare direttamente i residui in predizione dai residui ordinari.
Una misura della **differenza tra valori calcolati e valori predetti** è definita dalla seguente espressione:

$$DFFIT_i = \hat{y}_i - \hat{y}_{i/i} = r_i \cdot \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

mentre la corrispondente misura normalizzata è data dall'espressione:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i/i}}{s_{(i)} \cdot \sqrt{h_{ii}}} = r'_{i/i} \cdot \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

Valori di controllo per la rilevazione di un punto la cui influenza è anomala sono $2 \cdot \sqrt{p'/n}$ e $3 \cdot \sqrt{p'/n}$.
Un altro parametro utile per valutare **le differenze tra i coefficienti di regressione** calcolati con tutti i dati e senza l'*i*-esimo dato è definito come:

$$DFBETA_i = \hat{\mathbf{b}} - \hat{\mathbf{b}}_{i/i} = \frac{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} x_i r_i}{1 - h_{ii}}$$

Il corrispondente parametro scalato, relativo a ciascuna variabile, è:

$$DFBETAS_{ij} = \frac{\hat{b}_j - \hat{b}_{j(i)}}{s_{(i)} \cdot \sqrt{d^{jj}}} = \frac{\left\{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i\right\}_j r_i}{s_{(i)} \cdot \sqrt{d^{jj}}} = \frac{\left\{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i\right\}_j r'_{i/i}}{\sqrt{d^{jj}} \cdot \sqrt{1 - h_{ii}}}$$

dove il termine $\left\{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i\right\}_j$ rappresenta il *j*-esimo elemento del vettore $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i$.

Valori di controllo di questo parametro sono $2/\sqrt{n}$ e $3/\sqrt{n}$.
Una tecnica grafica molto efficace nell'analisi dettagliata del modello di regressione è il **grafico di Williams**, che consiste nel proiettare i valori diagonali della matrice **H** contro i valori dei residui standardizzati (Fig. 7-3).

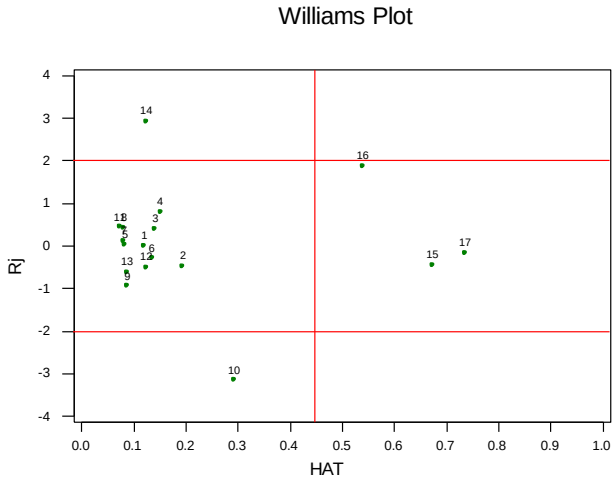


FIG. 7-3

Se i valori calcolati (o predetti) della risposta sperimentale per tutti gli oggetti sono accettabili, gli oggetti saranno concentrati intorno allo zero dei residui e comunque entro 2 o 3 unità di deviazione standard; inoltre se tutti gli oggetti hanno un'influenza simile nel determinare i parametri del modello (cioè hanno valori vicini al valor medio di *hat*, pari a *p'*/*n*, dove *p*' è il numero di parametri del modello), essi si troveranno nella parte più sinistra del grafico.
E' quindi facile individuare gli oggetti che sono ***outliers***, cioè gli oggetti che sono troppo in alto o troppo in basso nel grafico, al di fuori, ad esempio, di $\pm 2\sigma$ (le due linee orizzontali); nello stesso tempo, possiamo individuare gli oggetti che hanno una elevata influenza sul modello (***high leverage objects***), cioè gli oggetti più a destra nel grafico e, in particolare, oltre le linee di demarcazione pari a $2 \cdot \bar{h}$ (la linea verticale nel grafico) o $3 \cdot \bar{h}$.

🕮 Bibliografia

N. DRAPER E H. SMITH (1981). *Applied regression analysis*. Wiley, NY.

R.D. COOK (1982). *Residual and influence analysis*. Chapman & Hall, NY.

A.C. ATKINSON (1985). *Plots, transformations, and regression*. Oxford Univ. Press, Oxford.

8

I METODI DI REGRESSIONE: GLI SVILUPPI

8.1 - Oltre i minimi quadrati

Molti dati analizzati mediante metodi chemiometrici provengono dalla chimica analitica, dalla chimica organica, dalle ricerche su alimenti e farmaci, da analisi di sistemi ambientali. Spesso il numero di variabili supera largamente il numero di osservazioni; inoltre esiste in dati di questo tipo un alto grado di collinearità tra le variabili, come tipicamente accade per variabili che rappresentano segnali digitalizzati (ad esempio, spettri). La ricerca di modelli di regressione, così come per i modelli di classificazione, richiede strumenti evoluti, anche se spesso euristici, per trattare con finalità predittive questo tipo di problemi. Un certo numero di metodi di regressione, alcuni prodotti da statistici ed altri dagli stessi chemiometri, si è imposto all'attenzione per le potenzialità che essi hanno nella soluzione di problemi multivariati altamente complessi.

8.2 - Il metodo di regressione *Ridge* (RR)

Il *metodo di regressione Ridge*, proposto da Hoerl e Kennard nel 1962, si propone di superare i problemi dovuti a situazioni *malcondizionate* quando la correlazione tra le variabili del modello è sufficientemente elevata da rendere la matrice $\mathbf{X}^T\mathbf{X}$ quasi singolare (quindi invertibile con difficoltà).
La soluzione a questo problema viene cercata dal metodo Ridge con l'aggiunta di una piccola costante k a ciascun elemento diagonale della matrice di dispersione in modo tale da abbassare i valori più elevati della diagonale della sua matrice inversa. Questo consente quindi di ridurre l'incertezza nella stima dei parametri di regressione.

L'assunzione fondamentale su cui si basa il metodo **RR** nell'introduzione di un *bias* è che valori molto alti dei coefficienti di regressione sono probabilmente spuri.
L'aggiunta della costante *k* consente di forzare i coefficienti **b** verso zero.

$$\mathbf{b}_{RR}(k) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

dove *k* è la costante del metodo *RR* e **I** è la matrice identità.
Il parametro *k* è generalmente compreso tra 0 (il metodo coincide con *OLS*) e 1.
La regressione viene effettuata per diversi valori di *k* e viene selezionato il valore che rende massima la varianza spiegata in predizione.
Il grafico che riporta in ascisse i valori di *k* e in ordinate, in corrispondenza di ciascun *k*, i valori dei parametri del modello viene chiamato *ridge trace*.
L'esame di questo grafico consente di valutare il valore di *k* per il quale l'abbassamento complessivo dei valori dei diversi coefficienti di regressione verso valori piccoli si stabilizza.
La soluzione ricercata dal metodo *RR* minimizza la seguente espressione a due termini:

$$\min_k \ \sum_i e_i^2 + k\sum_j b_j^2$$

dove il primo termine è ancora il criterio di *OLS*, mentre il secondo termine penalizza le soluzioni in termini di coefficienti di regressione elevati.

Nota. La complessità di un modello è in relazione con il numero totale di variabili presenti nel modello; la somma dei quadrati dei coefficienti di regressione standardizzati è una differente misura della complessità del modello. In questo senso, il metodo *RR* minimizza contemporaneamente il quadrato dei residui e la complessità.

La varianza dei coefficienti calcolati col metodo RR è:

$$V\left(b_{RR}^{k}\right) = \sigma^2 \cdot \sum_j \frac{\lambda_j}{\left(\lambda_j + k\right)^2}$$

Come si può osservare, se il parametro *k* è uguale a zero, la varianza dei coefficienti coincide con quella calcolata per il metodo OLS.

Esempio

Per i dati dell'esempio REGTEST, nella Tab.8-1 sono riportati i diversi valori dei parametri di regressione calcolati per differenti valori del parametro *k* (compreso tra 0 e 0.2).

k	R^2	R^2_{cv}	*d.f.*	*RSS*	*PRESS*	
0.00	99.1	93.5	11.00	4535051	32195242	OLS
0.02	98.9	96.1	12.53	5248941	19193856	
0.04	98.8	96.8	12.86	6125079	15984377	
0.06	98.6	97.0	13.08	6865907	14807579	
0.08	98.5	97.1	13.24	7484340	14469406	RR
0.10	98.4	97.1	13.38	8017357	14555980	
0.12	98.3	97.0	13.49	8493606	14890134	
0.14	98.2	96.9	13.58	8932841	15383986	
0.16	98.1	96.8	13.66	9348507	15988917	
0.18	98.0	96.6	13.72	9749815	16675595	
0.20	97.9	96.5	13.79	10143190	17425126	

TAB. 8-1

In Fig. 8-1 viene riportato il grafico per la selezione del miglior modello predittivo (il modello di complessità ottimale. In funzione dei diversi valori di *k*, sono riportati i valori di R^2 in *fitting* e in predizione. Si può osservare che l'introduzione di un bias fa diminuire il valore in *fitting*.

Nella Tab.8-2 sono riportati i valori dei coefficienti di regressione calcolati per diversi valori del parametro *k* del metodo *RR*.
Il grafico di Fig. 8-2 viene chiamato *traccia di Ridge* (*Ridge trace*) e riporta i valori dei coefficienti di regressione dei diversi parametri in funzione di *k*. Come si può immediatamente osservare, il punto ottimale ($0.08 < k < 0.10$) è il punto in cui i valori dei coefficienti di regressione sono più piccoli, in accordo ai principi del metodo.

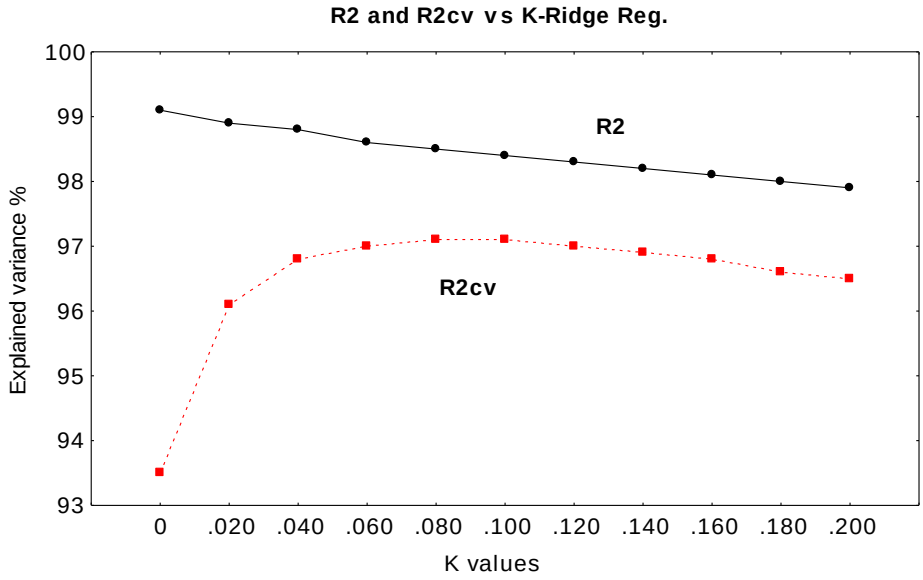


FIG. 8-1

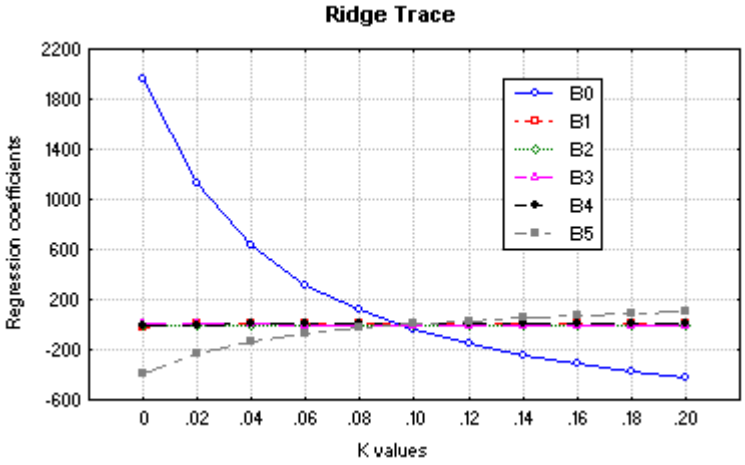


FIG. 8-2

k	b_0	b_1	b_2	b_3	b_4	b_5
0.00	1962.94	-15.85	0.06	1.59	-4.22	-394.31
0.02	1122.83	13.51	0.06	0.47	0.25	-236.20
0.04	624.89	12.10	0.06	0.41	3.29	-137.20
0.06	320.08	11.23	0.07	0.38	5.12	-74.75
0.08	113.17	10.64	0.07	0.36	6.32	-30.75
0.10	-36.91	10.20	0.07	0.34	7.16	2.50
0.12	-150.81	9.87	0.07	0.33	7.78	28.89
0.14	-240.04	9.60	0.06	0.32	8.24	50.56
0.16	-311.58	9.38	0.06	0.31	8.60	68.85
0.18	-369.91	9.19	0.06	0.31	8.88	84.59
0.20	-418.03	9.03	0.06	0.30	9.10	98.35

TAB. 8-2

8.3 - La selezione di un sottoinsieme ottimale di variabili

Una possibilità importante per ottenere modelli di regressione che, pur sacrificando qualche cosa alla capacità descrittiva dei dati, abbiano una migliore capacità predittiva, ci è data dalla ricerca di un sottoinsieme ottimale delle variabili originali (***VSS, Variable Subset Selection*** **)**.
La ricerca del migliore sottomodello o di un insieme di sottomodelli ottimali è giustificata da almeno 4 ragioni principali:

a. esprimere la relazione tra predittori e risposta nel modo più semplice possibile;
b. distinguere le variabili rilevanti da quelle non rilevanti;
c. aumentare la precisione delle stime statistiche e delle predizioni;
d. ridurre i costi di predizioni future;

Il primo e più ovvio metodo di *VSS* è quello che possiamo chiamare il **metodo di tutti i possibili modelli** (***All Subset Models*****)**. Questo metodo consiste nell'analizzare, di norma col metodo dei **minimi quadrati ordinari**, tutti i possibili modelli ottenuti da tutte le possibili combinazioni delle variabili considerate nel modello globale. Il numero totale di modelli è 2^p - 1.
Ad esempio, se il modello globale proposto è:

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + b_3\mathbf{x}_3$$

vengono provati, uno ad uno, tutti i modelli con le variabili indipendenti di Tab. 8-3.

numero di variabili	*variabili del modello*
1	x_1
1	x_2
1	x_3
2	x_1 , x_2
2	x_1 , x_3
2	x_2 , x_3
3	x_1 , x_2 , x_3

TAB. 8-3

Il modello finale viene selezionato valutando il valore massimo di R^2_{adj} o il valore massimo di R^2_{cv}.
Questo metodo, concettualmente molto semplice, pone in realtà gravi problemi di computabilità. Infatti, non appena il numero totale di variabili da prendere in considerazione supera alcune unità, il numero complessivo dei modelli da calcolare diviene straordinariamente grande.
Se *p* è il numero totale di variabili disponibili, il numero totale *t* di modelli ottenibili, di dimensione da 1 a *k* variabili, è dato dalla seguente formula:

$$t = \sum_k \left(\frac{p!}{k!(p-k)!} \right) \leq 2^p - 1$$

Ad esempio, se il numero di variabili disponibili è 32 ($p = 32$), il numero totale *t* di modelli con una variabile, con 2 variabili, fino a 8 variabili ($1 \leq k \leq 8$) è 1.5×10^7.
Utilizzando un computer in grado di calcolare in media un modello al secondo, occorrerebbero circa 174 giorni di calcolo!

Un metodo semplice e sufficientemente affidabile per la selezione delle variabili e che consente di effettuare un numero di calcoli molto inferiore è il **metodo di ricerca sequenziale**. Il metodo si basa sul seguente algoritmo:

1. si seleziona l'intervallo delle dimensioni del modello che interessano, cioè il numero minimo e massimo di variabili da utilizzare;

2. per ogni dimensione si estraggono alcuni modelli a caso, per i quali viene calcolata la risposta;

3. per ciascun modello di ciascuna dimensione k si effettua una ricerca sequenziale secondo il seguente schema: si provano i modelli costituiti scambiando la variabile estratta al primo posto con tutte le altre $p - 1$ variabili, escludendo quelle variabili che sono già presenti nelle posizioni successive alla prima, variabili che vengono tenute fisse ($p - k$). Se viene ottenuto un modello migliore questo viene memorizzato. Successivamente viene reinserita la variabile che originariamente era nella prima posizione e la stessa procedura viene effettuata sostituendo la variabile posta nella seconda posizione. La procedura viene effettuata per tutte le variabili del modello originale. Questa fase, per un modello di dimensione k e p variabili candidate, richiede il calcolo di k x ($p - k + 1$) modelli. Se in questa fase non viene trovato alcun modello migliore di quello iniziale, la procedura termina e viene esaminato un altro dei modelli iniziali estratti a caso. Diversamente, il modello iniziale viene sostituito dal migliore modello trovato e la procedura viene ripetuta nuovamente.

Ad esempio, se uno dei modelli estratti a caso è costituito dalle 4 variabili T, I, R, O, estratte tra le 26 variabili A-Z, vengono provati tutti i modelli in cui la variabile T viene sostituita da tutte le altre, escluse I, R e O. Supponiamo che il miglior modello trovato sia MIRO. Successivanmente, reinserita la variabile T, vengono provati tutti i modelli in cui viene sostituita la lettera I con tutte le altre, escluse T, R e O. In questo secondo caso non viene trovato alcun modello migliore rispetto a TIRO. Negli altri due restanti casi i migliori modelli trovati sono: TIAO e TIRA. Supponiamo che tra i 3 nuovi modelli trovati il migliore sia il modello TIAO. La procedura viene interamente ripetuta partendo da questo modello. Si può avere quindi la sequenza di Tabella 8-4:

modello iniziale	*modello trovato*
TIRO	TIAO
TIAO	GIAO
GIAO	GIAK
GIAK	GRAK
GRAK	GRAL
GRAL	GOAL
GOAL	nessuno

TAB. 8-4

Il miglior modello trovato è quindi costituito dalle variabili G, O, A e L.
Il numero totale di calcoli effettuati su un singolo modello tra quelli estratti a caso è 7 x 4 x (26 - 3) = 644 , mentre tutti i possibili modelli di dimensione 4 ammontano a 14950. Anche estraendo, per ciascuna dimensione, 4 o 5 modelli iniziali, il numero complessivo di calcoli risulta particolarmente conveniente. Questo metodo può essere considerato una buona alternativa al metodo basato sugli algoritmi genetici per la selezione delle variabili (v. capitolo 10), anche se, come in quest'ultimo, non vi è mai la certezza di trovare il modello migliore tra tutti quelli possibili.

Esempio

Nella Tab.8-5 sono riportati i migliori 10 modelli sui 31 totali ottenuti col metodo ***All Subset Models***. In carattere corsivo sono riportati i migliori modelli per ogni dimensione considerata.

ID	*n.var.*	R^2	R^2_{adj}	R^2_{cv}	*variabili*
1	2	98.48	98.26	*97.45*	3 5
2	2	98.40	98.17	97.38	1 5
3	3	98.50	98.16	*97.36*	1 3 5
4	3	98.50	98.15	96.71	3 4 5
5	3	98.47	98.11	96.40	1 4 5
6	3	99.01	98.78	96.39	2 3 5
7	2	98.67	98.48	96.39	2 3
8	2	98.61	98.41	96.35	1 2
9	3	98.94	98.78	96.35	1 2 5
10	4	98.51	98.01	*96.24*	1 3 4 5

TAB. 8-5

8.4 - I metodi di regressione S*tepWise*

Tra i metodi più noti, per superare gli sforzi computazionali posti dal metodo che prevede il calcolo di tutti i possibili modelli, sono stati proposti diversi metodi *VSS* che consentono di valutare solo un piccolo numero di possibili modelli. Questi metodi sono noti col nome di ***StepWise Regression*** (***SWR***) e sono basati su due strategie fondamentali che prendono il nome di ***Forward Selection*** (***FS***) e ***Backward Elimination*** (***BE***).
Il metodo **FS** parte con un modello costituito da nessuna variabile ed aggiunge via via una variabile alla volta fino a che tutte le variabili vengono inserite nel modello oppure fino a che non viene soddisfatto un determinato criterio di arresto della procedura.
La variabile che viene inclusa nel modello ad ogni passo è quella che fornisce il maggiore valore del rapporto *F* (*F* di Fisher) per singolo grado di libertà tra tutte le variabili non ancora incluse nel modello.
Cioè, la *j*-esima variabile viene aggiunta al modello costituito da *p* variabili già precedentemente incluse se

$$F_j^+ = max_j \left[\frac{RSS_p - RSS_{p+j}}{s^2_{p+j}} \right] > F_{in}$$

s^2_{p+j} è la varianza calcolata dal modello con *p* variabili con l'aggiunta della *j*-esima variabile.
Il valore di F_{in} viene selezionato in anticipo al fine di stabilire un criterio di arresto ed equivale al valore *F* di Fisher alla probabilità ♋, per 1 grado di libertà al numeratore e (*n* - *p* - 1) al denominatore. Normalmente si utilizza $F_{in} = 4.$
La quantità RSS_p rappresenta la somma dei quadrati dei residui del modello con *p* variabili, la quantità RSS_{p+j} rappresenta la somma dei quadrati dei residui del modello con *p* variabili più la *j*-esima variabile.
Il metodo **BE** parte con un modello contenente tutte le variabili ed elimina una variabile alla volta. Ad ogni passo, viene eliminata la *j*-esima variabile con il più piccolo valore del rapporto *F* se questo non è superiore ad un valore specificato F_{out} :

$$F_j^- = min_j \left[\frac{RSS_{p-j} - RSS_p}{s_p^2} \right] < F_{out}$$

s^2_p è la varianza calcolata dal modello con *p* variabili.
Normalmente si utilizza $F_{out} = 2.$
La quantità RSS_p rappresenta la somma dei quadrati dei residui del modello con *p* variabili, la quantità RSS_{p-j} rappresenta la somma dei quadrati dei residui del modello costituito dalle *p* variabili senza la *j*-esima variabile.
In pratica, nel metodo **FS** si aggiunge la variabile che più riduce la somma dei quadrati degli scarti (*RSS*), mentre nel metodo **BE** si elimina la variabile la cui esclusione comporta il minimo innalzamento di *RSS*.
Nel metodo **FS**, se il valore di F_{in} è scelto sufficientemente grande, non tutte le variabili verranno incluse nel modello finale; viceversa, nel metodo **BE**, se il valore di F_{out} viene scelto sufficientemente piccolo, non tutte le variabili saranno eliminate.
Una variante dei due metodi è il metodo di regressione proposto da Efroymson (1960) e denotato con ***Elimination-Selection*** (***ES***). L'idea base del metodo è quella del metodo **FS**, ma ad ogni passo viene anche presa in considerazione la possibilità di eliminare una delle variabili già inserite nel modello, seguendo la tecnica del metodo **BE**. Ovviamente questa seconda fase si effettua dal momento in cui nel modello compaiono almeno due variabili.
Nonostante il metodo di regressione *stepwise* sia ancora largamente utilizzato, grazie anche all'ampia disponibilità di software, sono stati ormai dimostrati i

notevoli limiti di questo metodo che, soprattutto in presenza di molte variabili, è di norma incapace di trovare i migliori modelli, fermandosi su massimi relativi.

Esempio

Nella Tab.8-6 sono riportati i migliori 10 modelli ottenuti col metodo ***StepWise Regression***. In carattere corsivo sono riportati i migliori modelli per ogni dimensione considerata.

ID	*n.var.*	R^2	R^2_{adj}	R^2_{cv}	*variabili*
1	2	98.48	98.26	*97.45*	3 5
2	3	98.50	98.16	*97.36*	1 3 5
3	3	98.50	98.15	96.71	3 4 5
4	3	99.01	98.78	96.39	2 3 5
5	2	98.67	98.48	96.39	2 3
6	1	97.22	97.03	*95.59*	3
7	1	95.15	96.96	95.50	1
8	2	95.44	96.86	95.44	1 3
9	2	93.44	97.18	93.44	3 4
10	1	85.84	87.66	85.84	4

TAB. 8-6

Si osservi che il miglior modello di dimensione 3 (ID = 2, $R^2_{cv} = 97.36$) è peggiore del miglior modello di dimensione 2 (ID = 1, $R^2_{cv} = 97.45$): questo fatto sta ad indicare che la variabile 1 - nel suo effetto combinato alle altre due variabili 3 e 5 - non apporta informazione utile e questo modello non dovrà essere preso in considerazione. Si può anche notare che il miglior modello, valutato tenendo conto del parametro R^2 *adjusted*, ha dimensione 3 (ID = 4, $R^2_{adj} = 98.78$).

8.5 - La regressione in Componenti Principali (PCR)

Il metodo di **regressione in componenti principali** (***PCR***, ***Principal Component Regression***) è un metodo di regressione basato sui principi generali di OLS in cui la matrice originale del modello (**X**) viene sostituita da una sua rappresentazione nello spazio delle prime *M* componenti principali ($\mathbf{T_M}$). Questo metodo si basa sull'idea di rappresentare lo spazio dei predittori (le variabili indipendenti) in un numero ridotto di componenti principali, eliminando l'informazione legata alle direzioni di minima varianza assunte come non importanti. L'impiego in regressione di componenti principali come variabili indipendenti ha il notevole vantaggio di bilanciare l'importanza delle variabili indipendenti originali, in quanto il contributo di molte variabili "simili" viene unificato in un'unica componente.
Uno dei vantaggi di questo approccio alla regressione è quello di permettere di trattare problemi in cui il numero di oggetti è inferiore a quello delle variabili, utilizzando come nuove variabili predittrici un piccolo numero di componenti principali, inferiori al numero di osservazioni. Inoltre PCR viene utilizzata per superare i problemi di multicollinearità che nel metodo OLS si riflettono nelle difficoltà di invertire la matrice $\mathbf{X}^T\mathbf{X}$.
Dall'analisi in componenti principali sappiamo che i campioni vengono definiti nello spazio delle componenti dall'espressione:

$$\mathbf{T_M} = \mathbf{X} \cdot \mathbf{L_M}$$

dove *M* è il numero di componenti ritenute nel modello e $\mathbf{L_M}$ è la matrice dei *loadings* di dimensione (*p* x *M*), con *p* il numero di variabili indipendenti totali nel modello. Il modello di regressione PCR è quindi definito come:

$$\mathbf{y} = \mathbf{T_M}\alpha + \varepsilon = \mathbf{X}\mathbf{L_M}\mathbf{L_M^T}\beta + \varepsilon$$

dove α è il vettore dei coefficienti di regressione veri nello spazio delle componenti (di dimensione *M*) ed ε il termine di errore. I coefficienti di regressione veri nello spazio delle variabili originali si ottengono direttamente dall'espressione:

$$\beta = \mathbf{L_M} \cdot \alpha \quad \text{ovvero} \quad \alpha = \mathbf{L_M^T}\beta$$

La soluzione per i coefficienti **a** (stima dei coefficienti veri α) nello spazio delle componenti e la corrispondente soluzione nello spazio delle variabili originali sono:

$$\mathbf{a}_{PCR} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \qquad \mathbf{b}_{PCR} = \mathbf{L}_{\mathbf{M}}\mathbf{a}_{PCR}$$

Ricordando che per l'inversa della matrice di dispersione vale la relazione generale:

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \mathbf{L}\Lambda^{-1}\mathbf{L}^{\mathbf{T}}$$

la soluzione per i coefficienti di regressione nello spazio delle variabili originali è:

$$\mathbf{b_{PCR}} = \mathbf{L_M}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} = \mathbf{L_M}\left(\mathbf{L_M^{\mathrm{T}}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{L_M}\right)^{-1}\mathbf{L_M^T}\mathbf{X^T}\mathbf{y} = \mathbf{L_M}\Lambda_{\mathbf{M}}^{-\mathbf{1}}\mathbf{L_M^T}\mathbf{X^T}\mathbf{y}$$

La varianza ed il *bias* dei coefficienti di regressione sono definiti come:

$$V\left(\mathbf{b_{PCR}}\right) = \sigma^2 tr\left(\mathbf{L_M}\Lambda_{\mathbf{M}}^{-1}\mathbf{L_M^T}\right) = \sigma^2\sum\nolimits_m\left(1/\lambda_m\right)$$

$$B^2\left(\mathbf{b_{PCR}}\right) = \left(\mathbf{L_L}\mathbf{L_L^T}\beta\right)^2$$

dove $\mathbf{L_L}$ sono gli autovettori del sottospazio a bassa varianza, non considerato nel modello.
Nell'utilizzare questa tecnica di regressione si deve porre molta attenzione nell'eliminazione delle componenti di minima varianza. Infatti, se è pur vero che in molte circostanze queste componenti sono legate alla varianza dovuta al rumore sperimentale e ad aspetti marginali o casuali, non vi è alcuna ragione a priori perchè le componenti correlate con la risposta siano necessariamente le prime, cioè quelle di massima varianza. Si è dimostrato infatti che in molti casi, una volta calcolate *tutte* le componenti principali (con autovalori diversi da zero), è più opportuno adottare metodi che selezionino le componenti più significative per riprodurre la riposta, senza quindi precostituire un modello definito da una sequenza ordinata delle prime *M* componenti (v. oltre, par.8.7).

Esempio

Nella Tab.8-7 sono riportati i risultati ottenuti col metodo *PCR*, con modelli da 1 a 5 componenti principali sui dati REGTEST.

Modello	R^2	R^2_{cv}	*SDEC*	*SDEP*	*F*
1 PC	96.40	95.97	1023.73	1197.86	289.2
2 PC	97.83	*97.16*	794.73	908.87	239.6
3 PC	97.87	90.08	787.30	1699.40	39.3
4 PC	99.07	94.97	519.06	1313.85	47.58
5 PC	99.08	93.49	516.50	1376.17	31.61
OLS	99.08	93.49	516.50	1376.17	31.61
RR	98.49	97.08	663.52	922.57	159.63

TAB. 8-7

N. PC	b_0	b_1	b_2	b_3	b_4	b_5
1	-2492.13	8.03	0.06	0.26	11.38	561.44
2	-731.41	8.04	0.07	0.26	13.72	104.04
3	-834.76	7.68	0.08	0.25	11.97	129.75
4	2098.06	16.47	0.06	0.57	-5.79	-425.28
5	1962.94	-15.85	0.06	1.59	-4.22	-394.31
OLS	1962.94	-15.85	0.06	1.59	-4.22	-394.31
RR	113.17	10.64	0.06	0.36	6.32	-30.75

TAB. 8-8

Nella Tab.8-8 sono riportati i coefficienti di regressione calcolati con modelli in componenti principali: da un modello con una sola componente fino ad un modello che utilizza tutte le 5 possibili componenti e, quindi, coincidente col

modello *OLS*. Insieme a questi modelli, sono riportati i valori dei coefficienti ottenuti col metodo *RR*.

8.6 - Il metodo Partial Least Squares regression (PLS)

Il metodo di **regressione parziale con minimi quadrati** (***PLS***, ***Partial Least Squares regression***) è un metodo di regressione *biased* che consente di ottenere la massima stabilità del modello calcolato dei predittori. Il metodo è utilizzabile sia quando il rapporto oggetti/variabili è inferiore a uno, sia in presenza di variabili molto correlate tra loro, sia quando sono contemporaneamente presenti anche più risposte.
Quest'ultima possibilità è particolarmente interessante quando le risposte Y sono tra loro correlate, in quanto è possibile ottenere un unico modello da interpretare e non tanti modelli separati quante sono le risposte. Tuttavia se le risposte misurano quantità effettivamente indipendenti tra loro non vi è alcun vantaggio a sviluppare un unico modello di regressione. In questo caso, infatti, PLS tende a dare un numero di componenti elevato, almeno pari al numero di risposte analizzate.
Dal punto di vista matematico, il metodo PLS utilizza - iterativamente - coppie di componenti principali su **X** e su **Y** (quando **Y** contiene più di una variabile risposta) e ricerca tra queste la massima correlazione possibile.
In particolare, si possono mettere in evidenza le seguenti caratteristiche di PLS:

a) Le componenti PLS sono selezionate per produrre la massima riduzione della matrice di covarianza $\mathbf{X}^T\mathbf{Y}$ (dimensione *p, r*), cioè PLS fornisce il minimo numero necessario di variabili.

b) L'algoritmo di PLS fornisce una sola componente alla volta (per ogni iterazione). L'algoritmo PLS può essere visto come una procedura *step-wise*, dove viene selezionata la coppia di componenti in **X** e in **Y** che sono tra loro il più vicino possibile e sono normalmente designate come **t** e **u**.

c) La procedura PLS continua fino a che esistono componenti significative (in grado di migliorare il poter predittivo del modello).

d) Il processo iterativo si arresta quando non esistono più coppie di componenti in **X** e in **Y** sufficientemente vicine tra loro (correlate) o quando non esiste più informazione utile estraibile da **X** che può essere utilizzata per predire **Y**.

e) Le componenti **t** e **u** di ciascuna coppia sono le componenti in **X** e in **Y** che hanno la massima covarianza tra tutte le componenti in **X** e in **Y.**

f) L'algoritmo PLS seleziona una coppia di componenti alla volta perchè la covarianza della successiva coppia di componenti è minore della covarianza massima nella iterazione successiva.

g) Il numero *M* di componenti ottimali di PLS viene determinato mediante validazione, massimizzando $R^2_{cv}(M)$.

🗀 L'algoritmo PLS

Data la matrice del modello **X** (*n* , *p*) e la matrice delle risposte **Y** (*n* , *r*) (normalmente la risposta y è una sola, cioè *r* = 1), l'algoritmo di calcolo del metodo *PLS* per ciascuna componente *m* è il seguente:

1. inizializzazione:	$\mathbf{u}_m = \mathbf{y}_1$
2. pesi di **X** :	$\mathbf{w}_m^{\mathrm{T}} = \mathbf{X}\mathbf{u}_m$ scala a lunghezza 1
3. variabile latente di **X** :	$\mathbf{t}_m = \mathbf{X}\mathbf{w}_m^{\mathrm{T}}$
4. pesi di **Y** :	$\mathbf{q}_m^{\mathrm{T}} = \mathbf{Y}^{\mathrm{T}}\mathbf{t}_m$ scala a lunghezza 1
5. variabile latente di **Y** :	$\mathbf{u}'_m = \mathbf{Y}\mathbf{q}_m^{\mathrm{T}}$
6. verifica convergenza:	se $(\delta\mathbf{u}_m < \varepsilon)$ prosegui / $\mathbf{u}_m = \mathbf{u}'_m$ e vai a 2
7. relazione interna (regr. u su t):	$b = \mathbf{u}_m^{\mathrm{T}}\mathbf{t}_m / (\mathbf{t}_m^{\mathrm{T}}\mathbf{t}_m)$ da $\mathbf{u} = b\mathbf{t}$
8. loadings di **X** :	$\mathbf{l}_m^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{t}_m / (\mathbf{t}_m^{\mathrm{T}}\mathbf{t}_m)$
9. residui di **X** :	$\mathbf{X}' = \mathbf{X} - \mathbf{t}_m\mathbf{l}_m$
10. residui di **Y** :	$\mathbf{Y}' = \mathbf{Y} - b\mathbf{t}_m\mathbf{q}_m$

I vettori **w** e **q** sono detti *pesi* rispettivamente di **X** e di **Y**. Essi sono i coefficienti delle due combinazioni lineari della matrice **X** (da cui si genera la componente **t**) e della matrice **Y** (da cui si genera la componente **u**).

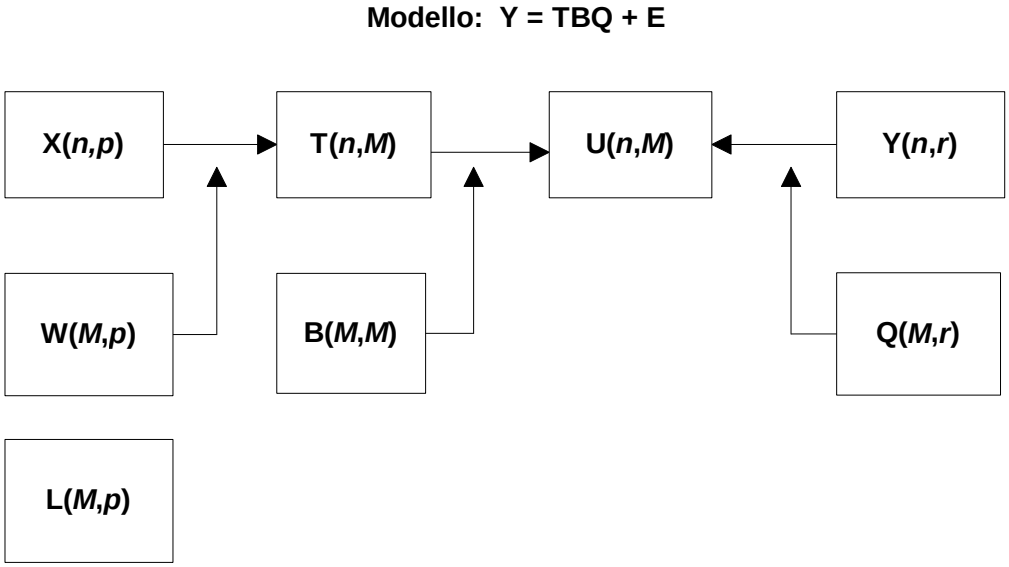


FIG. 8-3

Nota. Il vettore **q** viene scalato a lunghezza uno perchè, in caso contrario, si avrebbe:

$$\mathbf{u}^T\mathbf{t} = \mathbf{t}^T\mathbf{t}$$

cioè $b_m = 1$ per tutte le iterazioni.

Esempio

Nella Tab. 8-9 sono riportati i risultati ottenuti con i 5 modelli *PLS*.
Nella Tab. 8-10 sono riportati i coefficienti di regressione calcolati con modelli *PLS*: da un modello con una sola componente fino ad un modello che utilizza tutte le 5 possibili componenti e, quindi, coincidente col modello *OLS*. Insieme

a questi modelli, sono riportati i valori dei coefficienti ottenuti con i metod1 *RR* e *PCR*.

Modello	R^2	R^2_{cv}
1 PC	96.84	94.94
2 PC	98.01	96.26
3 PC	99.03	92.80
4 PC	99.07	94.07
5 PC	99.08	93.49
OLS	99.08	93.49
RR	98.49	97.08
PCR (2)	97.83	97.16

TAB. 8-9

N. PC	b_0	b_1	b_2	b_3	b_4	b_5
1	-2580.40	8.13	0.06	0.27	11.56	484.92
2	-399.26	8.63	0.07	0.29	12.24	39.93
3	1806.08	15.50	0.07	0.54	-6.35	-364.02
4	2097.86	16.40	0.06	0.57	-5.79	-425.23
5	1962.94	-15.85	0.06	1.59	-4.22	-394.31
OLS	1962.94	-15.85	0.06	1.59	-4.22	-394.31
RR	113.17	10.64	0.06	0.36	6.32	-30.75
PCR (2)	-731.41	8.04	0.07	0.26	13.72	104.04

TAB. 8-10

Qui di seguito viene riportato un esempio di *output* ottenuto applicando il metodo PLS ai dati REGTEST.

file : REGTEST inform. : I00 data-set : D00 17 unscaled data

Partial Least Squares (PLS) 05-13-1994 - 09:47

variables active : 1 2 3 4 5 6

---

c-v responses with fixed number of components: leave-one-out procedure

component 1

regression coefficient = 0.481

response : Y weight = 1.0000

ID	variable	coefficient	weight	modeling power
0	intercept	-2580.3967		
1	X1	8.1260	0.4894	98.750
2	X2	0.0590	0.0859	86.330
3	X3	0.2668	0.3988	98.610
4	X4	11.5648	0.2645	89.380
5	X5	484.9216	-0.7240	41.140

---

component 2

regression coefficient = 0.137

response : Y weight = 1.0000

ID	variable	coefficient	weight	modeling power
0	intercept	-399.2572		
1	X1	8.6296	0.4693	98.680
2	X2	0.0707	0.3077	93.060
3	X3	0.2851	-0.0808	98.530
4	X4	12.2433	-0.8237	94.260
5	X5	39.9338	0.0084	99.640

---

component 3

regression coefficient = 0.467

response : Y weight = 1.0000

ID	variable	coefficient	weight	modeling power
0	intercept	1806.0747		
1	X1	15.4994	0.4896	99.190
2	X2	0.0711	0.0974	92.650
3	X3	0.5365	0.4449	99.200
4	X4	-6.3507	0.2787	98.490
5	X5	-364.0150	0.6894	99.620

---

component 4

regression coefficient = 0.069

response : Y weight = 1.0000

ID	variable	coefficient	weight	modeling power
0	intercept	2097.8550		
1	X1	16.4041	0.4669	100.000
2	X2	0.0563	0.0766	100.000
3	X3	0.5684	-0.7978	100.000
4	X4	-5.7863	0.3731	100.000
5	X5	-425.2339	0.0216	100.000

---

component 5

regression coefficient = 1.298

response : Y weight = 1.0000

ID	variable	coefficient	weight	modeling power
0	intercept	1962.9449		
1	X1	-15.8525	0.2873	100.000
2	X2	0.0559	-0.9395	100.000
3	X3	1.5897	-0.0089	100.000
4	X4	-4.2186	-0.1863	100.000
5	X5	-394.3132	0.0097	100.000

---

response : Y

comp.	Rsq. %	SDEC	Rsq.cv %	SDEP
1	96.844	958.3992	94.943	1213.0996
2	98.011	760.7708	96.256	1043.8051
3	99.033	530.5300	92.802	1447.2574
4	99.074	519.0522	94.066	1314.1095
5	99.083	516.4955	93.492	1376.1666

8.7 - Confronto tra i metodi di regressione

Alla luce delle ampie possibilità di utilizzare metodi di regressione diversi, è interessante confrontare tra loro i diversi modelli ottenuti. Il confronto viene qui proposto utilizzando l'analisi delle componenti principali su una matrice di dati costituita dai diversi metodi di regressione (le righe della matrice) e dai coefficienti di regressione standardizzati (*p* colonne della matrice, se *p* sono le variabili considerate).
Le componenti principali più significative descrivono i coefficienti di regressione che variano di più tra i diversi modelli o, in altre parole, alla prima componente sono correlati i coefficienti che presentano la massima varianza, alla seconda i coeffcienti che presentano la seconda massima varianza, e così di seguito.
I grafici degli scores rappresentano i diversi metodi: punti che stanno tra loro vicini rappresentano quindi metodi che forniscono modelli simili per quanto riguarda i coefficienti di regressione correlati alle componenti principali analizzate e indipendentemente dalle capacità descrittive o predittive dei modelli.
Un esempio di confronto tra metodi di regressione viene discusso qui di seguito. Su un insieme di dati costituito da un certo numero di campioni rappresentati da 17 variabili ed una risposta sono stati calcolati i modelli di regressione riportati in Tab. 8-11. Non sono stati riportati i valori dei coefficienti di regressione ottenuti per ciascun modello.

ID	*Metodo*	R^2_{cv}	*ID*	*Metodo*	R^2_{cv}	*ID*	*Metodo*	R^2_{cv}
1	OLS	11.4	17	RR - 0.02	45.0	38	V-GA - 4A	58.4
2	PCR - 1	32.3	18	RR - 0.04	49.5	39	V-GA - 4B	57.0
3	PCR - 2	36.5	19	RR - 0.06	50.9	40	V-GA - 4C	56.0
4	PCR - 3	35.3	20	RR - 0.08	51.2	41	V-GA - 4D	54.8
5	PCR - 4	41.3	21	RR - 0.10	51.1	42	V-GA - 4E	54.6
6	PCR - 5	44.9	22	RR - 0.12	50.8	43	V-GA - 5A	58.5
7	PCR - 6	36.7	23	RR - 0.14	50.4	44	V-GA - 5B	58.4
8	PCR - 7	36.8	24	RR - 0.16	49.9	45	P-GA - 2A	40.5
9	PCR - 8	31.2	25	RR - 0.18	49.3	46	P-GA - 2B	39.6
10	PCR - 9	39.8	26	RR - 0.20	48.7	47	P-GA - 3A	44.5
11	PCR - 10	46.5	27	V-GA - 1A	43.5	48	P-GA - 3B	43.1
12	PLS - 1	45.5	28	V-GA - 2A	51.5	49	P-GA - 4A	48.0
13	PLS - 2	44.0	29	V-GA - 2B	48.3	50	P-GA - 4B	48.0
14	PLS - 3	40.4	30	V-GA - 2C	45.6	51	P-GA - 4C	46.4
15	PLS - 4	39.8	31	V-GA - 2D	45.2	52	P-GA - 5A	52.8
16	PLS - 5	37.9	32	V-GA - 2E	44.5	53	P-GA - 5B	50.2
			33	V-GA - 3A	55.0	54	P-GA - 6A	53.9
			34	V-GA - 3B	54.5	55	P-GA - 6B	53.4
			35	V-GA - 3C	52.7	56	P-GA - 6C	53.2
			36	V-GA - 3D	52.6	57	P-GA - 6D	53.1
			37	V-GA - 3E	52.4			

TAB. 8-11

Con il metodo PCR sono stati calcolati i modelli da 1 fino a 10 PC. Il potere predittivo maggiore è stato ottenuto per il modello PCR-11 (10 componenti) e un massimo relativo è stato ottenuto per il modello PCR-6 (5 componenti). Il metodo PLS fornisce il miglior modello con una sola variabile latente (modello PLS-12); sono stati calcolati anche i modelli PLS fino a 5 variabili latenti. Con il metodo RR sono stati calcolati 10 modelli ($0.02 \le k \le 0.20$): il modello migliore è il modello RR-20 ($k = 0.08$).
Diversi modelli sono inoltre stati ottenuti utilizzando metodi per la selezione delle variabili (v. capitolo 10 sugli algoritmi genetici). Questo metodo è stato utilizzato sia per selezionare i migliori sottoinsiemi di variabili originali (V-GA) sia per selezionare i migliori sottoinsiemi di componenti principali (P-GA). Quest'ultimo metodo differisce da PCR in quanto in PCR le componenti principali selezionate sono sempre in sequenza decrescente di varianza spiegata (cioè, in un modello PCR a 4 componenti vengono selezionate le componenti

dalla prima alla quarta); in P-GA le componenti sono liberamente selezionate tra tutte le componenti disponibili (cioè, in un modello a 4 componenti possono essere selezionate le componenti 2, 3, 5 e 7, se questo insieme di componenti fornisce il miglior risultato predittivo). Le lettere A, B, ..., E, che accompagnano i modelli V-GA e P-GA distinguono i diversi migliori modelli ottenuti con questi metodi; i numeri che precedono le lettere sono il numero di variabili del modello.

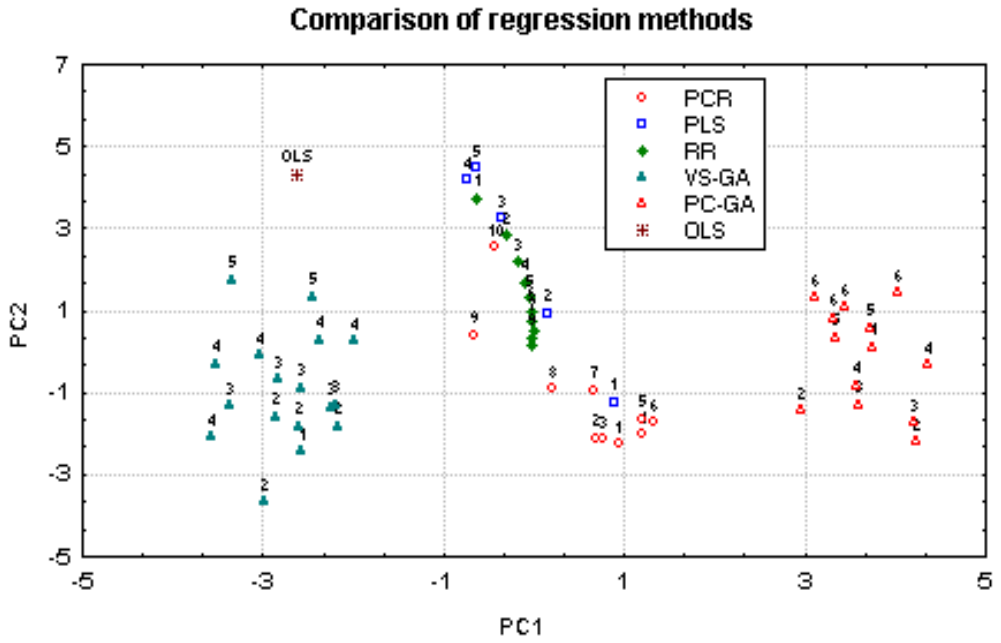


FIG. 8-4

Dall'analisi della Fig. 8-4 si può subito osservare che i modelli ottenuti con i *metodi asintoticamente unbiased* (PCR, PLS e RR) convergono verso OLS, rispettivamente, quando il numero di componenti considerate tende ad essere grande (PCR, PLS) o il valore della costante *k* tende a zero (RR). Ad esempio, si può osservare che PLS e PCR con 1 sola componente si discostano considerevolmente da OLS, diversamente da quanto fanno i rispettivi metodi con 5 (PLS) e 10 (PCR) componenti. Analogamente per il metodo RR, ove il

modello RR-17 (k = 0.02) è molto più vicino a OLS di quanto lo sia RR-25 (k = 0.20, l'ultimo verso il basso nella sequenza di RR).
I metodi di selezione delle variabili e di selezione delle componenti si discostano considerevolmente da tutti gli altri in quanto forniscono modelli diversi dai precedenti: in questi modelli tutte le variabili non presenti nei modelli hanno coefficienti di regressione uguali a zero.
In generale, ci si può aspettare che in presenza di grande correlazione vi siano notevoli distanze tra OLS e i migliori modelli ottenuti con i diversi metodi *biased*, indicando l'instabilità del modello OLS. Le diverse direzioni che uniscono OLS con i migliori modelli forniscono anche informazioni sulle diverse modalità con cui il *bias* interviene nello stabilizzare *predittivamente* i modelli.
Nella Fig. 8-5 sono riportati, per i diversi metodi, solo i migliori modelli (massima capacità predittiva) tra tutti quelli considerati, con il corrispondente valore di R^2_{cv}.

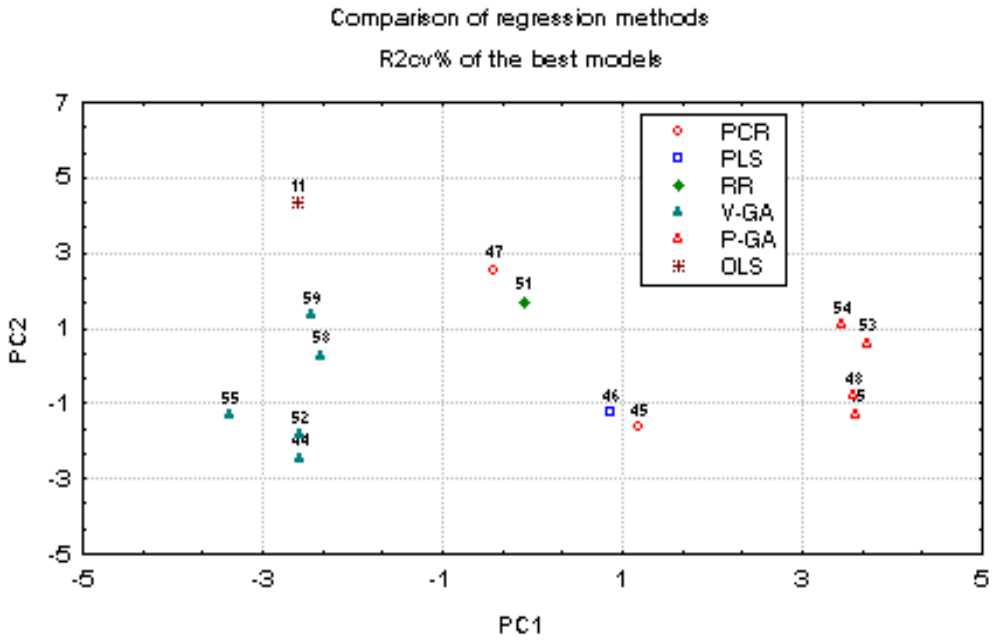


FIG. 8-5

Come è facile osservare, i modelli V-GA individuano una zona ottimale di massima predittività; i modelli P-GA, a loro volta individuano una diversa zona ottimale, se pur di qualità inferiore alla precedente.

8.8 - I metodi di regressione non-lineare

Un approccio del tutto naturale per trattare la non-linearità con il metodo dei minimi quadrati ordinari è quello di introdurre nel modello delle trasformazioni non-lineari delle variabili originali (***Quadratic Ordinary Least Squares,* QOLS**). Questo significa, nel caso più comune, aggiungere, ad esempio, tutti i termini quadratici delle p variabili originali e i loro $p(p-1)/2$ prodotti misti, ottenendo una matrice del modello $\mathbf{X_M}$ con un numero di colonne pari a $2p + p(p-1)/2 + 1$.

E' evidente che in molti casi vengono a cadere le condizioni di applicabilità del metodo OLS, sia in quanto aumenta la correlazione tra le variabili del modello sia perchè il rapporto oggetti/variabili diviene sempre più sfavorevole. Tuttavia l'applicazione di metodi di selezione di un sottoinsieme ottimale di variabili (v. par. 8-3 e Cap. 10) fornisce certamente nuove potenzialità al metodo QOLS.

In modo analogo a quanto sopra esposto, è possibile utilizzare la regressione in componenti principali (PCR) aggiungendo anche termini non lineari delle componenti selezionate.

8.9 - Il metodo PLS non lineare

Anche per quanto riguarda il metodo PLS, è possibile utilizzare la strategia precedente utilizzando delle trasformazioni non-lineari delle variabili originali. In questo caso, le limitazioni precedenti tipiche di QOLS sono senz'altro meno rilevanti.
Tuttavia, per l'ottenimento di modelli non-lineari con il metodo PLS, sono state proposte altre due possibili vie che introducono direttamente nel metodo la non-linearità.
La prima di queste vie, proposta da Svante Wold (***Quadratic Partial Least Squares***, **QPLS**), prevede la modifica del passo 9 dell'algortimo PLS, cioè la modifica da lineare in quadratica della relazione tra le coppie di variabili latenti :

$$u_{im} = bt_{im} \quad \Rightarrow \quad u_{im} = b_1 t_{im} + b_2 t_{im}^2$$

Una seconda versione, proposta da Ildiko Frank (***Non-Linear Partial Least Squares***, **NLPLS**), prevede di trattare la non-linearità utilizzando l'algoritmo di *smoothing* presente nel metodo ACE (v.oltre par. 8-10), secondo l'espressione:

$$u_{im} = f_m(t_{im})$$

dove *f* è una generica funzione univariata non-lineare tra le coppie di variabili latenti.

8.10 - Il metodo Alternating Conditional Expectations (ACE)

Il metodo *ACE* è un metodo di regressione non-lineare proposto da Breiman e Friedman nel 1985, dalle caratteristiche molto particolari e di cui si conoscono ancora poche applicazioni.
Il modello *ACE* stabilisce tra descrittori e risposta una relazione del tipo:

$$g(\mathbf{y}) = \sum_j t_j(\mathbf{x}_j) + \boldsymbol{\varepsilon} \qquad j = 1,\mathrm{K}\ ,p$$

dove *g* e *t* sono funzioni univariate non-lineari arbitrarie e ε è il termine di errore.

Il metodo ricerca le migliori trasformazioni lineari t_j delle variabili indipendenti $\mathbf{x}_j$ per riprodurre la risposta **y** o, più rigorosamente, una sua trasformata $g(\mathbf{y})$, in modo da minimizzare la somma dei quadrati degli scarti ***RSS***:

$$RSS = \sum_i \; g(y_i) - \sum_j t_j(x_{ij})^2$$

Le funzioni sono scelte in modo da ottimizzare l'accordo tra i valori calcolati e sperimentali della risposta (la somma dei quadrati dei residui), con l'unico vincolo di essere funzioni di *smoothing*. Si assume inoltre che il valor medio della funzione *g* e delle funzioni *t* sia nullo (cioè, le funzioni vengono centrate) e che la varianza della funzione *g* sia unitaria.
Per le funzioni non viene assunta nessuna forma particolare: esse vengono calcolate iterativamente partendo da una trasformazione lineare. Ciascuna funzione viene ottenuta in forma di tabella di *n* coppie di punti (tante quanti sono gli oggetti): la prima colonna rappresenta il valore sperimentale della variabile $\mathbf{x}_j$, mentre la seconda colonna rappresenta il corrispondente valore trasformato $t_j(\mathbf{x}_j)$ attraverso il processo di *smoothing*.
Queste funzioni puntuali possono essere analizzate graficamente ed eventualmente interpolate con i metodi convenzionali di *fitting*.
Ciascuna trasformata $t_j(\mathbf{x}_j)$ può essere, se lo si desidera, ristretta da vincoli di linearità o monotonicità.
Nel caso più semplice, $g(\mathbf{y})$ è una trasformazione lineare che lascia invariato l'andamento della risposta. Una trasformata $g(\mathbf{y})$ non lineare viene utilizzata soprattutto quando si desidera un alto grado di *fitting*, essendo disposti a rinunciare a un modello più semplice nella risposta (lineare) e, sovente, anche a un maggior potere predittivo del modello.
Il calcolo delle trasformate di $\mathbf{x}_j$ avviene attraverso una procedura di *smoothing*, che ricerca la migliore relazione lineare tra i punti sperimentali contenuti all'interno di una *finestra* definita da una frazione stabilita di dati o ottimizzata dall'algoritmo stesso.
Un parametro, detto *span*, viene selezionato dall'utente tra valori compresi tra 0 e 1: il valore selezionato rappresenta la *frazione dei dati ordinati* (e non l'ampiezza dell'intervallo, come avviene nei comuni metodi di *smoothing*) delle variabili $\mathbf{x}_j$ che viene considerata nel processo di *smoothing*. Questo processo avviene mediante una regressione locale col metodo di regressione OLS, effettuata sulla frazione di dati considerata (vedi Fig.8-6). Nell'esempio di Fig. 8-6, *span* è uguale a 0.33: essendo 15 il numero di dati (i cerchi), ogni passo del processo di *smoothing* viene affettuato su 5 dati alla volta, cioè sul 33% dei dati. Il primo insieme di dati è costituito dai primi 5 dati ordinati (1,2,3,4,5). Poichè

il processo di *smoothing* consente di calcolare il valore centrale dell'insieme dal modello di regressione ottenuto dai 5 punti, il valore della trasformata dei primi due dati coincide con i dati stessi, mentre il primo valore calcolato riguarda il punto 3. Il successivo insieme di dati è costituito dai 5 dati 2,3,4,5,6: da questo viene calcolato il valore della trasformata corrispondente al dato 4. La procedura prosegue fino all'ultima quintupla (11,12,13,14,15): il dato 13 viene calcolato come detto, mentre la trasformata degli ultimi due coincide con i dati stessi.

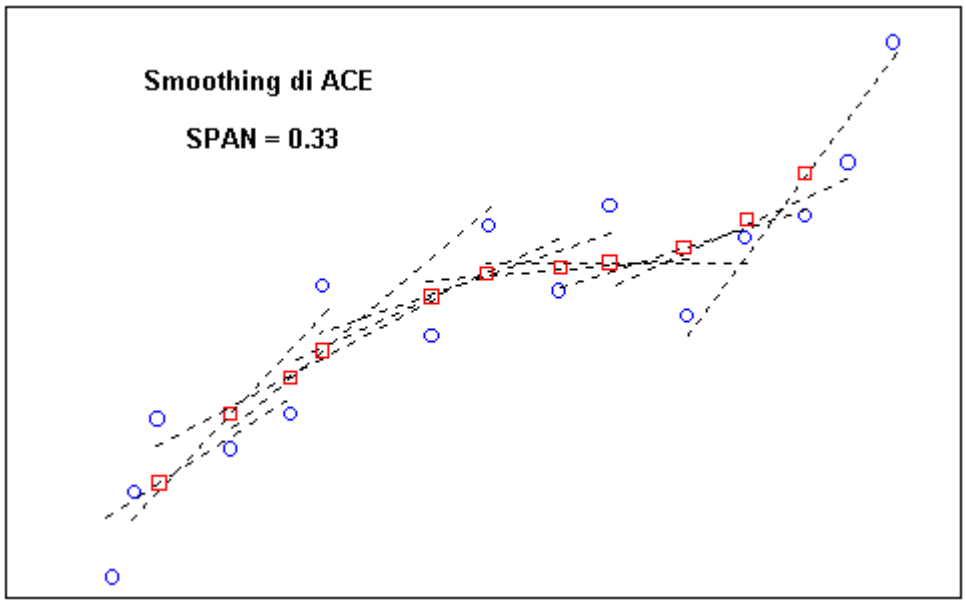


FIG. 8-6

Più è piccolo il valore di *span*, più piccola è la frazione di dati considerata e quindi più fine è il *fitting* dei dati, consentendo di rilevare accentuati livelli di non linearità. Viceversa, se *span* è grande, l'interpolazione lineare avviene su frazioni più grandi di dati e il processo di *smoothing* è più accentuato. Se il valore di span è zero, la frazione di dati per il processo di *smoothing* viene selezionata automaticamente tra i valori 0.05, 0.2 e 0.5 in funzione del miglior adattatamento locale ai dati.

I valori normalmente utilizzati per il parametro *span* sono 0.1, 0.2, 0.3, 0.4 e 0.5.
Il metodo è applicabile solo se il rapporto oggetti/variabili (*n/p*) è sufficientemente alto.
Il metodo fornisce, oltre ai valori di varianza spiegata in *fitting* o in predizione, le trasformate di ciascun predittore in modo grafico. Questo aspetto consente di 'vedere' la forma che ciascuna variabile dovrebbe assumere per riprodurre nel migliore dei modi la risposta. Ad ogni trasformata, è anche associata la varianza spiegata da ciascuna di esse, varianza che quindi può essere interpretata come il peso che ciascuna variabile ha nella regressione.

Esempio

Per differenti valori di *span*, il metodo *ACE* fornisce sui dati REGTEST i risultati riportati in Tab.8-12. In questo caso, i risultati in predizione non differiscono molto tra loro al variare del parametro *span*.

Span	R^2	R^2_{cv}	*SDEC*	*SDEP*
0.1	99.71	90.29	291.77	1681.08
0.3	99.05	90.95	525.59	1622.93
0.5	98.83	91.05	583.77	1613.53
0.7	98.90	91.25	565.06	1595.42
0.9	98.95	90.86	553.91	1631.26
ottimizzato	99.38	90.37	424.91	1674.16

TAB.8-12

Nella Tab.8-13 sono riportati i valori delle deviazioni standard relative alla varianza spiegata in predizione da ciascuna variabile.
Le Fig.8-7 - 8-11 riportano le trasformate delle 5 variabili indipendenti considerate; la Fig.8-12 riportata i valori sperimentali della risposta **y** contro quelli calcolati. Come si può osservare solo le variabili 2 e 3 hanno trasformate sostanzialmente lineari.

Span	x_1	x_2	x_3	x_4	x_5
0.1	0.042	0.291	0.724	0.021	0.075
0.3	0.031	0.273	0.721	0.021	0.030
0.5	0.012	0.281	0.715	0.024	0.036
0.7	0.017	0.279	0.726	0.028	0.041
0.9	0.040	0.262	0.740	0.018	0.061
ottimizzato	0.027	0.287	0.722	0.026	0.035

TAB. 8-13

Le Fig.8-13 e 8-14 riportano due possibili interpolazioni della trasformata della quarta variabile: un'interpolazione con un polinomio di quarto grado e un'interpolazione con una funzione logaritmica.

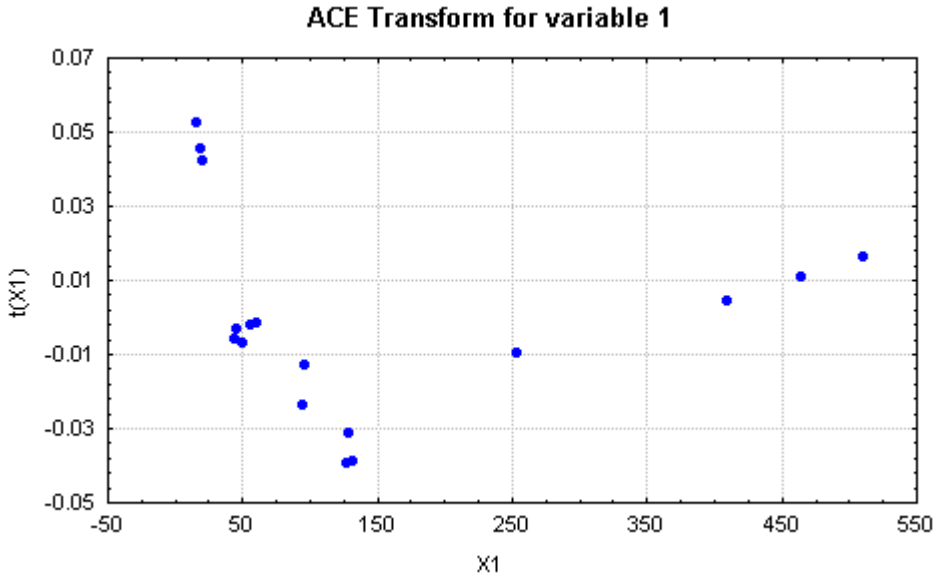


FIG.8-7

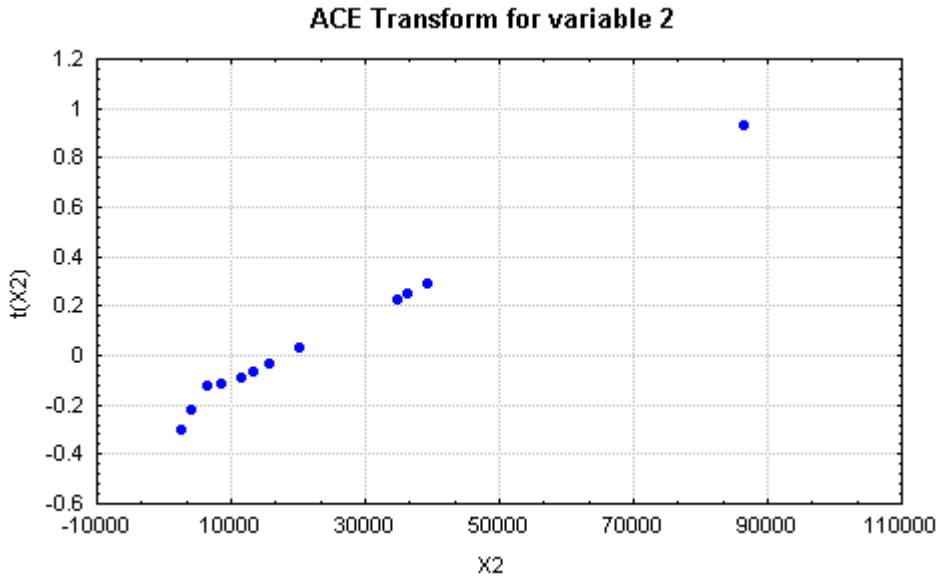

ACE Transform for variable 2
t(X2)
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-10000
10000
30000
50000
70000
90000
110000
X2

FIG.8-8

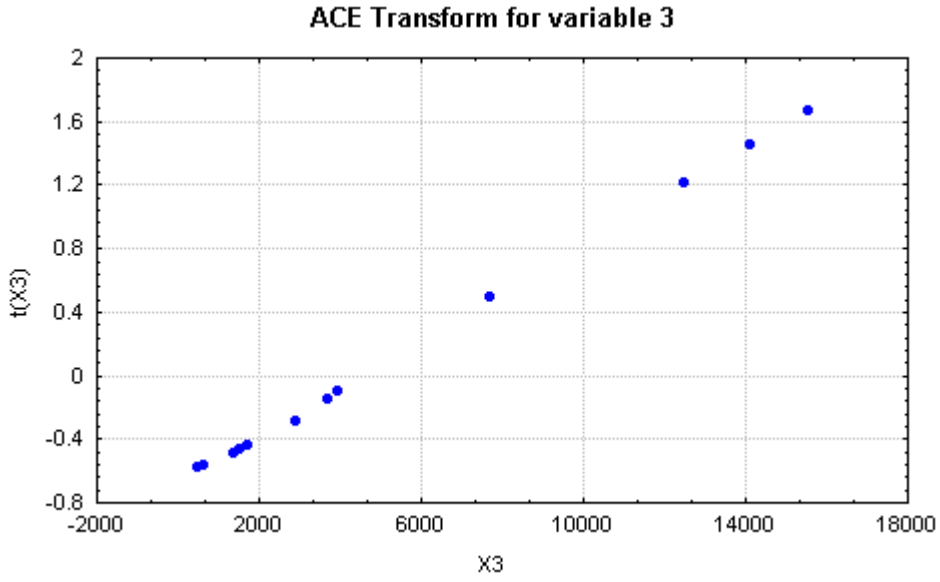

ACE Transform for variable 3
t(X3)
2
1.6
1.2
0.8
0.4
0
-0.4
-0.8
-2000
2000
6000
10000
14000
18000
X3

FIG.8-9

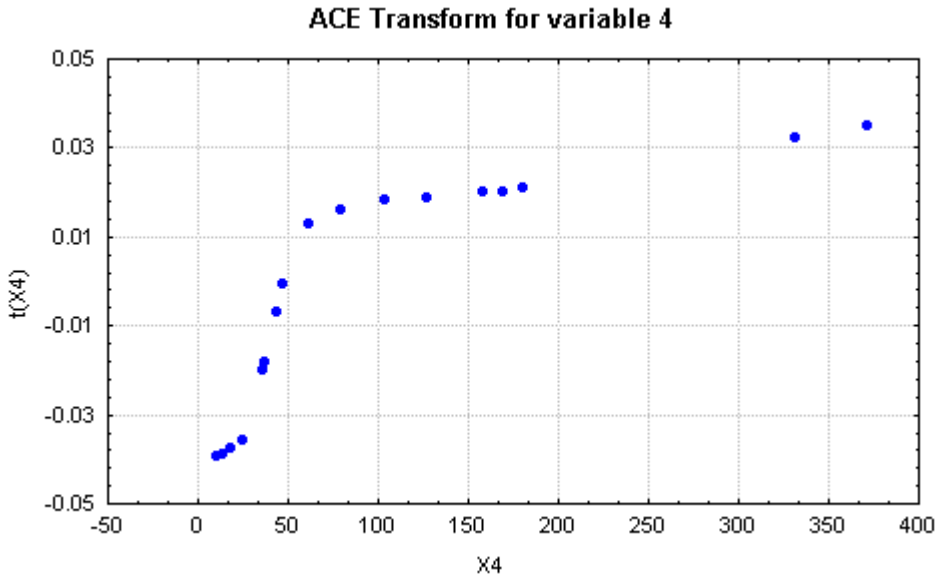


FIG.8-10

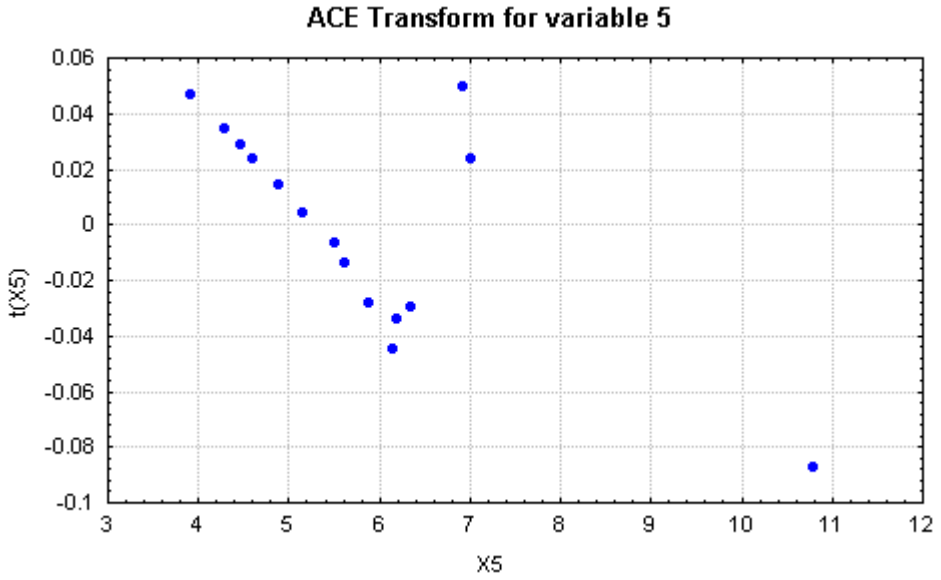


FIG.8-11

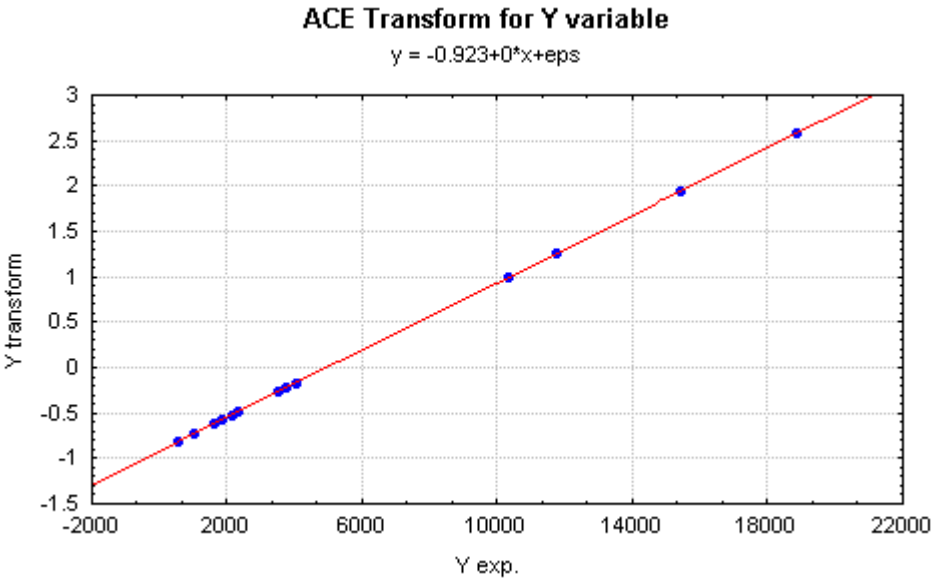


FIG.8-12

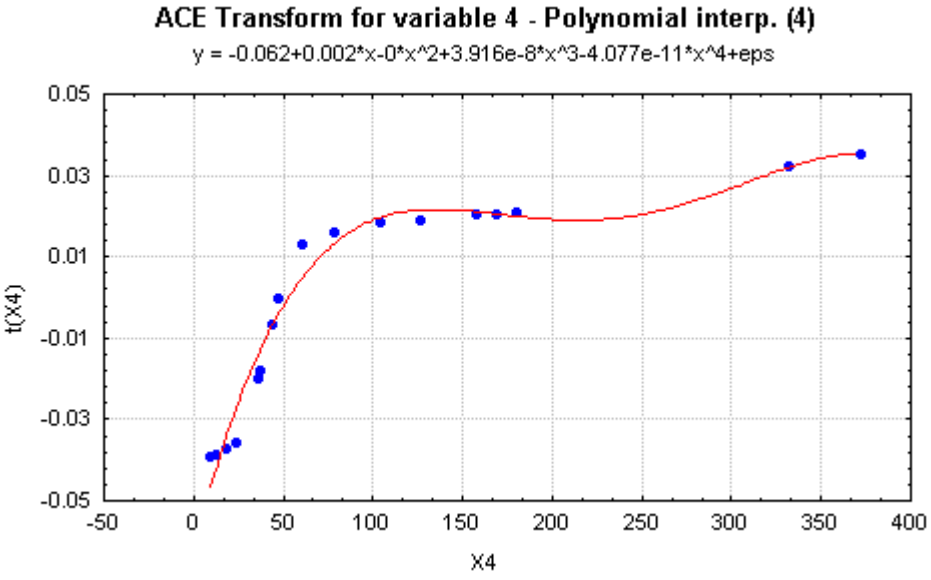


FIG.8-13

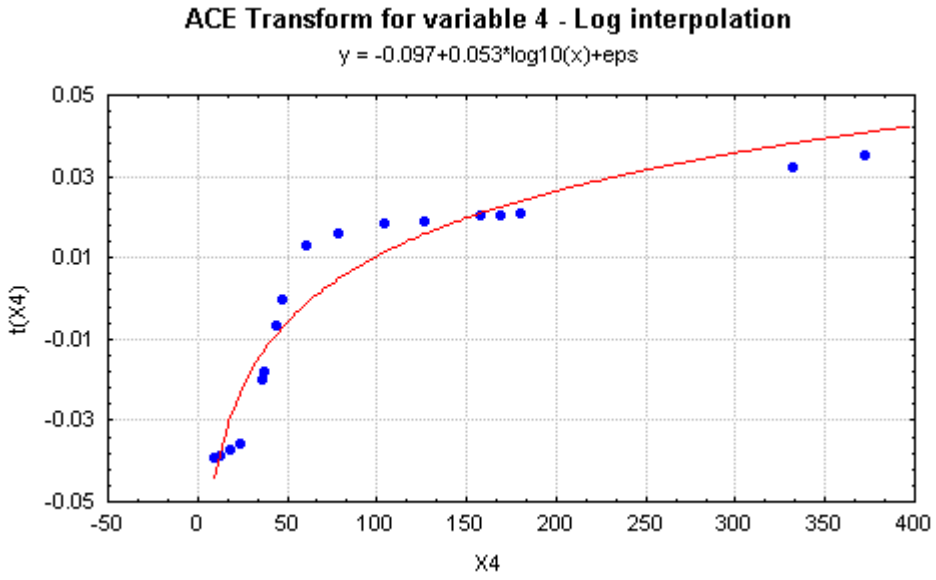


FIG. 8-14

8.11 - Modelli?!

Nonostante quanto è stato presentato fino a questo punto per ciò che riguarda le metodologie chemiometriche, è il caso di riassumere alcune questioni rilevanti per la costruzione di modelli e i termini per una loro corretta presentazione. Si fa qui riferimento esplicito ai modelli di regressione, che sono senz'altro i più utilizzati, ma molte delle considerazioni che seguono valgono anche per altri tipi di modelli.

In primo luogo, è bene sempre essere chiari sugli obiettivi che il modello cercato si prefigge: nella maggior parte dei casi, anche se non esplicitamente detto, il **modello ha finalità predittive**, cioè presuppone un suo utilizzo per applicazioni successive su dati la cui risposta non è ancora nota. Sia che la predizione di una risposta serva soltanto per uno *screening* orientativo, sia che in essa si cerchi un preciso valore quantitativo, la decisione per un suo utilizzo richiede la valutazione delle prestazioni predittive del modello. Come già detto, in nessuno dei parametri di regressione quali *RSS*, R^2, R^2_{adj} e *SDEC* è contenuto questo tipo di informazione.

Le procedure di validazione sono quindi un elemento imprescindibile per valutare la qualità predittiva del modello.

I parametri quali *PRESS*, R^2_{cv} e *SDEP* sono quelli indicati a questo scopo.
In secondo luogo, accade frequentemente che il numero di campioni utilizzabile per la costruzione di un modello non sia molto elevato: sono molto frequenti i casi in cui i campioni disponibili sono all'incirca compresi tra 8 e 15. In questi casi, le differenze tra la capacità descrittiva e la capacità predittiva del modello *possono* essere abbastanza frequentemente anche del 100%. Ad esempio, pochi sarebbero disposti ad accettare risultati ottenuti da modelli il cui valore di R^2 è inferiore al 50%. Ci si può chiedere come mai la maggior parte dei ricercatori è disposta ad accettare risultati ottenuti da un modello il cui valore di R^2 è uguale al 95%, ma il suo potere predittivo (ancorchè non valutato) potrebbe essere inferiore al 50% ed anche uguale allo 0%? Questa contraddizione, dovuta ad una non conoscenza del problema della predizione, diviene decisamente rilevante quando i modelli contengono più di una variabile indipendente.
Come dimostrato in precedenza, l'**aggiunta di variabili nel modello** - indipendentemente dalla loro rilevanza per la risposta - **non comporta mai un peggioramento delle qualità descrittive del modello**, ma, per la presenza di correlazione casuale, il valore di R^2 può solo crescere artificialmente. Al contrario, l'aggiunta di variabili non rilevanti per il modello comporta un abbassamento del valore di R^2_{cv} , evidenziando il **deterioramento delle qualità predittive del modello** stesso. Se, come gli statistici hanno dimostrato, questo è vero, deve essere evidente che alcune norme tradizionali per la certificazione di un modello di regressione non possono avere più un ruolo decisivo. Attenersi alle indicazioni sul rapporto oggetti/variabili (maggiore di 3), del parametro R^2 *adjusted* e agli esiti del test F di Fisher sono solo precauzioni per evitare la presenza di correlazioni casuali solo per coloro i quali non effettuano alcuna validazione del modello.
Infatti, inconsapevoli di questi aspetti, molti ritengono totalmente privi di significato modelli di regressione ove il **rapporto tra oggetti e variabili** sia inferiore a 3 (secondo le tradizioni statistiche classiche). In realtà, proprio mediante una validazione spinta (ad esempio, mediante procedure *leave-more-out*) è possibile produrre modelli con buone qualità predittive anche nei casi in cui ci si discosta in qualche misura da questa condizione. E' vero che questa non è una situazione auspicabile, ma è anche vero che modelli di questo tipo sono, in linea di principio, più affidabili in predizione di modelli più semplici (minor numero di variabili indipendenti), ma per i quali non è stata effettuata nessuna validazione.
In molti casi, la qualità del modello (quale qualità?) viene certificata mediante il **test F di Fisher**, cioè valori grandi del rapporto tra la varianza spiegata dal modello e quella dell'errore, relativamente ai gradi di libertà, indicherebbero

modelli dalle proprietà ottimali. Purtroppo, si trascura il fatto che il test F in regressione ci può soltanto dire **se esiste o non esiste un modello di regressione (in *fitting*)**, **e non quanto esso è buono e tanto meno quanto esso è predittivo**.
Il test F, che richiede come condizione di applicabilità che gli errori siano distribuiti normalmente (verifica di cui, normalmente, non viene data informazione alcuna), fornisce, ad un dato livello di probabilità, una risposta di tipo *si/no* alla domanda: *esiste un modello di regressione?*, e non una risposta quantitativa alla domanda: *quanto è buono il modello di regressione ottenuto?*.
Alla luce di queste considerazioni, si consiglia di riportare i modelli di regressione secondo uno schema, che nel caso più completo, potrebbe essere di questo tipo:

$$n = 13 \quad p = 3 \quad F_{(10,4)} = 54.5 \quad s = 0.611$$
$$R^2 = 94.3\,\% \quad SDEC = 0.543 \quad R^2_{cv} = 86.7\,\% \quad SDEP = 0.766$$
$$y = 1.23 + 4.57 \cdot x_1 - 0.55 \cdot x_2 + 13.45 \cdot x_3$$
$$y^s = 0.98 \cdot x_1 - 1.21 \cdot x_2 + 0.12 \cdot x_3$$

La prima riga contiene le informazioni sul numero di campioni (*n*), sul numero di variabili indipendenti (*p*), il valore del test F (con i relativi gradi di libertà) e il tradizionale valore dell'errore standard della stima (s). La seconda riga riporta i valori della varianza percentuale spiegata dal modello e dell'errore standard sia in *fitting* sia in predizione.
Le ultime due righe riportano il modello ottenuto con i coefficienti calcolati ed il corrispondente modello con i coefficienti standardizzati.
La terza riga descrive esplicitamente il modello con i coefficienti di regressione espressi nello spazio originale delle variabili al fine di poter utilizzare il modello per calcoli successivi. La quarta riga riporta il modello di regressione standardizzato, al fine di poter interpretare l'importanza delle diverse variabili nel modello mediante i loro coefficienti standardizzati. Questi coefficienti sono gli unici che consentono di valutare il peso (l'importanza) di ciascuna variabile nel modello, indipendentemente dalla scala in cui ciascuna variabile è rappresentata, e sono quindi di importanza fondamentale per gli aspetti interpretativi del modello. Nel caso in cui si siano utilizzati metodi di regressione come PCR o PLS, è necessario riportare, oltre al numero *p* delle variabili originali utilizzate, anche il numero *M* di componenti significative del modello.

Nota. Per quanto riguarda la validazione, è sempre opportuno riportare il dato relativo al metodo *leave-one-out* in quanto esso è l'unico parametro di validazione con cui è possibile confrontare i diversi modelli tra loro. Tuttavia, poichè per *n* abbastanza grande, $R^2_{LOO} \to R^2$, questo criterio di validazione può fornire risultati ancora troppo ottimistici. In molti casi è quindi importante effettuare validazioni più spinte e riportare anche tutti i diversi risultati ottenuti in validazione.

🕮 Bibliografia

I.E. FRANK E J. FRIEDMAN (1993): *A statistical view of some chemometrics regression tools.* Technometrics, **35**, 109.

9

LE COMPONENTI PRINCIPALI: GLI ALTRI METODI

9.1 - Introduzione

Sui principi fondamentali propri dell'Analisi delle Componenti Principali si fondano molte altre tecniche che differiscono tra loro in modo più o meno importante sia per l'obiettivo che si prefiggono sia per alcuni aspetti matematici. Tra queste le tecniche più note sono l'**Analisi dei Fattori**, l'**Analisi delle Corrispondenze**, l'**Analisi di Correlazione Canonica**, le **Scalature Multidimensionali**, l'**Analisi Evolutiva dei Fattori**.
A queste tecniche vengono dedicati solo alcuni cenni informativi di carattere teorico, senza alcuna pretesa di completezza. Inoltre, data l'importanza che rivestono i grafici degli scores e dei loadings visti in precedenza (Cap. 3), il paragrafo 9.2 è dedicato alla trattazione generale dei *biplots*.

9.2 - I biplots

Nell'analisi delle componenti principali, si è dimostrato che la matrice dei dati **X** può essere vista come la decomposizione:

$$\mathbf{X} = \mathbf{T}\mathbf{L}^{\mathbf{T}} = \mathbf{U}\Lambda^{1/2}\mathbf{V}^{\mathbf{T}}$$

dove $\mathbf{T} = \mathbf{U}\Lambda^{1/2}$ sono gli *scores*, $\mathbf{L} = \mathbf{V}$ sono i *loadings* e Λ è la matrice diagonale degli autovalori.
Poichè vale la seguente relazione:

$$\frac{1}{2} = \frac{\gamma}{2} + \frac{1-\gamma}{2} \qquad 0 \le \gamma \le 1$$

la precedente decomposizione può essere generalizzata come:

$$\mathbf{X} = \mathbf{U}\Lambda^{\gamma/2}\Lambda^{(1-\gamma)/2}\mathbf{V}^{\mathbf{T}} = \mathbf{A}\mathbf{B}^{\mathbf{T}}$$

dove **A** e **B** sono le due matrici definite come:

$$\mathbf{A} = \mathbf{U}\Lambda^{\gamma/2} \qquad \mathbf{B}^{\mathbf{T}} = \Lambda^{(1-\gamma)/2}\mathbf{V}^{\mathbf{T}}$$

In base alla scelta del valore di γ, possiamo ottenere diversi di tipi di proiezioni per il confronto tra oggetti e variabili.
Nel caso dell'analisi delle componenti principali, il *biplot* corrisponde al valore di $\gamma = 1$. Il *biplot* ottenuto in questo caso (**JK biplot**) preserva le distanze tra le righe (gli oggetti), cioè le distanze tra i vettori $\mathbf{a}_i$ sono *distanze euclidee*, mentre le distanze tra i vettori $\mathbf{b}_j$ sono le *distanze di Mahalanobis*.

Nota. Ricordiamo che le distanze euclidee sono distanze non pesate, mentre le distanze di Mahalanobis sono distanze pesate dalla matrice di covarianza e quindi pesano meno le distanze tra i punti che hanno un'elevata covarianza.

Nel caso opposto, $\gamma = 0$, i biblot ottenuti (**GH biplot**) preservano le distanze tra le colonne (le variabili), le cui differenze sono misurate da distanze euclidee, mentre riscalano in funzione della covarianza le distanze tra le righe (gli oggetti), le cui differenze sono misurate da distanze di Mahalanobis.
Il caso intermedio è definito per il valore di $\gamma = 0.5$ (**SQ biplot**).
.

9.3 - L'Analisi dei Fattori

Per quanto per molti problemi sia oggi più comunemente utilizzata l'analisi delle componenti principali, l'**Analisi dei Fattori (*Factor Analysis, FA*)** è la tecnica capostipite dei metodi di analisi dei fattori. Il proposito dell'analisi dei fattori è di descrivere, se possibile, le relazioni di covarianza tra molte variabili in termini di poche quantità sottostanti non osservabili, chiamate **fattori** o **variabili latenti**. L'idea sottesa da questo approccio è quella che un certo

gruppo di variabili, molto correlate tra loro e poco correlate con variabili di altri gruppi, possa essere descritto da un singolo fattore che coglie gli aspetti comuni al gruppo di variabili correlate tra loro.
L'analisi dei fattori, diversamente dall'analisi delle componenti principali, presuppone un modello matematico in cui è noto il termine di errore:

$$\mathbf{x}_{(1)} - \bar{x}_1 = f_{11}Z_1 + f_{12}Z_2 + \ldots + f_{1m}Z_m + \varepsilon_1$$
$$\mathbf{x}_{(2)} - \bar{x}_2 = f_{21}Z_1 + f_{22}Z_2 + \ldots + f_{2m}Z_m + \varepsilon_2$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\mathbf{x}_{(p)} - \bar{x}_p = f_{p1}Z_1 + f_{p2}Z_2 + \ldots + f_{pm}Z_m + \varepsilon_p$$

o, in termini matriciali:

$$\mathbf{x} - \bar{\mathbf{x}} = \mathbf{F}\,\mathbf{Z} + \boldsymbol{\varepsilon}$$

$$(p, 1) - (p, 1) = (p, M)\,(M, 1) + (p, 1)$$

dove **F** è la matrice dei *factor loadings*, **Z** è la matrice dei *factor scores* e **ε** è il vettore degli errori, detto **fattore di unicità** (*unique factor* o *specific factor*).
Questo modello è il più semplice modello matematico per cui i fattori $\boldsymbol{f_m}$ e gli errori $\boldsymbol{\varepsilon_i}$ sono variabili casuali e gli errori $\boldsymbol{\varepsilon_i}$ sono non correlati tra loro e con i fattori $\boldsymbol{f_m}$.
Un'importante differenza tra PCA e FA è che le componenti principali sono le quantità corrette per spiegare la varianza di uno spazio multivariato, mentre i fattori sono le entità appropriate a spiegare la covarianza del sistema.
Nell'analisi delle componenti principali i *loadings* rappresentano, come si è già detto, i coefficienti delle componenti principali, ma per la condizione di normalizzazione essi rappresentano le variabili in uno spazio a varianza unitaria.
L'**analisi dei fattori** (**Factor Analysis, FA**), i cui algoritmi matematici sono ancora quelli dell'analisi in componenti principali, capovolge quest'ultimo aspetto: i ***loadings dei fattori*** (*factor loadings*, *f*) hanno una varianza pari a quella del corrispondente autovalore nell'analisi delle componenti principali. La relazione tra *loadings* e *factor loadings* è la seguente:

$$f_{jm} = \mathrm{l}_{jm} \times \sqrt{\lambda_m}$$

dove ciascun *loading* è moltiplicato per la radice quadrata del corrispondente autovalore.
In questo caso

$$\lambda_m^2 = \sum_j f_{jm}^2$$

dove λ_m^2, corrispondente all'*m*-esimo autovalore della *m*-esima componente principale, rappresenta in questo caso il contributo dell'*m*-esimo fattore alla varianza totale delle variabili.
La varianza totale di una variabile rappresentata in un modello costituito da *M* fattori viene chiamata **communalità** (*communality*) ed è definita come:

$$h_j^2 = \sum_m f_{jm}^2$$

In questo caso, gli *scores* sono a loro volta divisi per la radice quadrata dei corrispondenti autovalori in modo che gli scores (*factor scores*) così ottenuti abbiano varianza unitaria:

$$\mathbf{X} = \mathbf{T}\mathbf{L}^{\mathbf{T}} = \mathbf{T}\Lambda^{-1/2}\Lambda^{1/2}\mathbf{L}^{\mathbf{T}} = \mathbf{Z}\mathbf{F}^{\mathbf{T}}$$

La matrice **Z** è la matrice degli scores calcolata dall'analisi dei fattori e la matrice **F** è la matrice dei *factor loadings*. In pratica, in questo caso, sono gli scores ad avere varianza unitaria, mentre i *factor loadings* di ciascun fattore hanno varianza pari all'autovalore corrispondente.
L'analisi dei fattori presuppone che le variabili $\boldsymbol{x_i}$ siano continue; nel caso in cui le variabili $\boldsymbol{x_i}$ siano discrete, il metodo prende il nome di **Latent-Structure Analysis**. In entrambi i casi, tuttavia, i fattori $\boldsymbol{f_m}$ vengono assunti continui. Nel caso in cui sia $\boldsymbol{x_i}$ che $\boldsymbol{f_m}$ siano discreti, il metodo prende il nome di **Latent-Class Analysis**.

9. 4 - L'Analisi delle Corrispondenze

L'**Analisi delle Corrispondenze (*Correspondence Factor Analysis, CFA*)** è un tipo di analisi multivariata, simile all'analisi delle componenti principali, che si applica di norma a dati rappresentati da *n* osservazioni descritte da variabili discrete, per ciascuna delle quali è nota la frequenza. L'informazione disponibile è l'insieme di frequenze

$$n_{ij}\,,\ i = 1,2,\cdots,r;\ \ j = 1,2,\cdots,c$$

dove n_{ij} è il numero di osservazioni (la frequenza) che per il campione *i*-esimo (la variabile di riga) hanno il valore *j*-esimo (la variabile di colonna). La matrice **N**, di dimensione r x c, ha come elementi n_{ij} e viene chiamata **tabella di contingenza**.
Si può quindi affermare che l'analisi delle corrispondenze coincide con l'analisi delle componenti principali effettuata su tabelle di contingenza; questo significa, matematicamente, che si utilizza la distanza χ^2 tra i vettori delle frequenze.
Obiettivo dell'analisi delle corrispondenze è quello di confrontare oggetti e variabili congiuntamente, cioè di trattare entrambi come punti nel medesimo spazio *M*-dimensionale.
E' possibile conseguire questo obiettivo solo se tutte le variabili sono misurate nelle stesse unità oppure se la matrice è stata pretrattata con un tipo particolare di scalatura in cui i singoli dati sono stati uniformati *contemporaneamente* sulle righe e sulle colonne, cioè trasformati in **profili**. Ciò significa che i dati originali, nell'analisi delle corrispondenze, vengono centrati sia sulle medie di colonna sia sulle medie di riga, secondo la relazione:

$$n'_{ij} = \frac{n_{ij}}{\sqrt{\bar{n}_i} \cdot \sqrt{\bar{n}_j}}$$

Il risultato della diagonalizzazione di **N** è una sequenza di coppie di vettori

$$\left(\mathbf{f}_1, \mathbf{g}_1\right), \left(\mathbf{f}_2, \mathbf{g}_2\right), \cdots, \left(\mathbf{f}_M, \mathbf{g}_M\right)$$

dove $\mathbf{f}_k$ (k = 1, 2, ... , r) sono gli r vettori *scores* delle righe di **N**, e $\mathbf{g}_k$ (k = 1, 2, ... , c) sono i c vettori *scores* delle colonne, M è il numero di componenti significative selezionate ed è al massimo uguale al valore minimo tra r e c. Poichè le coppie di vettori sono state definite nello stesso spazio, è possibile proiettare contemporaneamente sia i vettori di riga che i vettori di colonna in un grafico bidimensionale di due componenti: questa analisi globale dei dati consente di 'vedere' il comportamento delle righe, delle colonne e delle loro interrelazioni.
Caratteristica dell'analisi delle corrispondenze è che, a causa del tipo di scalatura effettuata, il primo autovalore è sempre uguale ad 1, cioè nella

diagonalizzazione viene persa l'informazione riguardante la prima componente, informazione che concerne l'aspetto quantitativo dei dati.
L'analisi delle corrispondenze nasce quindi per l'analisi delle tabelle di contingenza (dati costituiti da interi positivi), ma può venire applicata a matrici di dati i cui valori sono generici numeri reali non negativi. In questo caso è importante che le variabili siano omogenee, cioè rappresentino la stessa grandezza e siano descritte dalla stessa unità di misura. Nel caso in cui si vogliano analizzare matrici di dati tradizionali **X** (*n*, *p*), è possibile adottare direttamente la stessa procedura se le variabili sono omogenee (hanno tutte la stessa unità di misura); se invece le variabili sono tra loro eterogenee, è necessario trasformare tutte le variabili originali in categorie discrete (in classi) per ottenere una tabella di contingenza.
Questo tipo di analisi multivariata è particolarmente utile quando l'interesse è soprattutto incentrato al diretto confronto tra righe e colonne (ovvero, un **confronto tra oggetti e variabili** per dati di tipo **X** (*n*, *p*)). La proiezione contemporanea degli *scores* di riga e degli *scores* di colonna consente di studiare l'interazione tra oggetti e variabili, mediante l'utilizzo dei grafici **biplot**, così come avviene per l'analisi delle componenti principali, ove tuttavia la confrontabilità tra oggetti e variabili è più limitata.
Infatti una piccola distanza (euclidea) tra due oggetti indica che i loro profili sono tra loro simili; analogamente, una piccola distanza (euclidea) tra due variabili indica che le rispettive colonne espresse come percentuali sono tra loro simili. La corrispondenza tra oggetti e variabili viene valutata mediante l'angolo formato dalle rispettive congiungenti l'origine degli assi.
Tra PCA e CFA permane tuttavia una differenza significativa in quanto PCA viene effettuata di norma dopo aver sottratto dai dati il valor medio di ciascuna colonna (variabile), mentre CFA viene effettuata sottraendo a ciascun dato il valore atteso della frequenza $1/n_{..}$, assumendo l'indipendenza tra righe e colonne.
Strettamente collegata all'Analisi delle Corrispondenze è l'**Analisi delle Mappe Spettrali (*Spectral Map Analysis, SMA*)**, detto anche **Modello Log-Lineare (*LLM*)** o **Log-Linear Correspondence Factor Analysis (*LLCFA*)**.
Entrambi i metodi hanno gli stessi obiettivi, cioè il confronto tra l'informazione contenuta nelle righe e quella contenuta nelle colonne della matrice. Il tipo di dati analizzabile con questa tecnica è lo stesso analizzabile con CFA.
La differenza sostanziale consiste nel pretrattamento dei dati, ove il calcolo dei profili è qui sostituito dalla **doppia scalatura logaritmica** (logaritmo naturale) di riga e di colonna, secondo la relazione:

$$x'_{ij} = ln(x_{ij}) - \sum_i ln(x_{ij}) - \sum_j ln(x_{ij}) + \sum_i \sum_j x_{ij}$$

9.5 - L'Analisi di Correlazione Canonica

L'**Analisi di Correlazione Canonica (*Canonical Correlation Analysis, CCA*)** è un metodo che consente di studiare le correlazioni tra due blocchi di variabili **x** e **y**: questi blocchi sono designati generalmente come **X** (*n*, *p*) e **Y** (*n*, *r*).
Diversamente dai metodi di regressione, l'analisi di correlazione canonica non presume alcuna relazione causa-effetto tra il blocco **X** e il blocco **Y**.
La procedura consiste nel calcolo della matrice di correlazione totale **R**, decomposta nelle matrici $\mathbf{R}_{XX}$ (*p*, *p*) la matrice delle correlazioni interne al blocco **X**, $\mathbf{R}_{YY}$ (*r*, *r*) la matrice delle correlazioni interne al blocco **Y**, $\mathbf{R}_{XY}$ la matrice delle correlazioni incrociate tra i due blocchi ($\mathbf{R}_{XY} = \mathbf{R}_{YX}^{\mathbf{T}}$).
Obiettivo dell'analisi di correlazione canonica consiste nel determinare coppie di combinazioni lineari - una per ciascun blocco di variabili - aventi varianza unitaria, che siano tra loro massimamente correlate:

$$\mathbf{u}_k = \mathbf{X}\mathbf{a}_k$$
$$\mathbf{v}_k = \mathbf{Y}\mathbf{b}_k$$

e

$$V(\mathbf{u}_k) = \mathbf{a}_k^{\mathsf{T}} \mathbf{S}_{\mathbf{XX}} \mathbf{a}_k$$
$$V(\mathbf{v}_k) = \mathbf{b}_k^{\mathsf{T}} \mathbf{S}_{\mathbf{YY}} \mathbf{b}_k$$
$$Cov(\mathbf{u}_k, \mathbf{v}_k) = \mathbf{a}_k^{\mathsf{T}} \mathbf{S}_{\mathbf{XY}} \mathbf{b}$$

dove *V* e *Cov* indicano la varianza e la covarianza e **S** indica le rispettive matrici di covarianza.
Il metodo si propone di cercare i coefficienti **a** e **b** che rendano massima la correlazione *C* tra le coppie di combinazioni lineari **u** e **v**:

$$C(\mathbf{u},\mathbf{v}) = \frac{\mathbf{a}^{\mathsf{T}}\mathbf{S}_{\mathbf{XY}}\mathbf{b}}{\sqrt{\mathbf{a}^{\mathsf{T}}\mathbf{S}_{\mathbf{XX}}\mathbf{a}}\sqrt{\mathbf{b}^{\mathsf{T}}\mathbf{S}_{\mathbf{YY}}\mathbf{b}}}$$

per ogni *k*-esima coppia di combinazioni lineari e preservando il vincolo di ortogonalità tra le diverse coppie di combinazioni lineari.
Per conseguire questo obiettivo, viene effettuata la diagonalizzazione, cioè la ricerca degli autovalori λ_k e degli autovettori $\mathbf{e}_k$, della matrice:

$$\mathbf{R}_{\mathrm{XX}}^{-1/2} \cdot \mathbf{R}_{\mathrm{XY}} \cdot \mathbf{R}_{\mathrm{YY}}^{-1} \cdot \mathbf{R}_{\mathrm{YX}} \cdot \mathbf{R}_{\mathrm{XX}}^{-1/2}$$

La radice quadrata degli autovalori corrisponde alla correlazione tra ciascuna coppia di combinazioni lineari. Ciascuna coppia di coefficienti viene calcolata nel seguente modo:

$$\mathbf{a}_k = \mathbf{R}_{\mathrm{XX}}^{-1/2} \mathbf{e}_k$$

$$\mathbf{b}_k = \frac{\mathbf{R}_{\mathrm{YY}}^{-1} \mathbf{R}_{\mathrm{YX}} \mathbf{a}_k}{\mathbf{b}_k^{\mathrm{T}} \mathbf{R}_{\mathrm{YY}} \mathbf{b}_k}$$

dove il temine al denominatore è la deviazione standard (uno scalare), che normalizza il vettore **b** a varianza unitaria.

9.6 - Scalature multidimensionali

I metodi di questo tipo si basano sull'analisi della matrice delle distanze **D** (o di dissimilarità) ricavata dalla matrice dei dati **X**. Gli elementi d_{st} della matrice **D** rappresentano la distanze tra gli oggetti *s* e *t* (le righe di **X**).
Sia d_{st} la distanza tra gli oggetti *s* e *t* nello spazio originale *p*-dimensionale e $\hat{d}_{st}$ la corrispondente distanza nel sottospazio *M*-dimensionale, con $M < p$.
Da questo punto di vista, l'analisi delle componenti principali, nel sottospazio *M*-dimensionale, fornisce una soluzione per cui è minima la funzione *V* definita come:

$$V = min\left[\sum_s \sum_t \left(d_{st}^2 - \hat{d}_{st}^2\right)\right]$$

In altri casi, può essere preferibile ricercare sottospazi *M*-dimensionali che minimizzino funzioni differenti, quali, ad esempio,

$$L = min\left[\sum_s \sum_t \left(d_{st} - \hat{d}_{st}\right)^2\right]$$

oppure

$$L^* = min\left[\sum_s \sum_t w_{st}\left(d_{st} - \hat{d}_{st}\right)^2\right]$$

dove w_{st} sono gli elementi di una matrice di pesi statistici.
Questo tipo di analisi, detta ***Non-Linear Mapping*** (***NLM***) o ***Multi-Dimensional Scaling*** (***MDS***), ha lo scopo di ottenere proiezioni basso-dimensionali di spazi multivariati ad alte dimensioni, conservando nel miglior modo possibile, nello spazio ridotto, le distanze che sussistono tra gli oggetti nello spazio alto-dimensionale. Per questo motivo, di norma, si cercano proiezioni su 2 o 3 dimensioni, cioè $M = 2$ o 3. Questo metodo presuppone che tutte le variabili siano misurate su una scala di intervalli o di rapporti.
L'algoritmo più comune assume come caso particolare di L^* che $w_{st} = 1/d_{st}$ ed è basato sulla minimizzazione della seguente espressione:

$$min\left[\sum_{i<j} \frac{\left(d_{ij} - \hat{d}_{ij}\right)^2}{d_{ij}}\right]$$

dove d_{ij} e $\hat{d}_{ij}$ sono, rispettivamente, le distanze vere tra gli oggetti e le distanze nello spazio ridotto. Gli assi coordinati nello spazio ridotto non sono quindi combinazioni lineari delle variabili originali e non sono suscettibili di interpretazione. Nella ricostruzione delle coordinate degli *n* punti, l'origine delle coordinate vere dei punti è incognita e quindi è necessario considerare solo *n* - 1 punti, con l'*n*-esimo punto posto nell'origine.
Se **X** (*n*, *p*) è la matrice incognita delle coordinate, la matrice $\mathbf{XX^T}$ si ricava come rappresentazione della matrice $\mathbf{D^2}$. La matrice diagonale degli autovalori $\mathbf{\Lambda}$ viene calcolata da:

$$\mathbf{V^T}\left(\mathbf{XX^T}\right)\mathbf{V} = \mathbf{L}$$

dove **V** è la matrice degli autovettori di $\mathbf{XX^T}$ e gli autovalori uguali a zero sono *n* - *p*.
La matrice **X** (*n*, *p*) delle coordinate, cioè una loro rappresentazione centrata e lungo le direzioni di massima varianza, viene infine calcolata dall'espressione:

$$\mathbf{X} = \mathbf{V} \cdot \sqrt{\mathbf{\Lambda}}$$

Varianti delle scalature multidimensionali sono ***Analisi delle Coordinate Principali*** (***Principal Coordinates Analysis, PCoA***), il metodo ***Non-metric MultiDimensional Scaling*** (***NMDS***), utilizzato quando le variabili sono rappresentate su una scala ordinale.

9.7 - L'Analisi Evolutiva dei Fattori

L'analisi delle componenti principali applicata sequenzialmente a sottomatrici ordinate dei dati viene chiamata **Analisi Evolutiva dei Fattori (*Evolving Factor Analysis, EFA*)** ed è stata ideata per la soluzione di problemi tipicamente chimici, quali lo studio della comparsa e scomparsa nel tempo di diverse specie chimiche al variare nel tempo di alcune condizioni sperimentali (ad esempio, pH, lunghezza d'onda).
Se le *n* righe della matrice dei dati rappresentano i risultati ottenuti lungo una sequenza temporale, è possibile applicare l'analisi delle componenti principali, iterativamente, prima in avanti (*forward*) e successivamente all'indietro (*backward*). Nel primo caso, vengono effettuate *n* analisi in componenti principali su matrici costituite da *k* righe, con $1 \leq k \leq n$. Successivamente, vengono ripetute *n* analisi in componenti principali su matrici costituite da *k* righe, ove però le righe vengono via via aggiunte partendo dall'ultima.
L'analisi degli autovalori significativi viene effettuata graficamente proiettando i loro valori sull'asse delle ordinate in entrambe le fasi (*forward* e *backward*), relativamente all'asse delle ascisse che rappresenta la sequenza.
La procedura in avanti consente di rilevare l'emergere nel tempo di una nuova componente significativa, mentre la procedura all'indietro consente di individuare la scomparsa nel tempo di una componente significativa. La combinazione dei due grafici consente di determinare finestre temporali ciascuna delle quali viene associata alla presenza di una o più specie chimiche presenti.

🕮 **BIBLIOGRAFIA**

J.E. JACKSON (1991). *A User's Guide to Principal Components.* Wiley, New York (NY).

I.T. JOLLIFFE (1986). *Principal Component Analysis.* Springer-Verlag, New York (NY).

W.J. KRZANOWSKI (1988). *Principles of Multivariate Analysis. A User's Perspective*. Oxford Univ. Press, Oxford.

R.J. RUMMEL (1970), *Applied Factor Analysis.* Northwestern Univ. Press, Evanston (IL).

J.P. BENZECRI (1980). *L'Analyse des Correspondences.* Dunod, Parigi.

A. BASILEVSKY (1994). *Statistical factor analysis and related methods.* Wiley, New York (NY).

10

GLI ALGORITMI GENETICI

10.1 - Introduzione

Le tecniche di ottimizzazione sono molto spesso l'ultima risorsa quando è molto difficile o impossibile in pratica ottenere soluzioni ad un problema mediante metodi analitici, euristici o diretti. Le tecniche di ottimizzazione hanno lo scopo di ricercare una condizione ottimale di minimo o di massimo di una determinata risposta per un certo numero di parametri indipendenti da cui la risposta stessa dipende.
Esistono molti metodi di ottimizzazione il cui utilizzo è ben consolidato, quali il *metodo della discendente più ripida* (*steepest ascent*), i *metodi di Fletcher-Powell* e *di Davidon-Fletcher-Powell*, il *metodo di Marquart*, il *metodo Simplex*. Qui ci occupiamo esclusivamente di un metodo di ottimizzazione - gli **algoritmi genetici** (***Genetic Algorithms, GA***) - proposto recentemente e particolarmente idoneo a trattare anche i problemi connessi con la ricerca dei migliori modelli.
Gli algoritmi genetici sono una potente strategia di ricerca chemiometrica applicabile in problemi di ottimizzazione di larga scala. Proposti da Holland nel 1975, in questi ultimi anni sono stati largamente utilizzati per la soluzione di problemi di ottimizzazione in campi molto diversi tra loro:

- analisi delle immagini
- riconoscimento di modelli
- metodi di modellamento
- robotica
- progetto di reti
- ricerca operativa

Base per l'applicazione degli algoritmi genetici è la descrizione dei valori che ogni variabile può assumere mediante un codice binario, ove ciascun termine binario costituisce un **bit**, cioè un numero uguale a 0 o 1.
Nel linguaggio degli algoritmi genetici, ciascuna codifica binaria di un numero (una variabile, un parametro numerico) costituisce un **gene**, cioè un insieme di bits; l'insieme di più geni, cioè la concatenazione di gruppi di bits, costituisce un **cromosoma**. Ciascun cromosoma è una rappresentazione di un punto nello spazio *p*-dimensionale dei parametri indipendenti da ottimizzare, ove ciascun parametro è rappresentato da un gene.
A ciascun cromosoma, cioè all'insieme dei valori dei parametri da ottimizzare (il numero di geni) è associata una certa risposta (il valore della funzione i cui parametri hanno i valori numerici rappresentati nel cromosoma.
La strategia generale per l'ottimizzazione della funzione (la ricerca del minimo o del massimo della funzione) è rappresentata nella Fig. 10-1.

Una volta prefissata la dimensione *N* della popolazione, nella fase iniziale la popolazione viene generata utilizzando le migliori *N* risposte tra le tante ottenute con valori a caso dei parametri. Successivamente si passa alla fase evolutiva in cui una procedura di combinazione (accoppiamento) dei casi presenti nella popolazione viene alternata iterativamente ad una procedura di mutazione casuale dei casi stessi. In entrambe le procedure, per ogni nuovo caso ottenuto (il cromosoma), viene calcolata la risposta e, se questa è migliore di quelle attualmente presenti nella popolazione, il cromosoma entra a far parte della popolazione sostituendo il peggiore tra quelli presenti. La fase evolutiva, che alterna accoppiamenti e mutazioni, procede fino a che non viene soddisfatto qualche criterio di stop.

L'algoritmo più diffuso per la trasformazione di parametri definiti da numeri reali in un codice binario è noto col nome di ***Binary F6***. Ogni numero è codificato da 22 bit, potendo quindi rappresentare numeri che vanno da 0 a 2^{22}-1 (corrispondente al numero intero 4194303).
Pertanto, se dobbiamo rappresentare 3 numeri reali (i valori attuali di 3 variabili di una funzione risposta da ottimizzare) in codice binario, avremo un cromosoma di 66 bits, costituito da 3 geni di 22 bits ciascuno.

Il processo di ottimizzazione si effettua attraverso una codifica dei valori numerici reali in un determinato intervallo (ad esempio, tra -100 e +100).

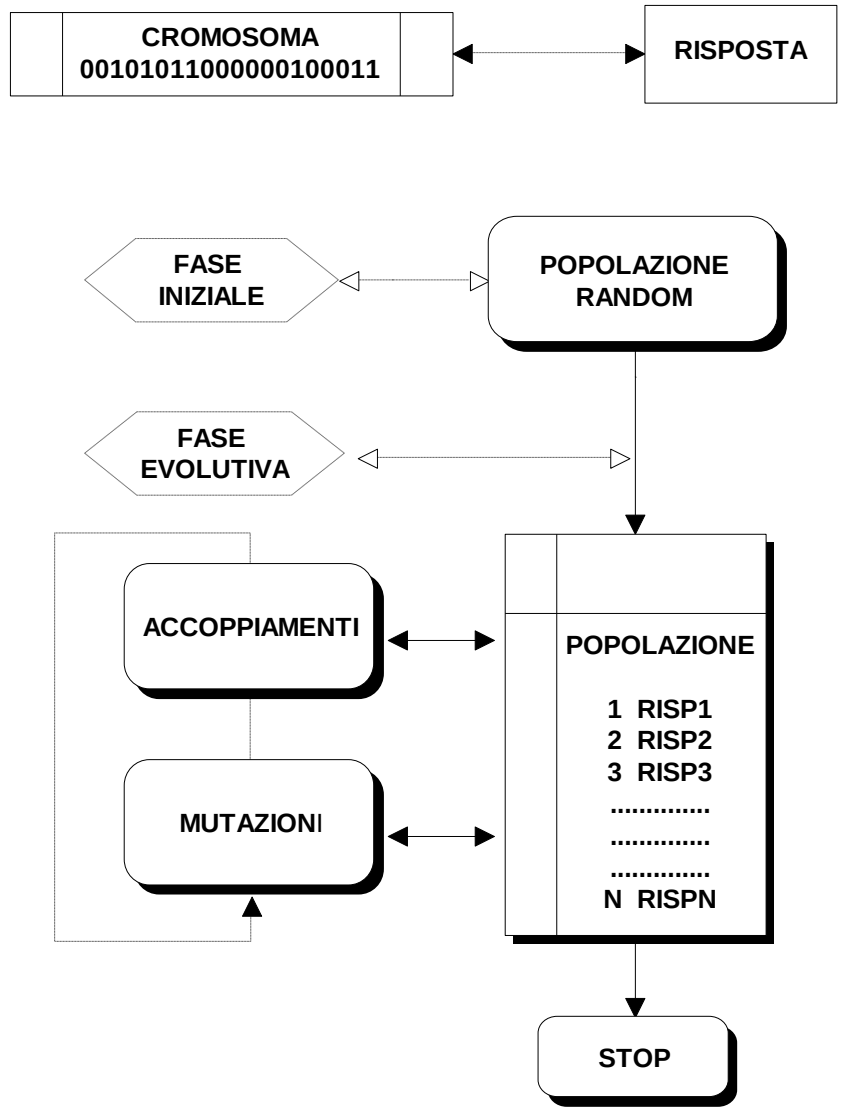


Fig. 10-1

In base all'algoritmo *F6*, la decodifica di un numero rappresentato in un codice binario avviene con i seguenti passi:

1. Si preleva la stringa di 22 bits relativa al numero considerato (un gene).

2. Si converte il codice binario nel numero intero corrispondente *x*:

$$x = \alpha_{21}2^{21} + \mathrm{L} \; + \alpha_1 2^1 + \alpha_0 2^0 = \sum_{k=21}^{0} \alpha_k \cdot 2^k$$

   dove ♋ può assumere i valori 0 o 1.

3. Si moltiplica il numero intero ottenuto per 0.00004768372718899898 e si sottrae 100 per ottenere il valore decimale del numero compreso tra -100 e 100.

Una volta decodificati tutti i valori delle variabili considerate, si valuta il valore della funzione risposta da ottimizzare, associandola al cromosoma considerato.

Ad esempio, il cromosoma di 22 bits:

21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1

corrisponde al numero intero decimale 165377**,** come risulta dalla Tab.10-1 ove sono riportati i valori per i bits non nulli.
Il numero ottenuto viene riscalato per ottenere un numero nell'intervallo definito: viene moltiplicato per il numero 0.00004768372718899898 per dare il valore 7.885791751335085; sottraendo 100, si ottiene il valore finale di -92.11420824866492.
Si osservi che se le variabili da ottimizzare sono già di tipo binario (si/no, presente/assente, etc.), ciascun bit (0/1) rappresenta direttamente il valore della variabile.

sequenza bit	termine binario	valore decimale	somma dei valori
17	2^{17}	131072	131072
15	2^{15}	32768	163840
10	2^{10}	1024	164864
9	2^{9}	512	165376
0	2^{0}	1	165377

TAB. 10-1

Nota. La procedura per riscalare il numero binario x corrispondente al valore decimale x' è una scalatura di intervallo generalizzata (v. capitolo 2).
Infatti, ad esempio, se il valore intero massimo è il numero 4194303, il valore minimo è 0 e l'intervallo in cui riscalare il dato è 200 (ottenuto da 100 - (-100)), il valore riscalato x' è dato dall'espressione:

$$x' = 200 \cdot \frac{x - 0}{4194303 - 0} - 100 = 0.00004768372719 \cdot x - 100$$

10.2 - Gli algoritmi genetici per la selezione delle variabili

I metodi basati sugli algoritmi genetici sono metodi di ottimizzazione in cui i parametri da ottimizzare sono definiti da numeri binari concatenati per formare un cromosoma.
Una volta definiti i parametri da ottimizzare e i loro intervalli di variazione, e i parametri necessari per la gestione del processo genetico di ottimizzazione, la procedura generale si basa sui seguenti passi:

fase iniziale:

viene estratto un certo numero di cromosomi a caso e per ciascuno di essi viene calcolata la risposta corrispondente (dopo decodifica dei singoli parametri); i migliori entrano a far parte della popolazione iniziale, in ordine decrescente di qualità della risposta.

fase evolutiva:

è costituita da due fasi distinte: una fase di generazione di cromosomi che derivano dall'accoppiamento dei cromosomi presenti nella popolazione ed il calcolo delle risposte corrispondenti; una fase di mutazione dove ciascun cromosoma della popolazione, a turno, può casualmente subire delle mutazioni: in questo caso, viene calcolata la corrispondente risposta.

Se le risposte calcolate sono migliori della risposta corrispondente all'ultimo cromosoma della popolazione (il peggiore della popolazione esistente), i cromosomi corrispondenti entrano a far parte della popolazione, nella posizione che loro compete in base alla qualità della risposta. In caso contrario, i cromosomi provati vengono scartati.

Le due fasi (di generazione e mutazione) si susseguono iterativamente, fino a che non viene soddisfatto un qualche criterio per porre termine alla ricerca della popolazione ottimale. I criteri di arresto più comuni sono definiti dal raggiungimento di un numero massimo prefissato di iterazioni o quando la popolazione attuale rimane inalterata per un numero prefissato di iterazioni.

Tra i diversi usi degli algoritmi genetici, quello di maggior interesse in questo contesto riguarda la loro applicazione nella ricerca del miglior sottoinsieme di variabili che ottimizzano un modello di classificazione o di regressione.
In questo caso, il singolo *gene* è costituito da un **1** o uno **0**, valori che rappresentano, rispettivamente, *la presenza o l'assenza della variabile nel modello*.
Il *cromosoma* è quindi costituito da tanti **1** e **0** quante sono le variabili.
Ad esempio, se abbiamo 10 variabili, il seguente cromosoma:

variabile	1	2	3	4	5	6	7	8	9	10
valore binario	0	1	1	1	0	0	1	1	0	0

indica la presenza nel modello delle variabili 2, 3, 4, 7 e 8, mentre le restanti variabili sono momentaneamente escluse.
Per ciascun cromosoma viene calcolata una risposta, cioè il valore del parametro da ottimizzare. Nel caso di un modello di regressione, il parametro da

massimizzare può essere R^2_{cv} (la percentuale di varianza spiegata dal modello in predizione), mentre nel caso di un modello di classificazione, il parametro da minimizzare può essere il rischio di errore $MR_{cv}\%$ (*misclassification risk* in predizione).

La procedura complessiva è la seguente:

🗀 **Fase iniziale**

- Si definisce la dimensione della popolazione (il numero di cromosomi) da cui procedere nell'evoluzione (ad esempio, 100)

- Si costruiscono *a caso* un certo numero di cromosomi, maggiore della dimensione della popolazione (ad esempio, 300)

- Si valuta la risposta di ciascuno dei cromosomi iniziali.

- Si inseriscono nella popolazione i cromosomi migliori, in ordine decrescente di risposta.

🗀 **Fase evolutiva**

- Si selezionano coppie di cromosomi dalla popolazione, con una probabilità di selezione proporzionale alla qualità del cromosoma e si procede, con una **probabilità di accoppiamento** (***cross-over probability***) prefissata, all'accoppiamento dei due cromosomi-genitori, generando due cromosomi-figli in quella che si chiama fase di **generazione**.

- Per ciascun cromosoma-figlio viene valutata la risposta e, se questa è migliore di una delle risposte associate ai cromosomi della popolazione, il cromosoma-figlio viene inserito nella popolazione, nella posizione che gli compete, mentre l'ultimo cromosoma della popolazione viene eliminato. Diversamente, il cromosoma-figlio viene scartato.

- Una volta effettuato un certo numero prefissato di accoppiamenti (e le relative valutazioni della risposta), si passa alla fase evolutiva che prevede la **mutazione** dei cromosomi della popolazione attualmente esistente. Ogni cromosoma viene analizzato e, in base ad una **probabilità di mutazione** (***mutation probability***) prefissata e in generale molto più piccola della

probabilità di accoppiamento, alcuni dei cromosomi vengono mutati in uno o più geni che lo compongono (da 0 in 1 o da 1 in 0).

La fase evolutiva, costituita dalla generazione di nuovi cromosomi e dalla mutazione di cromosomi, procede fino a che non si raggiunge una condizione che pone termine alla procedura.

🗀 **Fase finale**

La procedura viene fermata in base a criteri predefiniti, quali, ad esempio:

- Quando la procedura ha superato un numero massimo prefissato di iterazioni (da alcune centinaia di iterazioni fino ad alcune migliaia).

- Quando la popolazione non si rinnova per un certo numero di iterazioni (la qualità della popolazione non migliora).

10.3 - Lo sviluppo dell'algoritmo genetico

Partendo da due cromosomi-genitori (A e B), uno dei possibili algoritmi per la generazione dei cromosomi di base (SA e SB) dai quali ottenere due nuovi cromosomi (S1 e S2), è il seguente:

A	1	1	0	1	0	0	1	0	1	1
B	1	1	0	1	1	1	0	0	1	0
SA	1	1	0	1	**0**	**0**	**1**	0	1	**1**
SB	1	1	0	1	**1**	**1**	**0**	0	1	**0**

Gli elementi comuni ad entrambi i cromosomi A e B vengono mantenuti costanti in modo da conservare il patrimonio genetico (la qualità della risposta) presente. Gli elementi diversi nei due cromosomi A e B possono invece essere cambiati (4 in grassetto). Ciascuno degli elementi evidenziati viene mutato in base alla regola:

- viene estratto un valore a caso di probabilità per ciascun gene che può mutare;

- se questo valore è inferiore alla probabilità predefinita di *cross-over*, il valore del bit viene cambiato nel suo opposto, altrimenti rimane invariato.

Nell'esempio, se nel cromosoma SA cambiano il primo ed il terzo elemento (tra quelli che possono cambiare) e nel cromosoma SB cambiano il primo, il secondo ed il quarto elemento, otteniamo i seguenti due cromosomi-figli:

S1	1	1	0	1	**1**	**0**	**0**	0	1	**1**
S2	1	1	0	1	**0**	**0**	**0**	0	1	**1**

La probabilità di *cross-over* deve essere sufficientemente alta da consentire che avvengano alcune delle possibili inversioni.
Di norma, la selezione dei cromosomi-genitori avviene con una probabilità direttamente proporzionale alla qualità della risposta associata a ciascun cromosoma. A questo scopo, uno degli algoritmi più utilizzati per la selezione dei cromosomi-genitori è quello che si chiama ***roulette wheel***. Ognuno dei due cromosomi-genitori viene selezionato secondo l'algoritmo:

1. Si calcola la somma *t* delle risposte di tutti i cromosomi della popolazione (risposta totale).

2. Si genera un numero a caso compreso tra 0 e t.

3. Si seleziona il primo cromosoma della popolazione la cui risposta, cumulata alle risposte dei cromosomi che lo precedono, supera o è uguale al valore casuale ottenuto.

Supponiamo, ad esempio, che le risposte di 10 cromosomi che costituiscono la popolazione siano le seguenti:

cromosoma	1	2	3	4	5	6	7	8	9	10
risposta	78	72	70	57	48	45	27	23	10	8
risposta cum.	78	150	220	277	325	370	397	420	430	438

Supponiamo ore di estrarre 6 numeri a caso compresi tra 0 e 438. In funzione dei numeri estratti verranno selezionati i seguenti cromosomi:

numero casuale	28	251	135	344	432	71
cromosoma selezionato	1	4	2	6	10	1

Nella fase di mutazione genetica, ciascun cromosoma presente nella popolazione viene preso in considerazione una volta. Per ciascuno dei suoi

elementi viene estratto un valore a caso di probabilità: se questo valore è inferiore alla probabilità predefinita di mutazione, allora l'elemento viene scambiato.
Supponiamo, ad esempio, di avere 3 cromosomi di lunghezza 4. Per ciascun gene di ciascun cromosoma viene estratto un numero a caso compreso tra 0 e 1 e viene confrontato con la probabilità di mutazione (ad esempio, 0.005):

cromosoma prima della mutazione				*numeri casuali estratti*				*cromosoma dopo la mutazione*			
0	1	1	1	.125	.230	**.002**	**.005**	0	1	**0**	**0**
1	1	0	0	.976	.543	.234	.621	1	1	0	0
1	0	1	0	.780	**.001**	.312	.911	1	**1**	1	0

La probabilità di mutazione deve essere sufficientemente piccola da evitare di generare per mutazione cromosomi troppo casualmente diversi da quelli che costituiscono la popolazione attuale, allontanandosi esageratamente dalla probabile regione ottimale.

Un esempio di possibile inizializzazione dei parametri di un algoritmo genetico è il seguente:

numero di cromosomi iniziali creati a caso:	300
numero di cromosomi della popolazione:	100
probabilità di accoppiamento:	0.90
probabilità di mutazione:	0.005
numero massimo di iterazioni:	100000

10.4 - Varianti degli algoritmi genetici

Per quanto concerne l'algoritmo di selezione dei cromosomi-genitori, una variante utilizzata riguarda la linearizzazione della risposta. Questo significa che l'algoritmo di selezione *roulette wheel* viene modificato equispaziando linearmente le risposte.
Assumendo che il metodo originale *roulette wheel* introduca un bias del 100% nella selezione delle coppie di genitori, è possibile modulare con continuità il

metodo *roulette wheel*, trasformando, ad esempio, i valori della risposta attraverso la seguente funzione:

$$r_i' = r_i^{1/(101-bias\%)} \qquad 0 \le r_i \le 1$$

dove r_i è il valore delle singole risposte e *bias* è un valore predefinito. Se *bias* è uguale a 100 la selezione avviene secondo la modalità tradizionale in quanto le risposte vengono utilizzate così come sono ($r = r'$), mentre, per un *bias* uguale a 0, le singole risposte tendono a 1 e ciascun cromosoma della popolazione ha approssimativamente la stessa probabilità di essere selezionato (circa $1/N$, dove N è la dimensione della popolazione).
Per quanto riguarda la popolazione dei cromosomi in evoluzione, di norma, se vengono prodotti cromosomi già presenti nella popolazione, questi vengono esclusi. Tuttavia, una possibile variante del metodo consiste nel permettere che i cromosomi di buona qualità (cioè in grado di entrare nella popolazione), anzichè venire esclusi, possano invece ripetersi. Un'altra variante non consente di duplicare i cromosomi nella popolazione, ma, alla fine di ogni fase evolutiva, duplica semplicemente il migliore di essi (il primo della graduatoria, **elitismo**).

Esempio

Alle 5 variabili dell'esempio REGTEST trattato in regressione (v. cap. 7) sono stati aggiunti i quadrati ed i termini misti, per un totale di 20 variabili complessive.
Nella Tab.10-2 sono riportati i migliori 10 modelli ottenuti mediante l'applicazione di **algoritmi genetici per la selezione delle variabili**.
Come si può facilmente osservare, i primi 10 modelli riportati hanno prestazioni predittive molto simili tra loro (da 97.79 a 97.36). Il miglior modello è un modello di dimensione 2, costituito dalle variabili 1 e 10.
Le variabili 10, 16 e 20, trovate nei migliori modelli, sono rispettivamente:

$$x_5^2 \ , \ x_2x_4 \ , \ x_4x_5$$

ID	*n. var.*	R^2	R^2_{adj}	R^2_{cv}	*variabili*
1	2	98.56	98.35	97.79	3 10
2	3	98.60	98.28	97.33	1 3 10
3	3	98.69	98.38	97.70	1 10 20
4	2	98.47	98.26	97.69	1 10
5	3	98.69	98.35	97.49	3 10 20
6	3	98.61	98.29	97.47	1 10 16
7	3	98.68	98.44	97.45	3 10 16
8	2	98.48	98.26	97.45	3 5
9	2	98.40	98.17	97.38	1 5
10	3	98.50	98.16	97.36	1 3 5

TAB. 10-2

🕮 Bibliografia

L. DAVIS (Ed.) (1991). Handbook of Genetic Algorithms. Van Nostrand, New York, N.Y.

R. LEARDI, R. BOGGIA E M. TERRILE (1992). *Genetic algorithms as a strategy for feature selection*. J. of Chemometrics, **6**, 267-281.

R. TODESCHINI (1997) - *Moby Digs - Models By Descriptors In Genetic Selection* - Software per la selezione di variabili (Win95).

11

I CRITERI MULTIPLI DI DECISIONE

11.1 - Introduzione

In tempi recenti ha acquisito sempre più importanza l'analisi di strategie per la scelta di decisioni basate su criteri multipli (***MultiCriteria Decision Making***). E' questo un tipo di problema del tutto generale e molto comune, poichè di norma, moltissime delle nostre scelte sono basate su una serie di preferenze definite (consciamente o inconsciamente) sulla base di più criteri. Ad esempio, nella scelta di un'automobile, incidono, in diversa misura, il prezzo, la marca, la dimensione, il colore, il consumo, la cilindrata, i confort, eccetera. La decisione finale comporta la scelta di un'automobile tra tante possibili sulla base di un compromesso tra diversi criteri.
Le diverse metodologie per giungere ad una scelta ottimale (di un singolo campione o di un piccolo sottoinsieme ottimale) si fondano sulla costruzione di una graduatoria (*ranking*) delle diverse possibilità. Per poter ottenere questa graduatoria, dobbiamo essere in grado di esplicitare matematicamente per ogni singolo criterio le condizioni che si ritengono ottimali e l'importanza (il peso) che vogliamo assegnare ad ogni criterio.

I principali metodi per le scelte multicriterio sono:

- Diagrammi di Pareto
- Funzioni di desiderabilità
- Funzioni di utilità
- Funzioni di dominanza
- Funzioni di preferenza
- Metodo della minima distanza

Molti di questi metodi richiedono di esplicitare valori, caratteristiche e situazioni che si ritengono ottimali e quelli che si ritengono inaccettabili e la modalità con cui si passa da un caso all'altro. Utilizzando queste tecniche ci si

accorge che la vera difficoltà consiste proprio in questo passo: esso infatti richiede, in qualche modo, di matematizzare criteri di decisione che per lo più non sono stati completamente definiti, che hanno sempre inconsapevolmente avuto delle zone d'ombra o delle assunzioni non esplicitate.

Nota. I metodi basati sulle funzioni di dominanza e sulle funzioni di preferenza sono stati qui modificati rispetto alle proposte originali degli autori al fine di dare migliori proprietà matematiche alle funzioni definite.

11.2 - I diagrammi di Pareto

I diagrammi di Pareto costituiscono la metodologia più antica nell'ambito delle strategie decisionali. Quando l'analisi viene effettuata con un unico criterio, il diagramma di Pareto consiste in un semplice istogramma, ordinato in modo che da sinistra a destra vengano rappresentati i campioni in ordine decrescente di ottimalità.
In Fig. 11-1 viene riportata la velocità massima delle auto considerate nei dati AUTO, ordinata in modo decrescente. I primi campioni (20 e 21), rappresentano i punti ottimali di Pareto secondo l'unico criterio richiesto, cioè quello della massima velocità.
Quando i criteri da esaminare sono contemporaneamente due, il grafico di Pareto viene costruito riportando i valori dei campioni in un grafico di dispersione (*scatter plot*), dove ogni asse rappresenta un criterio (Fig. 11-2). In questo caso sono possibili quattro diverse situazioni di ottimalità sulla base delle combinazioni che i due criteri possono assumere: (+,+ ; +,- ; -,+ ; -,-), che rappresentano rispettivamente le seguenti richieste di ottimalità: valori massimi per entrambi i criteri; valore massimo per il primo criterio e minimo per il secondo; valore minimo per il primo criterio e massimo per il secondo; valori minimi per entrambi i criteri.
Nei corrispondenti quadranti del grafico è possibile inviduare i cosiddetti punti ottimali di Pareto (*Pareto optimal points*).

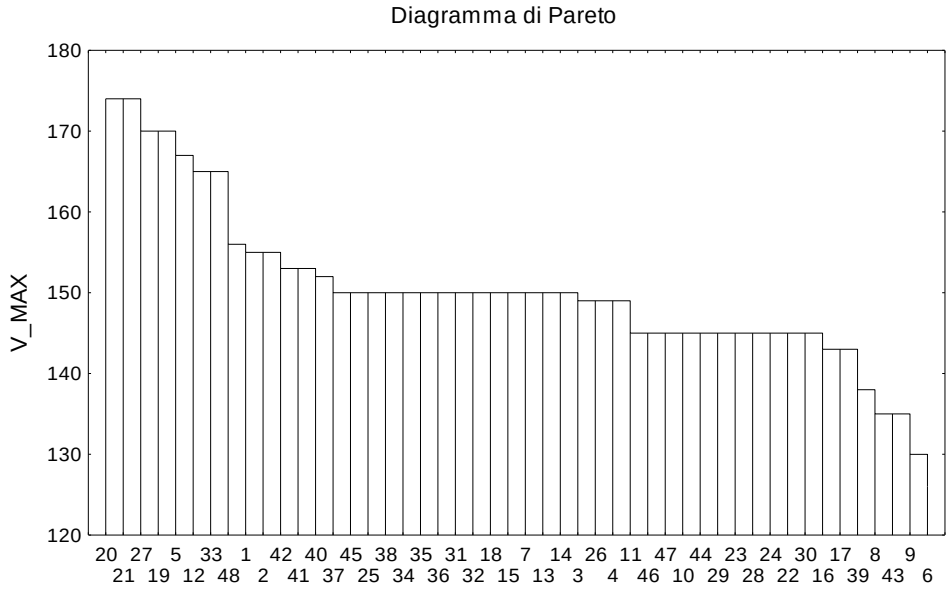


FIG. 11-1

In questo caso, assumendo come criteri la massima velocità e la massima percorrenza di chilometri con un litro di carburante, a velocità di 120 km/h, i punti ottimali corrispondono alle auto 34 e 20. Questi due punti sono detti tra loro incommensurabili, poichè sono in alternativa tra loro.
Per dimensionalità superiori, i diagrammi di Pareto non sono particolarmente utili, anche se l'utilizzo delle prime due componenti principali (v. analisi delle componenti principali, cap.3) può fornire ancora in diversi casi utili indicazioni. Tuttavia, l'utilizzo delle componenti principali non sembra particolarmente adatto in questo contesto.

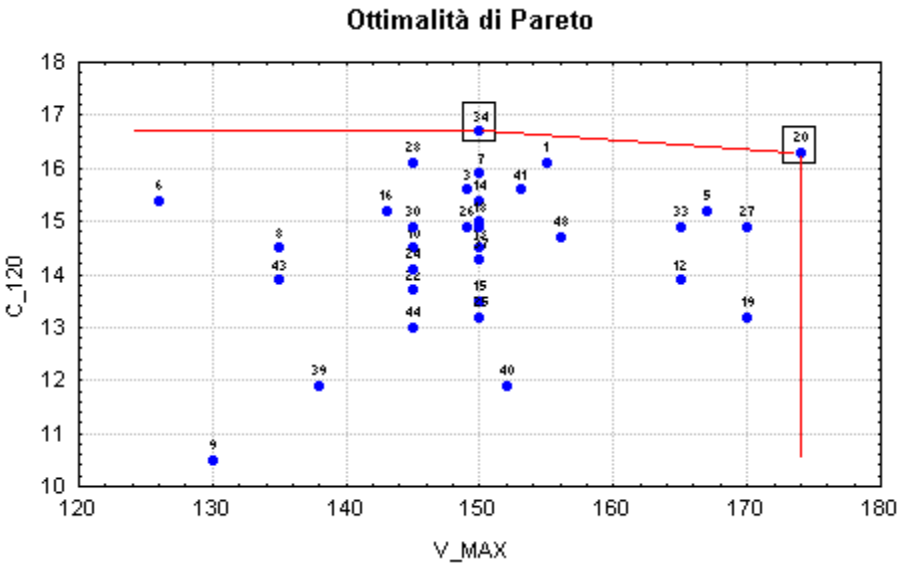


FIG. 11-2

11.3 - Le funzioni di desiderabilità

Le *funzioni di desiderabilità* d_k sono lo strumento fondamentale di una delle strategie decisionali multicriterio più note.
Questo tipo di approccio alle decisioni su criteri multipli si basa sulla definizione di una funzione di desiderabilità per ognuno dei criteri considerati. Queste funzioni possono essere dei tipi più diversi, anche se di norma la scelta avviene tra poche semplici funzioni (Fig. 11-3). Le funzioni di desiderabilità più comuni sono:

- lineare
- esponenziale
- logaritmica
- sigmoide
- a scalino
- triangolare
- normale

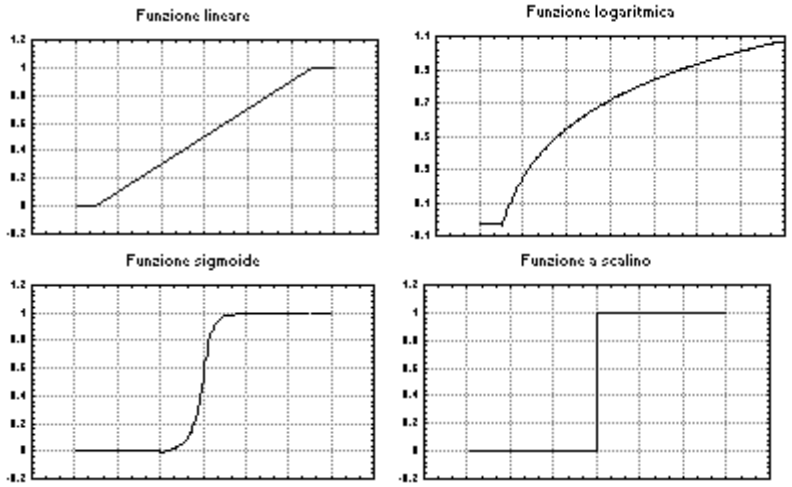
Qualunque sia la funzione prescelta, i valori della risposta - la desiderabilità - sono compresi tra zero (non accettabile) e uno (massima desiderabilità):

$$d_{ki} = f_k(y_{ki}) \qquad 0 \le d_{ki} \le 1$$

dove k è il criterio selezionato, f è il tipo di funzione prescelta e y_{ki} è il valore dell'i-esimo campione per il k-esimo criterio.
Ad esempio, se uno dei criteri selezionati è il prezzo dell'automobile, potremmo scegliere come funzione di desiderabilità una funzione lineare decrescente che assume i seguenti valori:

d = 1 se i prezzi sono minori o uguali a 10 milioni;
d = 0 se i prezzi sono maggiori o uguali a 25 milioni.
$0 < d < 1$ se i prezzi sono compresi tra 10 e 25 milioni.

Scegliendo un andamento lineare inverso, la desiderabilità d diminuisce linearmente, nell'intervallo considerato, all'aumentare del prezzo.



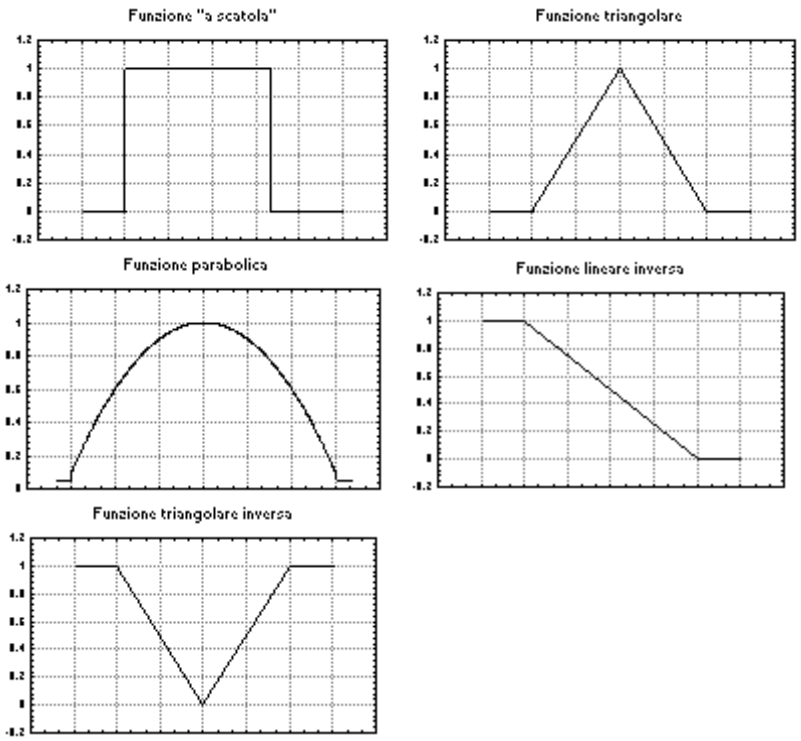


FIG. 11-3

Una volta definito il tipo di funzione e il suo andamento tra i limiti prefissati per tutti i criteri, cioè definite le singole funzioni di desiderabilità, possiamo valutare la funzione di desiderabilità globale *D* per l'*i*-esimo campione nel seguente modo:

$$D_i = \sqrt[k]{d_{1i} \cdot d_{2i} \cdot \ldots \cdot d_{ki}} \qquad 0 \le D_i \le 1$$

Dalla forma matematica di *D* (media geometrica delle singole funzioni di desiderabilità) si può osservare che la richiesta di desiderabilità è molto restrittiva poichè è *sufficiente* che una sola delle *k* funzioni di desiderabilità sia nulla per rendere nullo il valore di *D*. Viceversa, perchè il valore di *D* sia uguale a uno, è *necessario* che tutte le singole funzioni di desiderabilità siano massime (uguali a 1).

Calcolando il valore di *D* per tutti i campioni, siamo in grado di scegliere il campione corrispondente alla massima desiderabilità globale, se questa supera una certa soglia di accettabilità.
Comunemente le soglie di valutazione dei valori di *D* sono le seguenti (Tab.11-1):

valori di D	*valutazione della desiderabilità*
1.00 - 0.80	eccellente
0.80 - 0.63	buona
0.63 - 0.37	accettabile ma mediocre
0.37 - 0.20	limite di accettabilità
0.20 - 0.00	non accettabile

TAB. 11-1

Nel caso in cui si vogliano attribuire pesi w_k ai singoli criteri, cioè alle singole funzioni di desiderabilità, l'espressione per il calcolo di *D* diviene:

$$D_i = \left(d_{1i}^{w_1} \cdot d_{2i}^{w_2} \cdot \ldots \cdot d_{ki}^{w_k}\right)^{1/\Sigma w_k}$$

Per escludere una funzione come criterio di scelta, è sufficiente attribuirle un peso uguale a zero ($d^0 = 1$, indipendentemente dal valore di *d*).

11.4 - Le funzioni di utilità

Simile al criterio precedente delle funzioni di desiderabilità, le *funzioni di utilità* u_k hanno le stesse caratteristiche e la definizione di ciascuna di esse comporta la scelta a priori del tipo di andamento e dei limiti al di fuori dei quali la funzione è sempre minima o massima. Qualunque sia la funzione prescelta, i valori della risposta - l'utilità - sono compresi tra zero (non utile) e uno (massima utilità):

$$u_{ki} = f_k(y_{ki}) 0 \le u_{ki} \le 1$$

dove *k* è il criterio selezionato, *f* è il tipo di funzione prescelta e y_{ki} è il valore dell'*i*-esimo campione per il *k*-esimo criterio.
Ciò che cambia in questo caso è la valutazione globale, valutazione che viene effettuata secondo l'espressione:

$$U_i = \sum_k w_k u_{ki} \qquad 0 \le U_i \le 1$$

dove i pesi soddisfano la relazione $\sum_k w_k = 1$ e le funzioni u_{ki} sono scalate tra zero e uno. Le condizioni di ottimalità sono in questo caso meno restrittive di quelle relative alle funzioni di desiderabilità, in quanto, anche se per un criterio l'utilità è nulla, la funzione di utilità globale può ancora assumere valori alti per il contributo determinato dagli altri criteri.

11.5 - Le funzioni di dominanza

Il metodo della dominanza si basa sul confronto del comportamento dei diversi criteri per ogni coppia di campioni *i* e *j*. In questo caso, il metodo non richiede di stabilire nei termini di una funzione quantitativa l'andamento del criterio, ma solamente se il criterio considerato è ottimale per valori massimi o minimi. Per ogni coppia di campioni *i* e *j* viene calcolato il numero di criteri (k^+) per cui *i* è migliore di *j*, il numero di criteri per cui *i* e *j* sono uguali, ed il numero di criteri (k^-) per cui *i* è inferiore a *j*.
La *funzione di dominanza* tra i due campioni viene definita come:

$$C_{ij} = \frac{1 + \sum_{k^+} w_{k^+}}{1 + \sum_{k^-} w_{k^-}} \qquad 0.5 \le C_{ij} \le 2 \qquad k^+ + k^- = k$$

dove i pesi soddisfano la relazione $\sum_k w_k = 1$; la sommatoria al numeratore scorre su tutti i criteri ove *i* è migliore di *j* e quella al denominatore su tutti i criteri ove *i* è peggiore di *j* (i casi di parità vengono ignorati), tenendo conto del peso assegnato a ciascun criterio.

Un valore di $C_{ij} = 1$ indica parità di comportamento (cioè, $k^+ = k^-$, se tutti i pesi sono uguali); un valore di $C_{ij} > 1$ indica che nel complesso il campione *i* è

superiore al campione *j* ($k^+ > k^-$); un valore di $C_{ij} < 1$ indica che globalmente il campione *i* è inferiore al campione *j* ($k^+ < k^-$).
I valori ottenuti possono essere normalizzati secondo l'espressione:

$$C'_{ij} = \frac{C_{ij} - 0.5}{2 - 0.5}$$

Ad ogni *i*-esimo campione viene successivamente assegnato un punteggio globale dato da:

$$\phi_i = \sum_j C'_{ij}$$

Questo valore viene a sua volta normalizzato dividendo per il numero totale di confronti *n* - 1:

$$\phi'_i = \frac{\phi_i}{n-1}$$

I campioni con i valori maggiori di ϕ sono i punti ottimali di questo metodo, simile, nella proposta qui presentata, al metodo noto col nome di ELECTRE

11.6 - Le funzioni di preferenza

Questa metodologia richiede ancora la definizione di funzioni che rappresentano l'andamento voluto per ciascun criterio, come per le funzioni di desiderabilità e le funzioni di utilità. Tuttavia, in questo caso, l'andamento desiderato non modella direttamente i diversi valori che i campioni assumono per un dato criterio, ma modella la *differenza* tra i valori di coppie di campioni, in accordo con funzioni dette *funzioni di preferenza* $P_k(i,j)$.
Supponiamo che un criterio di scelta sia la resa percentuale di una reazione chimica, che avviene con diverse possibili modalità delle condizioni che controllano il processo. Questo significa, ad esempio, che se si richiede che la differenza d_{ij} di resa della reazione chimica tra l'*i*-esimo e il *j*-esimo gruppo di parametri che definiscono le condizioni di processo sia maggiore o uguale al 10% ($\delta_k = 10$), il valore della funzione di preferenza è uguale a uno se d_{ij} ¤ 10; se il valore d_{ij} è invece minore di zero o uguale a zero, il valore della funzione di preferenza è zero. In tutti gli altri casi il valore della funzione di preferenza è

dato dalla funzione stessa, il cui argomento è definito come $P_k(i,j) = f_k\left(d_{ij}/\delta_k\right)$.
Le tre differenti possibilità possono essere riassunte così:

$$P_k(i,j) = 0 \quad se \quad d_{ij} < 0$$
$$P_k(i,j) = 1 \quad se \quad d_{ij} \geq \delta_k$$
$$P_k(i,j) = f_k\left(d_{ij}/\delta_k\right) \quad se \quad 0 < d_{ij}/\delta_k < 1$$

La funzione *f* può essere una qualsiasi delle funzioni monotone crescenti o decrescenti definite precedentemente. Nel caso più semplice, la funzione viene definita come:

$$f_k = d_{ij}/\delta_k$$

Se la condizione di ottimalità è una funzione decrescente ($\delta_k < 0$), le precedenti condizioni divengono:

$$P_k(i,j) = 0 \quad se \quad d_{ij} > 0$$
$$P_k(i,j) = 1 \quad se \quad d_{ij} \leq \delta_k$$
$$P_k(i,j) = f_k\left(d_{ij}/\delta_k\right) \quad se \quad 0 < d_{ij}/\delta_k < 1$$

Per ogni coppia *i* e *j*, viene effettuato il confronto per tutti i criteri considerati. Definiamo quindi la seguente funzione :

$$\Pi(i,j) = \sum_k w_k \cdot P_k(i,j)$$

dove i pesi soddisfano la relazione $\sum_k w_k = 1$.
Per ogni campione *i* vengono quindi derivate le due seguenti funzioni:

$$\phi_i^+ = \sum_j \Pi(i,j) \qquad \phi_i^- = \sum_j \Pi(j,i)$$

definite come *positive flow outranking* e *negative flow outranking*. La prima sommatoria scorre su tutti i casi favorevoli a *i*; la seconda scorre su tutti i casi sfavorevoli a *i*. Queste rappresentano, rispettivamente, i successi e gli insuccessi di *i* con tutti gli altri campioni.

Per ogni campione *i* viene quindi valutata una funzione globale definita come:

$$\phi_i = \phi_i^+ - \phi_i^-$$

detta *net flow outranking.*
I valori calcolati possono essere normalizzati secondo l'espressione:

$$\phi_i' = \frac{\phi_i + (n-1)}{(n-1)+(n-1)} = \frac{\phi_i + n - 1}{2 \cdot (n-1)}$$

essendo *n* il numero totale di oggetti e +(*n* - 1) e -(*n* - 1), rispettivamente, il massimo ed il minimo valore di ϕ_i.
I campioni con i valori maggiori di ϕ sono i punti ottimali di questo metodo, noto anche come PROMETHEE.
Per quanto riguarda l'argomento d_{ij} / δ_k della funzione di preferenza $P_k(i,j) = f_k(d_{ij}/\delta_k)$, si può osservare che assumendo valori di δ_k maggiori della massima differenza tra i campioni (d_k^{max}), si verifica un appiattimento dei risultati, poichè nessun campione potrà soddisfare la richiesta di massima ottimalità.
Al contrario, valori di δ_k minori della minima differenza tra i campioni (d_k^{min}), qualora f_k siano funzioni lineari, portano ad un risultato identico a quello che si ottiene nel caso del criterio delle funzioni di dominanza. Infatti, in questo caso, o $P_k(i,j) = 0$ perchè $d_{ij} \leq 0$ oppure $P_k(i,j) = 1$ perchè $d_{ij} > \delta_k$.

11.7 - Il metodo della minima distanza

Questo metodo per le decisioni multicriterio si basa sull'idea semplice di misurare la distanza tra ogni campione e un *campione ideale* che si ritiene rappresentare la situazione ottimale sulla base di tutti i criteri considerati. L'applicazione di questo metodo richiede che per ogni criterio sia noto il valore ottimale e non solo, come accade per altri metodi, se un dato criterio deve essere minimo o massimo. Questa informazione è di fatto contenuta nei valori dei criteri che definiscono il campione ideale.

Una volta stabilita una misura di distanza, vengono calcolate tutte le *n* distanze tra i campioni e il riferimento.

I campioni vengono ordinati in modo crescente in funzione della loro distanza dal campione ideale, cioè in funzione della loro decrescente similarità con quest'ultimo. Questa tecnica coincide di fatto con l'analisi di similarità rispetto ad un riferimento, definita nel capitolo 4 nei metodi di *cluster analysis*.

ESEMPIO 1

Per una semplice applicazione dei metodi di decisione su criteri multipli si sono considerati i dati AUTO. I dati sono stati rilevati da una rivista specializzata (maggio 1995). Sono stati presi in considerazione i modelli di auto di piccola cilindrata, descritti dalla potenza (hp), dalla velocità massima (vmax), dal consumo urbano (curb) e a 10 km (c120), entambi in l/100 km, una valutazione del comfort (comf) e della sicurezza (sicur), il costo di esercizio (cese) e il costo totale (ctot).
In tutte le graduatorie, l'auto 9 è sempre all'ultimo posto. Questo effetto è dovuto al fatto che la caratteristica fondamentale di questa auto, cioè le 4 ruote motrici, comporta un cospicuo incremento del costo totale, ma è d'altra parte evidente che non è stato preso in considerazione nessun criterio in grado di apprezzare questa caratteristica.

Funzioni di desiderabilità

Le funzioni di desiderabilità sono state definite per 8 criteri, assumendo lo stesso peso per tutti i criteri, secondo le modalità definite in Tab.11-2:

ID	*criterio*	*peso*	*funzione*	*minimo*	*massimo*
1	hp	0.125	Lineare	25	100
2	vmax	0.125	Sigmoide	120	180
3	curb	0.125	Lineare	5	20
4	c120	0.125	Lineare	5	20
5	comf	0.125	Lineare	4	8
6	sicur	0.125	Lineare	4	7
7	cese	0.125	Lineare inversa	100	200
8	ctot	0.125	Lineare inversa	10000	25000

TAB. 11-2

Ord.	*Auto*	*punteggio*	*ID*
1	NISS-Micra_3P	0.536	26
2	HYUN-Accent_LS3P	0.536	20
3	ROVE-111_Si5P	0.533	42
4	RENL-Clio_RL5P	0.524	35
5	RENL-Clio_RTI5P	0.523	36
6	NISS-Micra_SLX5P	0.522	27
7	CITR-AX_I5P	0.516	4
8	RENL-Clio_RL3P	0.513	34
9	ROVE-111_Si3P	0.512	41
10	FORD-Fiesta_Caym5P	0.511	17
11	FIAT-Punto_S55_5P	0.509	14
12	CITR-AX_ITZX_5P	0.502	5
13	FORD-Fiesta_Caym3P	0.500	16
14	PEUG-106_XN5P	0.499	32
15	RENL-Twingo	0.499	37
16	FORD-Fiesta_Wind5P	0.498	18
17	CITR-AX_I3P	0.497	3
18	OPEL-Corsa_City5P	0.495	29
19	AUBI-Y10_Igloo	0.491	2
20	FIAT-Punto_6_Speed	0.487	13
21	RENL-Twingo_Spring	0.487	38
22	OPEL-Corsa_City3P	0.484	28
23	HYUN-Accent_HS5P	0.482	21
24	PEUG-106_Holl3P	0.481	31
25	SKOD-Felicia_GLXi	0.476	45
26	FIAT-Uno_Start5P	0.472	11
27	OPEL-Corsa_Swing5P	0.462	30
28	VOLK-Polo_1.3_5P	0.462	48
29	MAZD-LX4P	0.462	24
30	VOLK-Polo_1.0_5P	0.461	47
31	AUBI-Y10_Junior	0.459	1
32	PEUG-106_XT5P	0.458	33
33	MAZD-Cabrio_LX4P	0.458	25
34	FIAT-500Sporting	0.456	7
35	FIAT-Uno_Start3p	0.455	10
36	INNO-Mille_5P	0.454	23
37	FIAT-Punto_60_Sel5P	0.454	15
38	VOLK-Polo_1.0_3P	0.450	46
39	FIAT-Uno_Cond5P	0.443	12
40	INNO-Mille_3P	0.436	22
41	SKOD-Felicia_LX	0.431	44
42	ROVE-Mini_Cooper	0.375	40
43	HOND-Civic_EX3P	0.369	19
44	SEAT-Ibiza_Cli3P	0.367	43
45	FIAT-Panda_Young	0.347	8
46	ROVE-Mini_B_Open	0.326	39
47	FIAT-500_700ED	0.249	6
48	FIAT-Panda:C_Club	0.217	9

TAB. 11-3

Come si può osservare, sono stati scelti in tutti i casi degli andamenti lineari; unica eccezione, la velocità massima (vmax) per la quale si è scelta una funzione sigmoidale centrata su 150 km/h. In questo modo, si sono maggiormente penalizzate le auto con velocità inferiori e maggiormente apprezzate le auto con velocità superiori.
I risultati ottenuti sono presentati in Tab. 11-3.

Funzioni di dominanza

Anche per il criterio basato sulla dominanza sono stati selezionati gli stessi 8 criteri definiti in precedenza. In questo caso, non è necessario definire l'andamento delle funzioni, ma solamente la condizione di ottimalità, cioè se questa è un massimo o un minimo. Anche in questo caso, i pesi sono stati mantenuti tutti uguali. La Tab.11-4 riporta le definizioni dei criteri selezionati.

ID	*criterio*	*ottimalità*	*peso*
1	hp	massimo	0.125
2	vmax	massimo	0.125
3	curb	massimo	0.125
4	c120	massimo	0.125
5	comf	massimo	0.125
6	sicur	massimo	0.125
7	cese	minimo	0.125
8	ctot	minimo	0.125

TAB. 11-4

La Tab.11-5 riporta i risultati ottenuti secondo questo criterio.
Si può osservare, ad esempio, che l'auto 6 viene molto penalizzata con questo criterio poichè potenza e velocità massima rappresentano due minimi di desiderabilità, minimi che sono sufficienti ad abbattere la desiderabilità complessiva. Nei due metodi di dominanza e di *outranking*, basati sulle differenze, questo aspetto è molto meno accentuato.

Ord.	*Auto*	*punteggio*	*ID*
1	HYUN-Accent_HS5P	0.588	21
2	ROVE-111_Si5P	0.541	42
3	ROVE-111_Si3P	0.527	41
4	HYUN-Accent_LS3P	0.509	20
5	RENL-Clio_RTI5P	0.501	36
6	NISS-Micra_3P	0.489	26
7	NISS-Micra_SLX5P	0.477	27
8	AUBI-Y10_Junior	0.470	1
9	CITR-AX_ITZX_5P	0.466	5
10	RENL-Clio_RL5P	0.451	35
11	CITR-AX_I5P	0.449	4
12	CITR-AX_I3P	0.448	3
13	AUBI-Y10_Igloo	0.446	2
14	RENL-Clio_RL3P	0.442	34
15	FIAT-500Sporting	0.419	7
16	FORD-Fiesta_Caym5P	0.415	17
17	PEUG-106_Holl3P	0.401	31
18	FORD-Fiesta_Caym3P	0.401	16
19	OPEL-Corsa_City5P	0.397	29
20	PEUG-106_XN5P	0.397	32
21	RENL-Twingo	0.394	37
22	OPEL-Corsa_City3P	0.382	28
23	PEUG-106_XT5P	0.382	33
24	FORD-Fiesta_Wind5P	0.371	18
25	FIAT-Punto_S55_5P	0.361	14
26	FIAT-500_700ED	0.358	6
27	FIAT-Punto_6_Speed	0.356	13
28	RENL-Twingo_Spring	0.342	38
29	FIAT-Panda_Young	0.323	8
30	MAZD-Cabrio_LX4P	0.316	25
31	VOLK-Polo_1.3_5P	0.314	48
32	SKOD-Felicia_GLXi	0.310	45
33	FIAT-Punto_60_Sel5P	0.303	15
34	FIAT-Uno_Cond5P	0.291	12
35	FIAT-Uno_Start5P	0.282	11
36	HOND-Civic_EX3P	0.281	19
37	FIAT-Uno_Start3p	0.273	10
38	OPEL-Corsa_Swing5P	0.260	30
39	MAZD-LX4P	0.258	24
40	VOLK-Polo_1.0_5P	0.248	47
41	VOLK-Polo_1.0_3P	0.234	46
42	ROVE-Mini_Cooper	0.234	40
43	SKOD-Felicia_LX	0.226	44
44	INNO-Mille_5P	0.220	23
45	INNO-Mille_3P	0.199	22
46	SEAT-Ibiza_Cli3P	0.183	43
47	ROVE-Mini_B_Open	0.155	39
48	FIAT-Panda:C_Club	0.064	9

TAB. 11-5

Funzioni di preferenza

Anche in questo caso si sono selezionati gli stessi 8 criteri precedenti. Per l'andamento della funzione di preferenza, si scelto sempre un andamento lineare; gli intervalli oltre i quali considerare importante la differenza tra due campioni sono riportati nella Tab.11-6. Anche in questo caso, i pesi assegnati ai criteri sono tutti uguali.

ID	*criterio*	*peso*	*funzione*	*intervallo*
1	hp	0.125	Lineare	10
2	vmax	0.125	Lineare	20
3	curb	0.125	Lineare	1
4	c120	0.125	Lineare	1
5	comf	0.125	Lineare	.5
6	sicur	0.125	Lineare	.5
7	cese	0.125	Lineare inversa	10
8	ctot	0.125	Lineare inversa	2000

TAB. 11-6

ESEMPIO 2

I dati sono relativi all'ottimizzazione di un processo di rivestimento mediante materiale in grani. Sono state misurate 6 diverse caratteristiche di questi grani provenienti da 14 prove (definite da un disegno sperimentale centrato) ottenuto per diversi valori di 3 parametri indipendenti: temperatura, velocità di spray e pressione dell'aria di atomizzazione.

Le 6 caratteristiche misurate e le condizioni richieste sono (Tab. 11.8).
I risultati ottenuti con i diversi metodi sono i seguenti (Tab. 11-9).

Ord.	*Auto*	*punteggio*	*ID*
1	HYUN-Accent_HS5P	0.717	21
2	HYUN-Accent_LS3P	0.649	20
3	ROVE-111_Si5P	0.632	42
4	RENL-Clio_RTI5P	0.629	36
5	NISS-Micra_SLX5P	0.629	27
6	NISS-Micra_3P	0.620	26
7	ROVE-111_Si3P	0.613	41
8	CITR-AX_I5P	0.602	4
9	FORD-Fiesta_Caym5P	0.598	17
10	CITR-AX_ITZX_5P	0.589	5
11	CITR-AX_I3P	0.589	3
12	RENL-Clio_RL5P	0.587	35
13	FORD-Fiesta_Caym3P	0.581	16
14	RENL-Clio_RL3P	0.570	34
15	AUBI-Y10_Junior	0.567	1
16	OPEL-Corsa_City5P	0.567	29
17	AUBI-Y10_Igloo	0.555	2
18	OPEL-Corsa_City3P	0.548	28
19	FIAT-500Sporting	0.543	7
20	PEUG-106_XN5P	0.533	32
21	PEUG-106_Holl3P	0.522	31
22	FIAT-500_700ED	0.521	6
23	FIAT-Punto_S55_5P	0.520	14
24	FORD-Fiesta_Wind5P	0.512	18
25	PEUG-106_XT5P	0.511	33
26	RENL-Twingo	0.504	37
27	FIAT-Punto_6_Speed	0.499	13
28	FIAT-Panda_Young	0.488	8
29	RENL-Twingo_Spring	0.468	38
30	MAZD-Cabrio_LX4P	0.460	25
31	FIAT-Uno_Start5P	0.458	11
32	SKOD-Felicia_GLXi	0.451	45
33	FIAT-Uno_Cond5P	0.448	12
34	FIAT-Punto_60_Sel5P	0.446	15
35	HOND-Civic_EX3P	0.444	19
36	FIAT-Uno_Start3p	0.444	10
37	OPEL-Corsa_Swing5P	0.437	30
38	VOLK-Polo_1.3_5P	0.436	48
39	MAZD-LX4P	0.424	24
40	VOLK-Polo_1.0_5P	0.421	47
41	VOLK-Polo_1.0_3P	0.401	46
42	SKOD-Felicia_LX	0.387	44
43	INNO-Mille_5P	0.385	23
44	INNO-Mille_3P	0.364	22
45	ROVE-Mini_Cooper	0.358	40
46	SEAT-Ibiza_Cli3P	0.339	43
47	ROVE-Mini_B_Open	0.290	39
48	FIAT-Panda:C_Club	0.148	9

TAB. 11-7

Per quanto riguarda l'ordinamento dei 14 casi ottenuto con le funzioni di desiderabilità, si può osservare che i casi 1, 3, 5, 12 e 14 hanno punteggio zero perchè per ciascuno di essi esiste almeno un criterio che ha un punteggio zero: per i primi 3, il criterio è la percentuale di sostanza il cui limite inferiore di accettabilità è stato fissato al 13% (caso 1 = 12.8%; caso 3 = 12.8%; caso 5 = 12.6%); per i due restanti, il criterio non soddisfatto è la densità d'ingombro minima, fissata a 0.6 (caso 12 = 0.6; caso 14 = 0.6).
Nel caso in cui volessimo estendere l'intervallo di accettazione per quanto riguarda la percentuale di sostanza spostando i limiti da 13-14 a 12-15 (risultati non riportati in tabella), i casi 1, 3 e 5 (con valori di 12.8%, 12.8% e 12.6%, rispettivamente) non sarebbero esclusi da questo criterio. Lasciando immutate tutte le restanti condizioni, il caso 5 si inserirebbe al terzo posto, dopo il caso 10; il caso 1 si inserirebbe immediatamente dopo il caso 5 (al quarto posto), il caso 3 si inserirebbe dopo il caso 11 (al settimo posto). L'ordinamento di tutti gli altri casi rimane invariato, così come rimarrebbero con punteggio zero i casi 12 e 14 che non soddisfano il secondo criterio.

Proprietà	*min*	*max*	*ottimo*	*limite inf*	*limite sup*	*funzione*
uniformità di contenuto	0.98	3.31	< 1	1	5	lineare inv.
densità d'ingombro $\left(g/cm^3\right)$	0.60	0.67	> 0.7	0.6	0.7	lineare
densità sfruttata $\left(g/cm^3\right)$	0.68	0.78	> 0.8	0.0	0.8	lineare
dimensione particelle (μm)	94.9	139.4	< 95	95	145	lineare inv.
dissoluzione in acido (%/h)	25.2	74.1	> 80 %	0.0	80	sigmoide
percentuale di sostanza (%)	12.6	13.7	= 13.5	13	14	triangolare

TAB. 11-8

#	ID	Desid. F.	ID	Utility F.	ID	Pref. F.	ID	Dom. F.
1	13	0.835	13	0.856	5	0.621	9	0.671
2	10	0.822	10	0.833	3	0.610	5	0.612
3	9	0.749	9	0.796	10	0.539	3	0.551
4	11	0.747	11	0.781	6	0.532	13	0.538
5	7	0.713	7	0.766	11	0.518	10	0.513
6	6	0.694	2	0.764	13	0.499	1	0.510
7	2	0.678	6	0.740	1	0.490	2	0.428
8	8	0.547	12	0.726	7	0.482	11	0.331
9	4	0.283	5	0.714	8	0.477	12	0.315
10	1	0	8	0.702	14	0.468	7	0.255
11	3	0	1	0.680	4	0.461	6	0.253
12	5	0	3	0.629	2	0.461	8	0.233
13	12	0	14	0.595	9	0.461	4	0.137
14	14	0	4	0.471	12	0.382	14	0.090

TAB. 11-9

I precedenti vincoli influenzano negativamente anche i risultati ottenuti con le funzioni di utilità, dove, tuttavia, trattandosi di una somma di contributi, il risultato finale anche per i casi ora considerati è diverso da zero. Anche con questo metodo si conferma la buona qualità dei casi 13, 10, 9 e 11, già evidenziata con le funzioni di desiderabilità.
Per quanto riguarda le funzioni di preferenza e di dominanza, si sono presi in considerazione solo i primi 5 criteri che hanno tutti caratteristiche di ottimalità secondo funzioni monotone crescenti o decrescenti. Inoltre, per quanto riguarda le funzioni di preferenza, l'intervallo richiesto per ciascun criterio è stato definito come la metà della differenza tra il limite superiore e il limite inferiore (ad esempio, per l'uniformità di contenuto è stato selezionata una differenza di 2 ottenuta da (5 -1)/2).
Per l'ordinamento ottenuto con il metodo delle funzioni di preferenza, dove la valutazione avviene mediante il confronto di ogni caso con tutti gli altri, indipendentemente dai valori assoluti dei singoli criteri, si osserva che i casi 5 e 3, penalizzati con i precedenti due metodi, risultano in questo caso ai primi 2 posti. Questo metodo conferma inoltre la buona qualità dei casi 10, 11 e 13, mentre svaluta notevolmente il caso 9.
L'ordinamento ottenuto col criterio delle funzioni di dominanza è ovviamente più simile ai risultati ottenuti con le funzioni di preferenza. Si osservi, tuttavia,

che il caso 9, molto penalizzato con le funzioni di preferenza, è in questo caso al primo posto. Analoga invece la rivalutazione di quest'ultimo metodo dei casi 3 e 5 rispetto ai primi due metodi considerati.

🕮 BIBLIOGRAFIA

H.R.KELLER, D.L.MASSART E J.P.BRANS (1991). Multicriteria decision making: a case study. *Chemometrics and Intelligent Laboratory Systems*, **11**, 175-189.

P.J. LEWI, J. VAN HOOF E P. BOEY (1991). Multicriteria decision making using Pareto optimality and PROMETHEE preference ranking. *Chemometrics and Intelligent Laboratory Systems*, **16**, 139-144.

M.M.W.B. HENDRIKS, J.H. DE BOER, A.K. SMILDE E D.A. DOORNBOS (1992). Multicriteria decision making. *Chemometrics and Intelligent Laboratory Systems*, **16**, 175-191.

12

LE STRATEGIE QSAR

12.1 - Introduzione

Lo studio delle relazioni tra struttura molecolare e attività di un composto si fonda su un approccio razionale basato sull'assunzione che esistano certe relazioni tra la struttura molecolare (S) e l'attività biologica (A) dei composti. In altre parole, si cerca di determinare la relazione funzionale f(S, A) che mette in relazione l'attività A con la struttura molecolare S di un composto.
L'obiettivo finale generale degli studi di relazioni struttura-attività (***SAR***, ***Structure-Activity Relationships***) è quello di comprendere i meccanismi dell'azione farmacologica o tossicologica, suggerendo vie nuove per la sintesi di composti con attività biologica definita.
Le interazioni farmaco-organismo rappresentano un sistema complesso che coinvolge un grande numero di processi molti dei quali sono sconosciuti. L'effetto globale dei singoli processi può essere visto come il manifestarsi di un'attività del composto: questo aspetto è la causa della natura statistica delle strategie *SAR*.
Le assunzioni fondamentali degli studi *SAR* sono:

a) a composti simili corrispondono simili proprietà biologiche

b) a modifiche simili nella struttura molecolare corrispondono cambiamenti simili delle proprietà.

Sugli stessi principi generali si basano anche le strategie per lo studio delle relazioni tra struttura molecolare e proprietà chimico-fisiche, note come strategie ***SPR*** (***Structure-Property Relationships***).
La Fig.12-1 rappresenta lo schema generale entro cui si possono inquadrare gli studi delle relazioni attività/proprietà e struttura molecolare. Le funzioni α e μ rappresentano, rispettivamente, le *procedure sperimentali* mediante le quali

determiniamo le proprietà biologiche, farmacologiche o tossicologiche e le proprietà chimico-fisiche di un insieme di composti C:

$$\alpha: C \rightarrow A \qquad \mu: C \rightarrow M$$

ovvero

$$\alpha(C) = A \qquad \mu(C) = M$$

L'insieme M è costituito da proprietà chimico-fisiche quali, ad esempio, il punto di ebollizione, il punto di fusione, il volume molare, il momento dipolare, la solubilità in acqua, i coefficienti di ripartizione ottanolo/acqua, aria/acqua, sedimento/acqua, i parametri di reattività chimica, ecc. L'insieme A è a sua volta costituito da quantità che sono legate all'attività biologica, farmacologica, tossicologica dei composti studiati. Le attività biologiche esplicano le loro attività con effetti e a livelli molto diversi tra loro. Si possono, ad esempio, distinguere:

Livello macromolecolare:
forze di legame col recettore, costanti di inibizione, costanti di Michaelis

Livello cellulare:
mutagenicità, trasformazioni cellulari

Livello di organismi (effetti acuti):
risposta bologica di alghe, invertebrati, pesci, uccelli, mammiferi

Livello di organismi (effetti cronici):
neurotossicità ritardata, bioconcentrazione, biodegradazione, carcinogenicità, tossicità sulla funzionalità riproduttiva.

La funzione γ_3 rappresenta l'insieme dei modelli con cui siamo in grado di calcolare le attività biologiche da proprietà chimico-fisiche dei composti:

$$\gamma_3(M) = A$$

Per molto tempo la strategia QSAR più frequente (prima dell'attuale riconosciuta importanza dei descrittori molecolari teorici) è stata improntata alla

ricerca di una correlazione diretta tra proprietà chimico-fische sperimentali e misure sperimentali di attività biologiche:

$$\gamma_3\mu(C) = \alpha(C)$$

In questo caso, evidentemente, non si fa ricorso ai descrittori di struttura molecolare e quindi è sempre assente ogni relazione funzionale sulla struttura molecolare.
L'utilizzo della funzione γ_1 riflette le strategie secondo le quali si vuole predire una proprietà chimico-fisica da altre proprietà chimico-fisiche:

$$\gamma_1(M) = M'$$

dove M' rappresenta una proprietà in M, diversa dalle proprietà in M utilizzate come descrittori in γ_1.

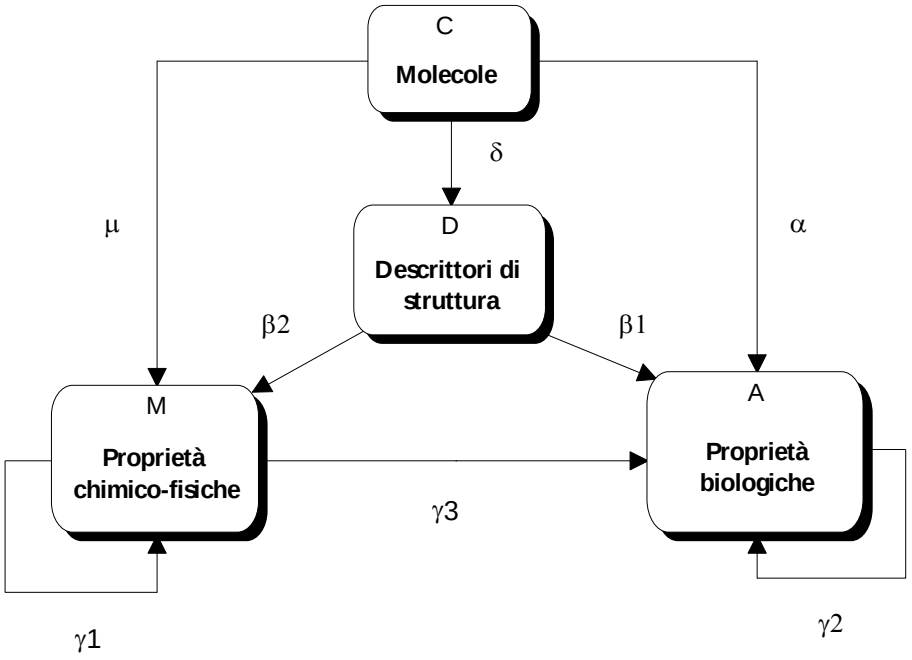


FIG. 12-1

Quest'ultimo aspetto è di notevole interesse quando la proprietà che si vuole predire è, in qualche senso, meno conveniente delle proprietà utilizzate per predire la prima (cioè la proprietà misurata è affetta da elevato rumore sperimentale, richiede una procedura sperimentale complessa, costosa, lunga; richiede costi alti per la sua determinazione; non è accessibile per alcune condizioni sperimentali).
In modo del tutto analogo, si opera con la funzione γ_2 quando si vuole predire, ad esempio, un'attività biologica da un'altra attività biologica, quest'ultima determinabile in modo più conveniente.

$$\gamma_2(\mathrm{A}) = \mathrm{A}'$$

In questo caso, di norma, i modelli di interesse sono semplici relazioni bivariate tra le due grandezze considerate, cioè l'attenzione è incentrata sulla loro correlazione.
La classe più importante delle strategie *SAR* e *SPR* si basa sulla creazione di funzioni δ in grado di trasformare l'informazione chimica contenuta nelle formule di struttura, nella topologia molecolare, nelle rappresentazioni 3D delle molecole, cioè nei termini delle loro coordinate spaziali x,y,z, in descrittori di struttura (D):

$$\delta(\mathrm{C}) = \mathrm{D}$$

Una volta ottenuti i descrittori di struttura, la seconda fase delle strategie *SAR/SPR* si basa sulla ricerca di modelli matematici capaci di predire da questi le attività biologiche (A) e/o le proprietà chimico-fisiche (M), cioè

$$\beta_1(\mathrm{D}) = \mathrm{A} \qquad \beta_2(\mathrm{D}) = \mathrm{M}$$

Quando le differenze

$$\alpha(\mathrm{C}) - \beta_1\delta(\mathrm{C}) \quad \text{e} \quad \mu(\mathrm{C}) - \beta_2\delta(\mathrm{C})$$

sono dell'ordine di grandezza dell'errore sperimentale, si perviene a modelli adeguati struttura-attività (nel primo caso) e struttura-proprietà (nel secondo caso).
Naturalmente è possibile predire l'attività biologica anche seguendo la via definita dalla relazione

$$\alpha(\mathrm{C}) - \gamma_3\beta_2\delta(\mathrm{C})$$

per cui l'attività biologica viene predetta dalle proprietà chimico-fisiche calcolate dai descrittori strutturali.

12.2 - L'approccio di Hansch ai problemi QSAR

Sebbene le prime relazioni tra le proprietà chimico-fisiche delle molecole e la loro attività biologica si debbano far risalire ai lavori di Meyer (1899) e di Overton (1901), il grande sviluppo di questo campo deriva dalle ricerche sviluppate da Hansch e collaboratori a partire dal 1963. Il postulato sul quale si basa l'approccio di Hansch afferma che quando una sostanza biologicamente attiva entra in contatto con il sistema molecolare di un organismo vivente, la probabilità che essa raggiunga un sito recettore e induca quindi una determinata risposta sull'organismo - l'effetto - è funzione delle proprietà della sostanza stessa.
Il postulato può essere schematizzato come segue:

sistema biologico + sostanza attiva = risposta

ovvero

$$risposta \quad = \quad f_1(L) + f_2(E) + f_3(S) + f_4(M)$$

dove le quattro funzioni sono rispettivamente funzioni di **proprietà lipofile** (L), di **proprietà elettroniche** (E), di **proprietà steriche** (S) e di eventuali **altre proprietà molecolari** (M) necessarie per una completa descrizione dell'effetto biologico considerato. Tutte queste proprietà sono le *proprietà della molecola* considerata e *la risposta dipende additivamente* da esse.
Una volta note le diverse funzioni, l'equazione precedente ci permette di valutare quantitativamente la relazione tra la risposta biologica e la struttura molecolare, utilizzando metodi di regressione multivariata.
La validità di questo postulato è indissolubilmente legata ad un altro postulato, noto come **principio di congenericità**. Secondo questo principio, le strategie *QSAR* sono applicabili soltanto a classi di composti "simili", ove per composti simili si intendono:

a) composti che abbiano uno scheletro-base comune

b) i sostituenti dello scheletro-base differiscano tra loro in modo da non influenzare in modo decisivo le proprietà globali della molecola.

In base ai principi su cui si basa l'approccio di Hansch, lo sviluppo di questa strategia si è decisamente indirizzato verso la parametrizzazione delle proprietà dei gruppi sostituenti (gruppi funzionali, frammenti molecolari), piuttosto che sulle misure delle proprietà di tutta la molecola. Questa impostazione, una volta note le proprietà dei gruppi funzionali, consente un'ampia e facile applicabilità a moltissimi problemi *QSAR*. Tuttavia, appaiono anche evidenti e indiscutibili i limiti di questo approccio che presume la possibilità di modellare le risposte biologiche mediante uno schema lineare puramente additivo utilizzando solo l'informazione locale insita nei gruppi sostituenti, indipendente quindi dalle proprietà globali di ciascuna molecola.

12.3 - Le strategie QSAR

Per lo studio delle relazioni tra struttura molecolare e attività sono state proposte molte differenti metodologie che si distinguono tra loro sia per l'impostazione generale nell'approccio al problema, sia per le modalità con cui sono descritte le molecole (diverse tipologie di descrittori), sia per i metodi matematici utilizzati per estrarre l'informazione necessaria.
In Fig. 12-2, è riportato uno schema generale che rappresenta la strategia di lavoro tipica nella ricerca delle relazioni attività-struttura molecolare. Nella figura sono anche evidenziati i due momenti riguardanti la costruzione del modello (*fitting*) e la fase di utilizzo del modello per fini predittivi (predizione).
Le principali strategie *QSAR* sono brevemente descritte qui di seguito.

🗀 *Relazioni Quantitative Struttura-Attività (QSAR)*

Gli studi *QSAR* (***Quantitative Structure-Activity Relationships***) sono basati su una serie di approcci matematici e statistici con lo scopo di trovare *modelli quantitativi* nelle relazioni tra struttura molecolare e attività.
Le strategie principali si fondano comunemente sulle seguenti fasi:

a) una rappresentazione delle strutture molecolari mediante opportuni descrittori.

b) la ricerca di relazioni quantitative specifiche tra descrittori ed attività (biologica, farmacologica, tossicologica) utilizzando principalmente l'analisi statistica multivariata e i metodi chemiometrici.

c) la predizione dell'attività di nuovi composti con una struttura predefinita utilizzando i modelli matematici trovati.

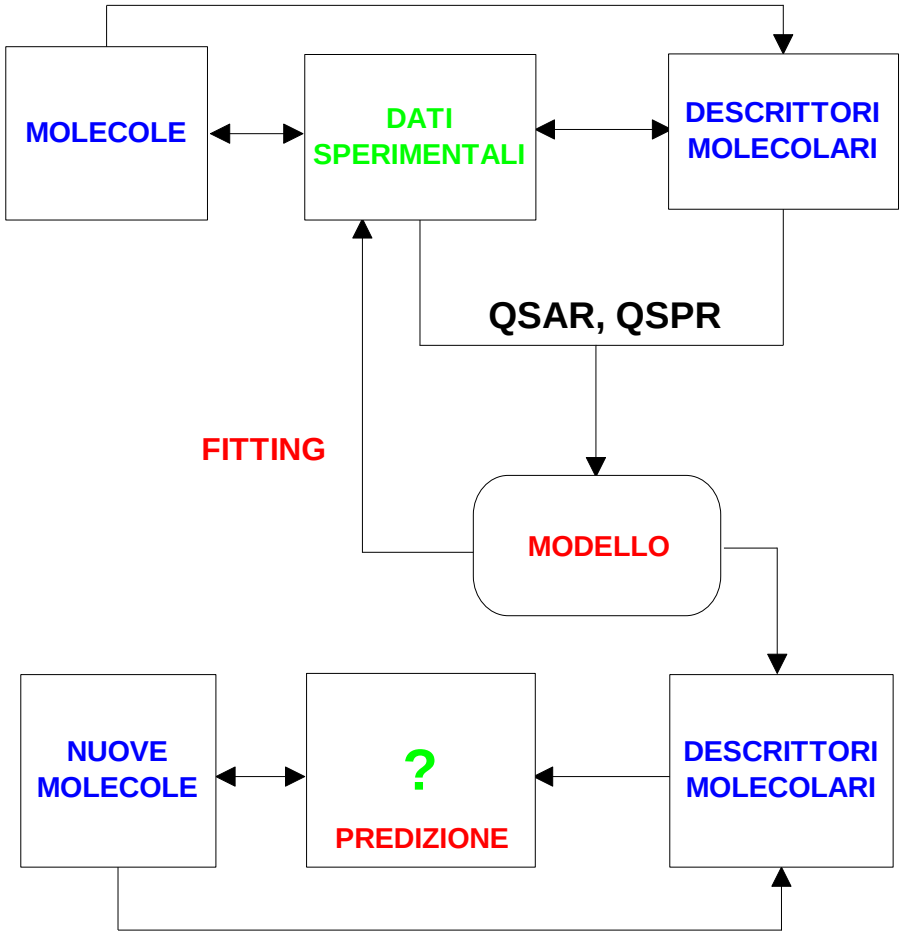


FIG. 12-2

La teoria dei grafi, i metodi della chimica teorica (ad esempio, l'analisi conformazionale), i metodi chemiometrici ed i metodi di *molecular modelling*

sono strumenti importanti e comunemente utilizzati negli studi *QSAR*. Questo tipo di strategie viene applicato soprattutto a problemi chimici, farmacologici, ambientali, tossicologici.
Il termine *QSAR* viene sovente utilizzato in senso più generale per denotare tutte le strategie di ricerca delle relazioni attività-struttura.

🗀 *Relazioni Quantitative Struttura-Proprietà (QSPR)*

Simile alle metodologie proprie dei problemi *QSAR*, nelle relazioni tra struttura molecolare e proprietà del composto (***Quantitative Structure-Property Relationships***) l'accento viene soprattuto posto sulle proprietà chimico-fisiche intrinseche del composto, quali, ad esempio, il punto di fusione, il punto di ebollizione, l'idrofobicità, ecc. In diversi casi le proprietà studiate sono rappresentate come variabili categoriche e quindi l'attenzione viene in questi casi spostata dai modelli di regressione ai modelli di classificazione, dove le classi indicano differenti livelli della proprietà studiata.

🗀 *Linear Free Energy Relationships (LFER)*

Un dizione datata che si riferisce alle relazioni quantitative tra struttura e attività (*QSAR*) quando l'obiettivo principale era quello di determinare correlazioni e modelli di regressione capaci di rendere conto di piccoli cambiamenti delle risposte in classi di reazioni simili. *LFER* è quindi un approccio quantitativo sviluppato soprattutto nel contesto delle chimica organica, basato sul principio che sostanze simili dovrebbero reagire in modo simile (a parità di condizioni di reazione) e che variazioni simili nella loro struttura dovrebbero produrre cambiamenti simili nella reattività. Molte di queste metodologie fanno ricorso ad una descrizione molecolare basata sui frammenti molecolari o sui sostituenti.
L'equazione di Hammett è l'esempio più tipico di questo approccio alle relazioni struttura-reattività:

$$log\left(\frac{K_X}{K_H}\right) = \rho \cdot \sigma$$

dove *K* sono velocità di reazione o costanti di equilibrio relative al sostituente X e al riferimento non sostituito H, ρ è un parametro riguardante le condizioni sperimentali e σ è un parametro legato all'effetto elettronico di un sostituente.

🗀 *Relazioni Quantitative Struttura-Reattività (QSRR)*

Si tratta di un'estensione della metodologia *LFER* (in inglese, ***Quantitative Structure-Reactivity Relationships***) ove, come in *QSAR*, si fa largo uso strumenti statistici multivariati, utilizzando anche descrittori di processo e descrittori binari che indicano la presenza/assenza di particolari condizioni di reazione (ad esempio, del catalizzatore).

🗀 *Approccio di Free-Willson*

Si tratta di un particolare approccio alla costruzione di modelli di regressione in cui i descrittori sono variabili indicatrici (*dummy variables*) del tipo di sostituente e della sua posizione.
Ad esempio, si consideri il toluene e i due siti di sostituzione adiacenti al metile (posizioni orto o 2 e meta o 3) e supponiamo che i sostituenti siano il fluoro, il bromo e lo iodio. In Tab.12-1 sono riportati cinque possibili composti. La variabile sito1 - F indica la presenza (1) o l'assenza (0) del fluoro in posizione orto rispetto al metile del toluene; la variabile sito1 - Br indica la presenza (1) o l'assenza (0) del bromo in posizione orto rispetto al metile del toluene; e così via.

	sito 1: orto / pos.2			*sito 2: meta / pos.3*		
composto	*F*	*Br*	*I*	*F*	*Br*	*I*
toluene	0	0	0	0	0	0
2-iodo-toluene	0	0	1	0	0	0
3-iodo-toluene	0	0	0	0	0	1
2,3-difluoro-toluene	1	0	0	1	0	0
2-bromo, 3-fluoro-toluene	0	1	0	1	0	0

TAB. 12-1

Il numero di variabili indipendenti è dato dal totale dei gruppi sostituenti considerati per i siti di sostituzione (nell'esempio: 3 sostituenti x 2 siti = 6). Ogni composto è definito da un vettore di zeri e uno e la matrice dei descrittori è quindi interamente costituita da valori zero e uno; il modello di regressione

cerca di mettere in relazione la risposta sperimentale Y con le posizioni e il tipo di sostituenti.
Questo tipo di approccio presenta il vantaggio di essere facilmente applicabile e di non dipendere dalla conoscenza di alcuna proprietà chimico-fisica; tuttavia la tipologia delle variabili è tale da presentare normalmente notevoli problemi di predittività.

🗀 *Relazioni Attività-Composizione (CARE)*

Si tratta di un'applicazione particolare delle metodologie *QSAR*, che pone l'accento sulle relazioni tra la composizione chimica di un singolo composto e un effetto biologico (***Composition-Activity RElationships***).
Si incontra spesso questo tipo di problema in campo ambientale, ove si pressupponga che l'effetto considerato dipende dalla variazione di concentrazione dei costituenti chimici presenti in un composto.
In questo caso, insieme ai comuni metodi utilizzati nell'analisi multivariata (PCA, *cluster analysis*, regressione, ecc.), particolare attenzione viene data ai metodi di decomposizione della varianza e covarianza.

12.4 - Chemiometria e modelli QSAR

Lo sviluppo della chemiometria ha messo in luce delle nuove potenzialità nello sviluppo dei modelli *QSAR*.
In primo luogo, la logica della validazione consente di sviluppare modelli anche complessi (presenza di molte variabili, modelli non-lineari) per i quali è possibile valutare realisticamente la qualità predittiva. Accanto alle procedure di validazione, il cui scopo è quello di trovare la complessità ottimale del modello in grado di fornire il massimo potere predittivo, in molti metodi chemiometrici si è posto l'accento sulla ricerca della *rilevanza di ciascuna variabile nel modello* (*loadings* in PCA, coefficienti standardizzati in regressione, potere modellante di una variabile in PLS, metodi di selezione di un sottoinsieme di variabili, ecc.).
Da tutto ciò emerge la possibilità di utilizzare nella fase iniziale della ricerca di un modello non più alcune variabili preselezionate *ad-hoc*, ma un numero più elevato (anche molte centinaia) di variabili candidate: le variabili che si manifesteranno come poco o per nulla rilevanti nel modello saranno successivamente eliminate a favore di quelle variabili specifiche che sono correlate con la risposta studiata.

Questa impostazione dà una rilevanza notevole alla ricerca di nuovi descrittori dei sistemi studiati. In particolare, nei problemi *QSAR*, lo sviluppo di descrittori molecolari teorici (ad esempio, topologici e tridimensionali) gioca un ruolo fondamentale nella ricerca rivolta a predire risposte sperimentali complesse da grandezze calcolabili teoricamente.
La costruzione di adeguato *training set* per costruire modelli sufficientemente rappresentativi del problema esaminato costituisce una fase di importanza fondamentale per le strategie *QSAR*. In Fig. 12-3 viene schematizzato il problema.
I dati vengono reperiti non solo dalle banche dati esistenti e dalla letteratura, ma possono anche essere calcolati per via teorica.
I *fattori limitanti* più importanti nella costruzione del *training set* riguardano la disponibilità di banche date aggiornate, l'omogeneità e l'accuratezza dei dati sperimentali, la disponibilità e la rappresentatività di descrittori teorici.
Il *training set* deve essere costituito da un numero di dati sufficientemente numerosi da garantire di rappresentare il problema in modo adeguato e devono essere sufficientemente accurati al fine di evitare che il rumore in essi presente possa sovrastare l'informazione che si vuole estrare da essi.

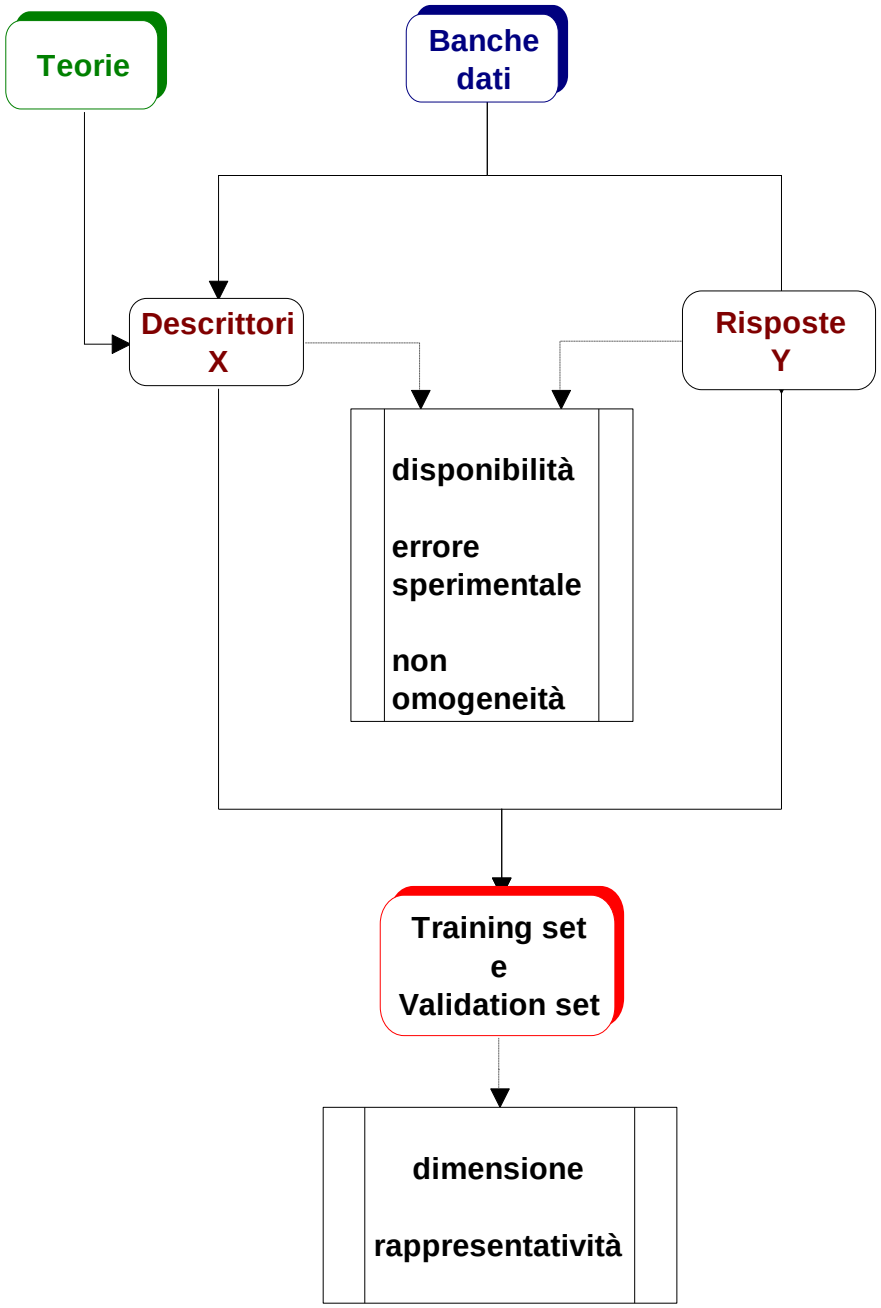

Teorie
Banche dati
Descrittori X
Risposte Y
disponibilità
errore sperimentale
non omogeneità
Training set e Validation set
dimensione
rappresentatività

FIG. 12-3

12.5 - *3D-QSAR* in generale

L'approccio tridimensionale allo studio delle relazioni attività-struttura è fortemente legato all'approccio tradizionale *QSAR*, ove viene fatto un largo uso dei metodi chemiometrici. Esso tuttavia si differenzia dalle strategie tradizionali per il fatto che i descrittori molecolari di cui si fa uso considerano in qualche modo gli aspetti 3D della molecola.
Viene quindi dato un largo spazio alle informazioni che tengono conto degli aspetti tridimensionali delle molecole, della geometria molecolare e degli aspetti conformazionali. Questo comporta inevitabilmente un pesante aggravio di lavoro in quanto non è più possibile definire la molecola semplicemente mediante il suo grafo molecolare (2D), ma è necessario effettuare calcoli che portino ad una "vera" struttura tridimensionale di minima energia. Inoltre, per molecole che presentano una certa flessibilità conformazionale, si deve tener conto che si possono avere più conformazioni di minima energia, geometricamente anche molto diverse tra loro, e che l'azione biologica potrebbe essere espletata non dalla conformazione di minima energia, ma da un suo stato di transizione.
A questo scopo sono stati sviluppati molti programmi che utilizzano per il calcolo delle proprietà molecolari diversi metodi caratteristici della chimica teorica (metodi quantistici semi-empirici, metodi di meccanica molecolare, eccetera). Utilizzando i risultati di questi calcoli è possibile "vedere" la molecola rappresentata da superfici di risposta che modellano le proprietà molecolari e permettono di studiare visivamente le singole molecole (Fig. 12-4).
Questo approccio trova un largo impiego nei settori in cui l'interesse per le relazioni tra struttura e attività è dettato soprattutto per la progettazione di farmaci (***drug design***). Come è noto, infatti, la progettazione di un nuovo composto con proprietà o attività farmacologiche predefinite, la sua sintesi, la ricerca sperimentale farmacologica ed i test clinici necessari sono un impegno estremamente rilevante e i cui esiti sono comunque incerti nella maggior parte dei casi.
Molte strategie teoriche basate sui principi su cui si fondano gli studi delle relazioni struttura-attività vengono oggi utilizzate col proposito di evitare, almeno in parte, questa pesante e costosa attività sperimentale.

L'attuale rilevanza delle strategie *3D-QSAR* è legata allo sviluppo delle stazioni di lavoro grafiche (*work stations*), con potenze di calcolo e disponibilità di memoria tali da poter essere in molti casi confrontabili con quelle dei grandi computers. Ciò ha permesso di adottare in modo sistematico metodologie *QSAR*

che consentono di studiare i problemi a livello molecolare e di interazione molecola-recettore anche mediante visualizzazioni grafiche (***Computer-Aided Molecular Design, CAMD***).

I due metodi più noti sono *GRID* e *CoMFA*: i campi di interazione di ***GRID*** rappresentano le energie totali (somma delle interazioni date dai potenziali elettrostatici, di Lennard-Jones, di legame idrogeno), mentre nel metodo ***CoMFA*** (***COmparative Molecular Field Analyis***) i potenziali sono di tipo elettrostatico o sterico. In entrambi i casi, i campi otttenuti possono essere utilizzati come descrittori puntuali della struttura molecolare e delle sue interazioni, particolarmente adatti allo studio di interazioni di *binding*.

Si tratta di descrittori ad alta dimensionalità in quanto i punti della griglia sono generalmente molte migliaia. Questo permette una rappresentazione grafica delle interazioni molecolari che consente di evidenziare le regioni di massima e minima interazione con il *probe*, permettendo in molti casi delle semplici interpretazioni. Tuttavia, quando lo studio è rivolto al confronto tra più molecole, viene richiesto il postulato di congenericità e un'attenta analisi preliminare che consenta di orientare opportunamente nella griglia le molecole (*alignment*) al fine di poter confrontare tra loro i campi di potenziale ottenuti. Inoltre, proprio l'alta dimensionalità di questo tipo di descrizione molecolare non consente sempre un immediato utilizzo di questi descrittori nella ricerca di relazioni tra struttura e attività/proprietà della molecola.

Per ovviare a quest'ultimo inconveniente, il metodo CoMFA è stato integrato dal metodo ***GOLPE*** (***Generating Optimal Linear PLS Estimations***), il cui scopo è quello di selezionare - tra le migliaia di punti della griglia - un sottoinsieme di punti di massima informazione mediante un opportuno disegno sperimentale (*D-ottimale*). Questi punti vengono successivamente utilizzati come descrittori per modellare la risposta biologica mediante il metodo di regressione *PLS*.

Recentemente è stata proposta una via alternativa che si propone di superare molte delle difficoltà incontrate nell'utilizzo del metodo CoMFA ed, in particolare, i problemi dovuti all'allineamento delle molecole nel campo e all'alto numero di variabili. Il metodo è basato sull'utilizzo dei descrittori ***G-WHIM*** (***Grid-Weighted Holistic Invariant Molecular descriptors***) e consiste nel generare il campo di energie potenziali, come nei metodi precedenti, ma considerando ogni molecola indipendentemente dalle altre, evitando quindi i problemi dovuti all'allineamento. Poichè in questo caso i singoli punti corrispondenti per molecole diverse non sono più tra loro confrontabili, l'informazione contenuta nei potenziali della griglia viene trattata globalmente mediante una variante pesata dell'analisi delle componenti principali.

L'informazione contenuta nel campo viene condensata in un insieme ridotto di descrittori 3D calcolati secondo gli algoritmi di calcolo dei descrittori ***WHIM*** (vedi Capitolo 13, Descrittori WHIM). Tuttavia, mentre ogni gruppo di descrittori WHIM viene calcolato pesando in diverso modo le coordinate degli atomi della molecola (utilizzando come pesi le masse atomiche, le polarizzabilità atomiche, le elettronegatività atomiche di Mulliken, i volumi atomici di van der Waals, ecc.), i descrittori G-WHIM vengono calcolati pesando le coordinate di ciascun punto di griglia con il suo valore di potenziale. Nel caso più semplice, si ottengono due gruppi di descrittori, uno ottenuto pesando i punti della griglia soltanto con i potenziali positivi, l'altro ottenuto pesando gli stessi punti con i valori assoluti dei potenziali negativi. Se il campo di potenziale è calcolato in modo sufficientemente denso, cioè in punti sufficientemente vicini tra loro, si dimostra che i descrittori G-WHIM hanno le stesse proprietà di invarianza roto-traslazionale dei descrittori WHIM e consentono una efficace rappresentazione del campo di potenziale molecolare generato e delle sue proprietà.
Questo metodo può essere utilizzato in modo estremamente flessibile, calcolando, ad esempio, campi di interazione con *probes* diversi, selezionando i punti della griglia in base a diversi *cut-off* di energia, o considerando solo i punti della griglia che appartengono a superfici determinate.
Per la ricerca delle relazioni attività-struttura, l'utilizzo combinato dei descrittori WHIM e dei descrittori G-WHIM appare una strategia molto promettente, potendosi combinare le informazioni contenute nella struttura 3D della molecola (WHIM) con quelle relative alle interazioni della molecola con l'esterno (G-WHIM).

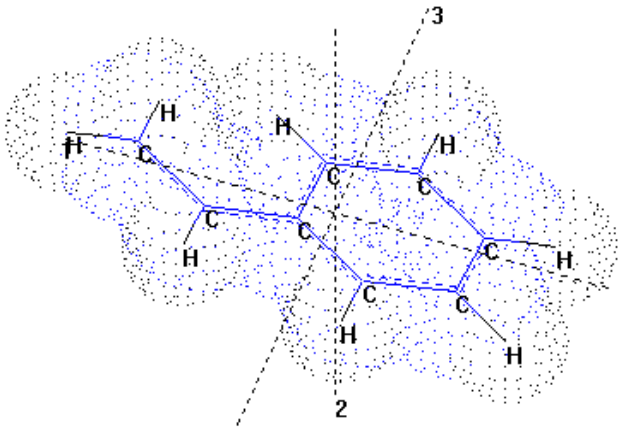


FIG. 12-4

🕮 Bibliografia

M. BARONI, G. COSTANTINO, G. CRUCIANI, D. RIGANELLI, R. VALIGI E S. CLEMENTI (1993). *Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems*. Quant. Struct.-Act. Relat., **12**, 9-20.

R.D. CRAMER III, D.E. PATTERSON E J.D. BUNCE (1988). Comparative Field Analysis (CoMFA). 1. *Effect of Shape on Binding os Steroids to Carrier Proteins*. J. Am. Chem. Soc., **110**, 5959-5967.

P.J. GOODFORD (1985). *A computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules*. J. Med. Chem., **28**, 849-857.

C. HANSCH E A.J. LEO, Eds. (1979). *Substituents Constants for Correlation Analysis in Chemistry and Biology*. Wiley, New York, NY.

R. TODESCHINI, M. LASAGNI E E. MARENGO (1994). *New Molecular Descriptors for 2D- and 3D-structures. Theory*. J.Chemometrics, **8**, 263-272

R. TODESCHINI, P. GRAMATICA E R. PROVENZANI (1995). *Weighted Holistic Invariant Molecular descriptors. Part 2. Theory development and applications on modeling physico-chemical properties of PolyAromatic Hydrocarbons (PAH)*. Chemometrics and Intelligent Laboratory Systems, **27**, 221-229

13

I DESCRITTORI NEI PROBLEMI QSAR

13.1 - Introduzione

Il linguaggio con cui esprimiamo i concetti e descriviamo la realtà non coincide con la realtà stessa, ma svolge tuttavia una parte attiva nel modellare la realtà di cui intendiamo parlare. L'interpretazione di ogni "fatto" è quindi indissolubilmente legata ai modi con cui il fatto viene descritto attraverso la mediazione del linguaggio ed il significato di ogni termine utilizzato dipende dal contesto teorico in cui si trova.
In questo contesto, i descrittori costituiscono gli elementi del linguaggio con cui rappresentiamo l'oggetto studiato, sia esso un sistema chimico, fisico o biologico.
Come è facilmente immaginabile, il tipo di descrittori utilizzati nella costruzione di modelli per le relazioni tra attività/proprietà e struttura molecolare è quindi decisivo nel determinare la qualità dei modelli ottenuti.
L'importanza del problema è evidente se si pensa che al momento di ipotizzare l'utilizzo di un nuovo composto sarebbe di grande rilevanza avere la possibilità di predire alcune proprietà (chimico-fisiche o farmacologiche) prima della sintesi del composto stesso o comunque prima di effettuare un'estesa sperimentazione. Ad esempio, avere una stima attendibile della solubilità, del punto di fusione, del punto di ebollizione, della tensione superficiale, di un'attività biologica definita, della capacità di penetrare membrane lipidiche di un composto, utilizzando altre proprietà già note del composto, sarebbe di grande importanza per il formulatore.

La più comune classificazione dei descrittori, in accordo con l'approccio di Hansch, fa riferimento a tre gruppi di proprietà fondamentali:

a) idrofobicità (ad esempio, logP)
b) parametri elettronici (ad esempio, le energie HOMO e LUMO)
c) parametri sterici, di forma e dimensione (ad esempio, il peso molecolare)

I descrittori molecolari possono presentare le caratteristiche e le tipologie più diverse: in particolare, possono (a) provenire da misure sperimentali, da calcoli teorici e da semplici operazioni di conto o di somma; (b) rappresentare l'intera molecola o un particolare frammento molecolare o un sostituente in un sito definito; (c) richiedere la conoscenza della struttura 3D della molecola, oppure il suo grafo molecolare o semplicemente la formula bruta; (d) essere definiti da uno scalare, da un vettore, da un campo di scalari, etc.

I descrittori molecolari si possono dividere, ad esempio, in descrittori monodimensionali (1D), bidimensionali (2D), tridimensionali (3D), a livello microscopico e macroscopico, ecc.
I descrittori a livello 1D sono i descrittori più semplici, derivabili direttamente dalla formula bruta della molecola; si tratta di descrittori scalari quali, ad esempio, il peso molecolare, il numero di atomi di carbonio e di legami. A livello 2D si collocano i descrittori derivati dal grafo molecolare e dalla matrice di connettività, cioè tipicamente i descrittori topologici. A livello 3D sono definiti tutti i descrittori che provengono dalla struttura tridimensionale della molecola, cioè dalla conoscenza delle sue effettive coordinate spaziali.
Nel livello successivo, in riferimento al mondo microscopico, sono definiti tutti i descrittori che tengono conto degli aspetti conformazionali delle molecole e, nel mondo macroscopico, i descrittori geometrici come volumi e superficie delle molecole. Nella direzione del microscopico, il livello successivo è il livello elettronico, ove possiamo collocare descrittori come le energie degli orbitali molecolari, le mappe di densità elettronica, i parametri quanto-meccanici, ecc. Infine, a livello globale, definiamo i descrittori che tengono conto anche delle interazioni con l'intorno, come sono di fatto la maggior parte delle grandezze misurate sperimentalmente.
I descrittori molecolari costituiscono un importante sottoinsieme di descrittori di diverso genere (chimici, fisici, biologici, ambientali, geografici e geologici, etc.) che vengono utilizzati nella costruzione dei modelli matematici.
Alla luce di queste considerazioni, i descrittori sono qui stati suddivisi nei seguenti gruppi:

Descrittori	*Fonte principale*
COMPOSIZIONALI	SPERIMENTALI
CHIMICO-FISICI	SPERIMENTALI
QUANTO-MECCANICI	TEORICI
DI GRUPPI SOSTITUENTI	CALCOLATI DA DATI SPERIM.
GLOBALI GEOMETRICI	CALCOLATI
LOCALI GEOMETRICI	CALCOLATI
BINARI	CALCOLATI
DI PUNTEGGIO	CALCOLATI
ENUMERATIVI	DEFINITI DALLA STRUTTURA
MATRICIALI DI CONNETTIVITÀ	DEFINITI DALLA STRUTTURA
TOPOLOGICI	TEORICI
DI CORRELAZIONE STRUTTURALE	TEORICI
WHIM E G-WHIM	TEORICI
DIFFERENZIALI	TEORICI
CROMATOGRAFICI	SPERIMENTALI
SPETTROSCOPICI	SPERIMENTALI
DI INTERAZIONE A CAMPI SCALARI	TEORICI
DI SIMILARITÀ MOLECOLARE	TEORICI
DI REATTIVITÀ CHIMICA E PROCESSO	SPERIMENTALI
DI ATTIVITÀ BIOLOGICA	SPERIMENTALI
CHEMO-AMBIENTALI	SPERIMENTALI
DI PROPRIETA' PRINCIPALI	CALCOLATI

13.2 - Descrittori composizionali

Questi descrittori sono i più caratteristici descrittori chimici, sono cioè descrittori delle concentrazioni relative alle diverse sostanze che compaiono in un campione. Si tratta quindi grandezze legate alla concentrazione di una sostanza in un campione e le unità di misura più utilizzate sono grammi/mole, milligrammi/mole, grammi/litro, milligrammi/litro, moli/litro, millimoli/litro, ppm. Il numero di variabili composizionali dipende dal numero di composti potenzialmente presenti nei campioni che si vogliono studiare.

🗀 concentrazione assoluta

E' la concentrazione di un composto espressa direttamente come la quantità, in peso o in moli, per unità di volume o in unità molari.

🗀 concentrazione relativa o percentuale

Quando nei campioni in esame l'interesse si concentra su un numero definito di composti o di specie atomiche, si usa comunemente trasformare le loro concentrazioni assolute in concentrazioni relative o percentuali. Si ottengono così, per ogni campione, dei *profili* di concentrazione. Se le percentuali sono calcolate su tutti i composti considerati, si deve tenere conto della condizione di chiusura imposta (la somma delle percentuali di un campione, cioè una riga della matrice dei dati, è uguale a 100), per cui le variabili indipendenti sono p - 1. Questo tipo di rappresentazione è tipico delle miscele.

🗀 recupero % (*recovery* %)

E' la quantità percentuale di sostanza recuperata mediante un opportuno processo di estrazione rispetto alla quantità totale di sostanza presente nella matrice iniziale.

13.3 DESCRITTORI CHIMICO-FISICI

I descrittori chimico-fisici sono il gruppo più importante di descrittori per il loro numero, la loro grande varietà e soprattutto per le loro rilevanza nella descrizione delle caratteristiche molecolari.

🗀 peso molecolare (MW)

Il peso molecolare di un composto è definito come la somma dei pesi atomici di tutti gli atomi che lo compongono. Il peso molecolare è quindi calcolabile semplicemente dalla formula bruta di un composto e, rappresentando il valore totale della massa del composto, è un indice dimensionale sempre disponibile.

🗀 punto di ebollizione (bp, T_{pb})

Il punto di ebollizione è definito come la temperatura alla quale un liquido è in equilibrio con la propria pressione di vapore ed è quindi un indice della volatilità del composto e delle forze intermolecolari in fase liquida.
Normalmente i valori dei punti di ebollizione si riferiscono a 760 torr (1 atm).

🗀 punto di fusione (mp, T_{mp})

Il punto di fusione è definito come la temperatura alla quale un solido è in equilibrio con la fase liquida e rappresenta un indice delle forze intermolecolari in fase solida, delle forze di impaccamento del cristallo e del suo grado di simmetria.

🗀 pressione di vapore (p)

La pressione di vapore di un liquido è la pressione esercitata dal vapore quando esso è in equilibrio con la fase liquida o solida. La pressione di vapore è una proprietà specifica del composto.

🗀 densità (d o ρ)

La densità di una sostanza è definita come la massa per unità di volume ed è espressa in grammi per centimetro cubico in unità cgs o in kilogrammi per metro cubo in unità SI. Sovente la densità viene espressa in grammi per millilitro; un millilitro corrisponde a 1.000028 cm^3.

🗀 indice di rifrazione (n, n_D)

L'indice di rifrazione è il rapporto tra la velocità della luce in una sostanza e la velocità della luce nel vuoto.
L'indice di rifrazione è una funzione sia della temperatura (viene misurato normalmente tra 15 °C e 30 °C), sia della lunghezza d'onda della luce incidente. Di norma, vengono utilizzate le linee del sodio D_1 e D_2 , la cui media pesata è pari a 5892.6 Angstrom è viene identificata con D.
L'equazione empirica di Eykman consente di correlare l'indice di rifrazione con la densità ρ:

$$\left(\frac{n^2-1}{n+0.4}\right)\cdot\frac{1}{\rho}=C$$

dove *C* è una costante indipendente dalla temperatura.
L'equazione di Lorenz-Lorentz ha basi teoriche più fondate e correla l'indice di rifrazione con la densità, il peso molecolare e la rifrazione molare:

$$\frac{n^2-1}{n^2+2}\cdot\frac{MW}{\rho}=MR$$

Questa espressione può essere scritta evidenziando l'indice di rifrazione nel seguente modo:

$$n=\left[\frac{V_m+2\cdot MR}{V_m-MR}\right]^{1/2}$$

dove V_m è il volume molare.

🗀 coefficiente di viscosità (η)

La viscosità di un liquido è un indice delle forze che si oppongono al movimento o al suo scorrimento quando viene applicata una forza di taglio; la viscosità è quindi un fattore rilevante per tutti i fenomeni di trasporto dei liquidi. Il coefficiente di viscosità è definito come la forza per unità di area necessaria per conservare un gradiente di velocità unitario tra due piani paralleli a distanza unitaria.
L'unità di misura cgs è il *poise* (dyne s/cm^2 = g/s cm); sono utilizzate anche il *millipoise* (mP) e il *centipoise* (cP). L'unità di viscosità in unita SI è Newton s/cm^2 (N s/m^2 = kg/s m).
Recentemente Hildebrand e Bachinski hanno proposto la seguente relazione, valida per i composti non-polari, con il volume molare:

$$\frac{1}{\eta}=\phi=B_2\left(\frac{V_m-V_0}{V_0}\right)$$

dove ↗ è la **fluidità**, V_m è il volume molare del composto, V_0 è il volume molare quando $\phi=0$ e B_2 è una costante.

🗀 osmolarità e osmolalità (Π)

La pressione osmotica Π è definita come la pressione richiesta per innalzare la pressione di vapore di una soluzione a quella del solvente puro ed è definita come:

$$\Pi = \frac{RT}{V} \cdot n_B$$

dove n_B sono le moli di soluto nella soluzione.
L'osmolarità è la pressione osmotica di una soluzione 1 molare (una mole / litro di solvente); l'osmolalità è la pressione osmotica di una soluzione 1 molale (una mole / 1000 g di solvente).

🗀 tensione superficiale (σ)

La tensione superficiale è definita come la forza che agisce perpendicolarmente a una linea di 1 cm sulla superficie di un liquido.
L'energia richiesta per aumentare l'area superficiale di un'area unitaria di 1 cm^2 o di 1 m^2 è chiamata *energia superficiale* ed è numericamente uguale alla tensione superficiale che si oppone all'aumento.
La relazione più importante per stimare la tensione superficiale è data dall'equazione di MacLeod e Sugden:

$$\sigma = \left[Pr \cdot \frac{(\rho_L - \rho_V)}{MW} \right]^4$$

dove *Pr* è il paracoro, ρ_L e ρ_V sono rispettivamente le densità del liquido e del vapore (quest'ultima è sovente trascurabile).

🗀 rifrazione molare (MR)

La rifrazione molare ha le unità di un volume molare ed è definita come

$$MR = \frac{n^2 - 1}{n^2 + 2} \cdot \frac{MW}{\rho}$$

dove *n* è l'indice di rifrazione molare, *MW* il peso molecolare e ρ la densità.
La rifrazione molare è senz'altro correlata al volume molecolare. Tuttavia, essa è anche proporzionale alla *polarizzabilità elettronica* α_E: infatti, la *polarizzazione elettronica* P_E è numericamente uguale alla rifrazione molare per la luce visibile:

$$MR = \frac{n^2 - 1}{n^2 + 2} \cdot \frac{MW}{\rho} = P_E = \frac{4}{3} \cdot \pi N \alpha_E$$

dove N è il numero di Avogadro. Per questo motivo, si considera *MR* contemporaneamente un parametro dimensionale ed elettronico.

🗀 calore di vaporizzazione (ΔH_v)

Il calore di vaporizzazione è la quantità di calore richiesta per vaporizzare una quantità definita (una mole) di un liquido, alla temperatura di ebollizione. Viene anche chiamato *calore latente di vaporizzazione*.
Una stima del calore di vaporizzazione può essere ottenuta dalla temperatura di ebollizione mediante l'equazione di Kistiakowsky:

$$\Delta H_v = 8.75 \cdot T_{bp} + 4.571 \cdot \log T_{bp}^2$$

🗀 calore di fusione (ΔH_m)

Il calore di fusione è la quantità di calore richiesta per liquefare una quantità definita (una mole) di un solido alla temperatura di fusione. Viene anche chiamato *calore latente di fusione*.

🗀 calore di sublimazione (ΔH_s)

Il calore di sublimazione è la quantità di calore richiesta per trasformare una quantità definita di un solido (una mole), ad una temperatura definita, direttamente allo stato di vapore alla pressione di vapore di equilibrio .

🗀 calore di formazione (ΔH_f)

Il calore di formazione è la quantità di calore assorbito (reazione endotermica, $\Delta H_f > 0$) o emesso (reazione esotermica, $\Delta H_f < 0$) nella formazione di un composto partendo dagli elementi che lo compongono. Il calore di formazione è un indice della stabilità termodinamica di un composto.
Quando il composto ed i suoi elementi sono in uno stato standard (1 atm e 25 °C), esso viene chiamato *calore di formazione standard* (ΔH_f^0).

🗀 entropia standard (S^0)

L'entropia standard di un composto, a temperatura T e nel suo stato standard, è un indice dello stato di disordine della sostanza riferito ad un'entropia uguale a zero del cristallo perfetto allo zero assoluto.

🗀 entropia standard di reazione (ΔS^0)

L'entropia standard di reazione è un indice (insieme al fattore entalpico) della tendenza della reazione (processo) a procedere spontaneamente nella direzione definita dalla reazione (processo). Valori negativi dell'entropia di reazione indicano che la reazione procede spontaneamente da sinistra a destra; valori positivi indicano che la reazione procede spontaneamente in senso inverso.

🗀 energia libera standard di reazione (ΔG^0, energia libera di Gibbs)

E' l'energia libera di reazione ed è un indice globale della direzione della reazione chimica. E' legata all'entalpia e all'entropia dalla relazione generale:

$$\Delta G = \Delta H - T\Delta S$$

🗀 temperatura critica (T_c)

La temperatura critica è definita come la temperatura al di sopra della quale un gas non può essere liquefatto.

🗀 pressione critica (P_c)

La pressione critica è la minima pressione richiesta perché avvenga la liquefazione di un gas ad una data temperatura.

🗀 volume critico (V_c)

Il volume critico è il volume occupato da 1 mole di sostanza alla sua temperatura critica e alla sua pressione critica.

🗀 **volume molare (V_m)**

Il volume molare è il volume occupato da una mole di sostanza ed è calcolato come $V_m = MW/\rho$, cioè come il rapporto tra il peso molecolare di un composto e la sua densità. I valori del volume molare sono influenzati dalle interazioni intermolecolari.

🗀 **paracoro (Pr)**

Il paracoro, come definito da Sugden, è il volume molare per una funzione della tensione superficiale σ del composto:

$$Pr = \frac{MW}{\rho} \cdot \sigma^{1/4} = V_m \cdot \sigma^{1/4}$$

🗀 **capacità termica specifica a pressione costante (C_p)**

La capacità termica specifica a pressione costante è la quantità di calore richiesta per innalzare di 1 °C la temperatura di 1 grammo di sostanza, a pressione costante.

🗀 **capacità termica specifica a volume costante (C_v)**

La capacità termica specifica a volume costante è la quantità di calore richiesta per innalzare di 1 °C la temperatura di 1 grammo di sostanza, a volume costante.

🗀 conduttività termica (k)

La conduttività termica è una proprietà di trasporto di una sostanza e descrive la capacità di una sostanza di condurre calore. Questa proprietà viene espressa come quantità di calore che passa nell'unità di tempo per unita di area, per una differenza di temperatura di 1°C per uno spessore unitario di sostanza. La conduttività termica dei liquidi è sempre maggiore di quella dei gas, alla stessa temperatura.

🗀 coefficiente di conduttività termica (λ)

Il coefficiente di conduttività termica è la velocità di flusso di calore moltiplicata per l'area da cui avviene il trasferimento di calore e per il gradiente di temperatura.
In unità SI è dato da $J\,s^{-1}\,m^{-1}\,deg^{-1}$, mentre in unità cgs è dato da $cal\,s^{-1}\,cm^{-1}\,deg^{-1}$.

🗀 costante crioscopica (K_f)

La costante crioscopica di un solvente è definita come

$$K_f = \frac{MW_{solv} R T_{mp}^2}{1000 \cdot \Delta H_m}$$

dove $MW_{solv}, T_{mp}, \Delta H_m$ sono, rispettivamente, il peso molecolare, la temperatura di fusione e il calore molare di fusione del solvente e R è la costante dei gas (1.9872 cal K^{-1} mol^{-1}). Il valore di K_f è generalmente espresso gradi per mole per 1000 g di solvente.

🗀 costante ebullioscopica (K_B)

La costante ebullioscopica di un solvente è definita come

$$K_B = \frac{MW_{solv} R T_{bp}^2}{1000 \cdot \Delta H_v}$$

dove $MW_{solv}, T_{bp}, \Delta H_v$ sono rispettivamente, il peso molecolare, la temperatura di ebollizione e il calore molare di vaporizzazione del solvente e *R* è la costante dei gas (1.9872 cal K^{-1} mol^{-1}). Il valore di K_B è generalmente espresso gradi per mole per 1000 g di solvente.

🗀 fattore acentrico (ω)

E' un parametro, detto anche fattore acentrico di Pitzer, comunemente usato per definire la non-sfericità (o acentricità) di un composto. Viene calcolato come:

$$\omega = -log\ P_r - 1$$

dove P_r è la pressione di vapore ridotta (P/P_C) calcolata alla temperatura $T_r = T/T_c = 0.7$.
Valori approssimati del fattore acentrico possono essere calcolati utilizzando la seguente espressione:

$$\omega = \frac{3}{7} \cdot \frac{T_{bp}/T_c}{1 - T_{bp}/T_c} log\ P_c - 1$$

Il fattore acentrico è zero per i gas monoatomici. Per molecole di grande dimensione esso cresce significativamente non solo con la non-sfericità del composto, ma anche con la sua polarità. Quindi valori grandi del fattore acentrico misurano contemporaneamente il grado di non-sfericità del composto e il suo grado di polarità.

🗀 energia di prima ionizzazione (I_p)

E' l'energia necessaria per strappare il primo elettrone ad una molecola neutra ed è un indice della sua capacità di ionizzarsi. Normalmente viene espressa in $kJ\ mol^{-1}$.

🗀 affinità protonica (PA)

E' l'energia rilasciata in una reazione in fase gassosa nella reazione (reale o teorica) tra un protone ed un composto per fornire l'acido coniugato corrispondente. E' una misura del potere nucleofilo del composto.

🗀 conduttività elettrica (κ)

La conduttività (o *conduttanza specifica*) è il reciproco della resistenza specifica, che è la resistenza che offre una sostanza quando viene posta tra due elettrodi paralleli piani di 1 cm^2 di area, posti a una distanza di 1 cm. La conduttività elettrica è una misura della capacità del composto di condurre corrente.

🗀 costante dielettrica (ε)

La costante dielettrica è una misura dell'effetto relativo che una sostanza ha sulla forza con cui due cariche opposte, immerse nella sostanza, si attraggono l'una con l'altra. La costante dielettrica nel vuoto è definita uguale a 1; per motivi pratici, le misure sono comunque effettuate in aria.
Qualitativamente, sostanze con un'elevata costante dielettrica sono buoni mezzi ionizzanti. Le solubilità possono essere stimate dalla costante dielettrica.

🗀 momento dipolare (μ)

Il momento dipolare è una misura della separazione del centro di gravità delle cariche positive e negative di una molecola. Esso è definito come il prodotto tra una carica (espressa in Coulomb) e una distanza (espressa in metri).
Il momento dipolare è comunemente espresso anche in *Debye* (*D*):

$$1D = 3.336 \times 10^{-30} \ \ C\,m$$

Il momento dipolare è importante per lo studio della polarità e polarizzabilità di una molecola e si correla alla solubilità, alla simmetria molecolare e a molte proprietà dei solventi.
In generale, sostanze non polari sono volatili e solubili in altri mezzi non polari, mentre sostanze polari sono meno volatili e solubili sono in mezzi polari.

🗀 suscettibilità elettrica (χ_e)

La suscettibilità elettrica è una grandezza adimensionale che rappresenta il fattore di proporzionalità tra la polarizzazione totale *P* di un composto ed il campo elettrico efficace E^*, secondo l'espressione:

$$P = \varepsilon_0 \chi_e E^*$$

dove ε_0 è la costante dielettrica del vuoto.
Essa viene calcolata generalmente dall'espressione:

$$\chi_e = 3 \cdot \frac{\varepsilon - 1}{\varepsilon + 2}$$

🗀 polarizzazione elettronica molare (P_E)

La polarizzazione molare di una sostanza è definita dall'equazione di Clausius-Mosotti:

$$P_E = \frac{\varepsilon - 1}{\varepsilon + 2} \cdot \frac{MW}{\rho}$$

dove ε, MW, ρ sono rispettivamente la costante dielettrica, il peso molecolare e la densità della sostanza.

🗀 polarizzabilità (α_E)

Poiché anche le molecole non polari possono acquisire un momento dipolare (indotto) se immerse in un campo elettrico *E*, è possibile definire la polarizzabilità come il fattore di proporzionalità tra il momento dipolare indotto ed il campo elettrico:

$$\mu_{ind} = \alpha_E \cdot E$$

L'unità di misura della polarizzabilità è $J^{-1} C^2 m^2$ quando il campo elettrico è espresso in $V\ m^{-1}$ e il momento dipolare in $C\ m$.

🗀 volume di polarizzabilità (α')

Le polarizzabilità α_E sono generalmente espresse come volume di polarizzabilità α'_E, secondo l'espressione

$$\alpha'_E = \frac{\alpha_E}{4\pi\varepsilon_0}$$

dove $4\pi\varepsilon_0$ ha unità $J^{-1}C^2\ m^{-1}$. Spesso α'_E è espresso in cm^3 o $Angstrom^3$.

🗀 costante molare di Kerr ($_mK$)

La costante molare di Kerr è la *birifrangenza ottica*, cioè la misura di valori diversi dell'indice di rifrazione, indotta in una sostanza isotropa per effetto dell'applicazione di un campo elettrico. La costante molare di Kerr è definita come

$$_mK = \frac{N \cdot (\vartheta_1 + \vartheta_2)}{18\varepsilon_0}$$

dove N è il numero di Avogadro, ε_0 è la permittività del vuoto, ϑ_1 è un termine che dipende dalla anisotropia di polarizzabilità e ϑ_2 è un termine che dipende anche dal momento di dipolo permanente. Questa grandezza è legata alla polarizzabilità della molecola e, quindi, anche alla sua simmetria.

🗀 suscettibilità magnetica (χ_m)

La suscettibilità magnetica è il fattore di proporzionalità tra la magnetizzazione M (cioè il *momento di dipolo magnetico* per unità di volume) di un composto e il campo magnetico applicato H:

$$M = \chi_m \cdot H$$

La suscettibilità magnetica, come quella elettrica, è una grandezza adimensionale.

🗀 parametro di polarità di Reichardt-Dimroth (E_T e E_T^N)

E_T un parametro empirico di polarità, definito come energia di transizione, alla temperatura di 25 °C e alla pressione di 1 bar; della banda di assorbimento UV/Vis del composto N-fenossi betaina piridinio quando dissolto in un solvente; la sua unità di misura è *kcal* mol^{-1}. $\mathbf{E}_T^N$ è il parametro di polarità di Reichardt-Dimroth normalizzato, cioè riferito al valore ottenuto per il tetrametilsilano (TMS).

🗀 fattore elettrostatico (EF)

Utilizzato soprattutto per la classificazione dei solventi, è definito come il prodotto della costante dielettrica per il momento dipolare:

$$EF = \varepsilon \cdot \mu$$

In generale, valori di *EF* tra 0 e 2 indicano solventi idrocarburi, valori compresi tra 2 e 20 solventi elettron-donatori, valori tra 15 e 50 solventi idrossilici, valori maggiori di 50 solventi aprotici dipolari.

🗀 solubilità in acqua (S_W o W)

La solubilità in acqua è chiaramente un parametro di idrofilicità e ci si deve quindi aspettare un comportamento inverso alla parametro di lipofilicità (*coefficiente di ripartizione ottanolo/acqua*).
La solubilità viene generalmente espressa in $mol\ l^{-1}$ o in $kg\ m^{-3}$; sovente viene definita come logaritmo ($\log S_W$). La solubilità in acqua può essere messa in relazione anche con i parametri solvatocromici.

🗀 parametro di solubilità di Hildebrand (δ)

Il parametro di solubilità è una costante fisica che descrive la relazione tra le proprietà chimico-fisiche del solvente e la sua efficacia nel disciogliere soluti specifici.
Il parametro di solubilità, come definito da Hildebrand, è:

$$\delta = \sqrt{\frac{\Delta E_v}{V_m}}$$

dove ΔE_V è il calore di vaporizzazione e V_m il volume molare.

🗀 parametri AQUAFAC (q)

AQUAFAC è un metodo per la determinazione dei coefficienti di attività in acqua mediante contributi di gruppo ed è basato sulla seguente equazione generale per la solubilità:

$$log\, S_w^{oss} = log\, S_w^t + log\, \gamma_w$$

dove γ_W è il coefficiente di attività in acqua.
La solubilità teorica considera il contributo alla solubilità dovuto allo stato cristallino e, a temperatura ambiente, è definito come

$$log\, S_w^t = -\frac{\Delta S_{mp}}{2.303 \cdot 298 \cdot R}\left(T_{mp} - 25\right)$$

Dalle due equazioni precedenti, si ottiene:

$$log\, \gamma_w = log\, S_w^{oss} - \frac{\Delta S_{mp}}{2.303 \cdot 298 \cdot R}\left(T_{mp} - 25\right)$$

Secondo il metodo AQUAFAC, i coefficienti di attività, determinati con quest'ultima equazione, possono essere calcolati come contributo di gruppi funzionali, secondo l'espressione:

$$log\, \gamma_w = \sum n_i q_i$$

dove n_i sono il numero di volte in cui ciascun gruppo funzionale compare nel composto e q_i i contributi di ciascun gruppo al coefficiente di attività.

🗀 parametri UNIFAC

UNIFAC è un metodo per la determinazione dei coefficienti di attività mediante contributi di gruppo ed è basato sulla seguente equazione generale:

$$\log \gamma = \log \gamma^C + \log \gamma^R$$

ove ambedue i contributi sono calcolati come contributi di gruppi funzionali. Il primo contributo - la frazione combinatoria - è legato all'area superficiale e al volume del gruppo funzionale; il secondo - la frazione residua - è legato alle energie di interazione tra il gruppo funzionale e la miscela.

Come per il metodo AQUAFAC, i coefficienti di attività vengono utilizzati per calcolare la solubilità in acqua.

🗀 punto di flash (fp, T_{fp})

Il punto di flash di un liquido è la temperatura minima a cui il suo vapore all'equilibrio con un'aggiunta di aria, ad una pressione totale di 760 Torr, si incendia se innescato da una sorgente di fiamma applicata secondo norme prestabilite.

🗀 funzione di Kirkwood (*K*)

La funzione di Kirkwood, utilizzata soprattutto nella classificazione di solventi, è definita dalla seguente funzione della costante dielettrica:

$$K = \frac{\varepsilon - 1}{2\varepsilon + 1}$$

🗀 coefficiente di ripartizione ottanolo/acqua (K_{ow})

Il coefficiente di ripartizione $\mathbf{K_{ow}}$ è definito come il rapporto tra le concentrazioni all'equilibrio di un soluto distribuito tra due fasi liquidi immiscibili costituite da ottanolo e acqua:

$$K_{ow} = \frac{C_{ottanolo}}{C_w}$$

La concentrazione della fase idrofobica è posta per convenzione al numeratore.
Il largo impiego di questo descrittore e la sua importanza sono legati al fatto che in molti problemi di relazioni tra struttura molecolare e attività biologica esso rappresenta la capacità del composto di penetrare membrane lipidiche, cioè la sua affinità relativa a sostanze lipidiche o polari.
Normalmente, viene utilizzato il logaritmo del coefficiente di ripartizione, cioè $\mathbf{logK_{ow}}$, altrimenti noto in letteratura come **logP**.

🗀 lipolo (*L*)

Il lipolo di una molecola è una misura della distribuzione della lipofilicità. Viene calcolato come somma dei valori atomici di logP, come avviene per il calcolo del momento dipolare:

$$\mathbf{L} = \sum_i \mathbf{r}_i \cdot \mathrm{l}_i$$

dove $\mathbf{r}_i$ è la distanza dell'atomo *i* dall'origine della molecola e l_i è la lipofilicità dell'atomo *i*-esimo.

🗀 coefficiente di ripartizione aria/acqua (K_{aw})

Il coefficiente di ripartizione aria/acqua è definito come il rapporto tra la concentrazione di un composto in aria e la sua concentrazione in acqua, in condizioni di equilibrio, alla temperatura di 25 °C:

$$K_{aw} = \frac{C_a}{C_w}$$

Esso è quindi una grandezza adimensionale. Sovente si utilizza il suo logaritmo **logK_{aw}**.

🗀 costante di Henry (H)

La costante di Henry è il fattore di proporzionalità che mette in relazione la pressione parziale P di un composto con la sua concentrazione in acqua C_w, in condizioni di equilibrio:

$$P = \mathrm{H} \cdot C_w$$

La costante di Henry è uno dei parametri importanti per la valutazione della *volatilità dall'acqua* di un composto ed è direttamente correlata con il coefficiente di ripartizione aria/acqua K_{aw} mediante la relazione:

$$\mathrm{H} = RT \cdot K_{aw}$$

H ha unità di misura $Pa\ m^3\ mol^{-1}$.

🗀 **costante di prima dissociazione acida (K_a)**

La costante di dissociazione K_a di un solvente organico è una misura della forza di un acido quando esso è dissolto in un altro solvente, normalmente l'acqua. Di norma si utilizza il logaritmo: $pK_a = \log(K_a)$.

🗀 **costante di prima dissociazione basica (K_b)**

L costante di dissociazione K_b di un solvente organico è una misura della forza di una base quando essa è dissolta in un altro solvente, normalmente l'acqua. Di norma si utilizza il logaritmo: $pK_b = \log(K_b)$.

13.4 - Descrittori quanto-meccanici

Questo tipo di descrittori deriva da calcoli effettuati con i metodi caratteristici della chimica teorica, siano essi metodi quantistici (metodi *ab-initio*, metodi semi-empirici come MNDO e AM1) o metodi di meccanica molecolare (MM3 di Allinger). Con questi metodi è possibile stimare le geometrie molecolari, le energie ad esse associate, parametri quantistici legati alla distribuzione elettronica nella molecola ed diverse grandezze sperimentali. Qui riportiamo alcune delle grandezze più caratteristiche che sono ottenibili da questi metodi.

🗀 **energia dell'ultimo orbitale occupato (HOMO)**

E' l'energia relativa all'orbitale occupato (o parzialmente occupato) di massima energia.

🗀 **energia del primo orbitale non occupato (LUMO)**

E' l'energia relativa all'orbitale libero di minima energia, cioè l'energia del primo orbitale libero.

🗀 **differenza LUMO-HOMO (GAP)**

E' la differenza tra le energie degli orbitali LUMO e HOMO ed è una misura della capacità di ionizzarsi della molecola.

🗀 **superdelocalizzabilità (S)**

Grandezza introdotta da Fukui, è definita come la somma degli orbitali di frontiera (HOMO o LUMO) di un atomo divisa per l'energia HOMO o LUMO.

🗀 **massima carica netta atomica (q_i)**

E' la massima carica netta totale sull'atomo *i*-esimo.

🗀 **carica netta totale (Q)**

E' la somma delle cariche nette assolute degli atomi che compongono una molecola o un particolare gruppo.

🗀 **ordine di legame (*bond order*, b)**

E' una misura della forza del legame tra una coppia di atomi legati. Comunemente viene considerato l'ordine di legame di tipo □; ad esempio, l'ordine di legame C-C nell'etano è 0, nell'etilene è 1, nell'acetilene è 2, nel benzene è 0.67.

🗀 **minimo ordine di legame totale (MTB)**

E' l'ordine di legame totale minimo su un atomo di carbonio.

🗀 **polarizzabilità molecolare media (POLM)**

E' il valore calcolato della polarizzabilità molecolare media.

🗀 **anisotropia della polarizzabilità molecolare (ANIS)**

E' il valore calcolato cone somme delle differenze tra la polarizzabilità molecolare media e la polarizzabilità lungo ciascuno degli assi di polarizzazione.

🗀 **durezza (*hardness*, ♒)**

E' la resistenza del potenziale chimico a modificare il numero di elettroni.

13. 5 - Descrittori di gruppi sostituenti

Da tempo hanno avuto larga diffusione descrittori relativi a frammenti molecolari o, più esattamente, a gruppi sostituenti. Chiaramente, questo approccio consente una forte semplificazione della rappresentazione del composto e assume implicitamente che, nel confronto tra composti diversi, la parte restante della molecola abbia un comportamento sostanzialmente eguale in tutti i composti considerati. Questo tipo di descrittori ha riscosso, per la sua semplicità, un notevole successo nel trattare molti problemi *QSAR*.
Si deve tuttavia tenere presente che (a) viene presupposta la validità di un comportamento additivo che in realtà deve essere valutato di volta in volta, (b) la descrizione dei frammenti non tiene conto né degli aspetti conformazionali della molecola né degli effetti olistici della stessa e (c) in diversi casi la descrizione di ciò che va considerato come frammento rispetto al resto della molecola non è univoca.
Nonostante i suddetti limiti di questo approccio, la forte semplificazione del problema chimico considerato ha permesso di conseguire dei vantaggi pratici che molto spesso hanno ampiamente compensato gli svantaggi.

🗀 costante idrofobica di Hansch-Fujita (π)

La costante idrofobica di un sostituente, proposta da Hansch e Fujita, è definita come

$$\pi_X = log\ P(RX) - log\ P(RH)$$

dove R è il residuo e X è il sostituente. $\pi_H = 0$ (costante idrofobica dell'idrogeno) viene assunto come riferimento.
Questa costante è quindi riferita ad un sostituente e non al composto. Tuttavia, poiché esiste una dipendenza dell'idrofobicità dal composto nella sua globalità, questa costante non è strettamente additiva, cioè la costanti idrofobiche di un sostituente calcolate da composti di riferimento diversi sono in generale diverse tra loro.
Come descrittore globale viene sovente utilizzata la somma della costanti di Hansch-Fujita dei sostituenti in una molecola.

🗀 costante di Hammett (σ)

L'effetto elettronico di un sostituente X sulla costante di dissociazione dell'acido benzoico è stato razionalizzato da Hammett secondo la relazione:

$$\log\left(Ka_X / Ka_H\right) = \rho \cdot \sigma_X$$

dove ρ è una costante che rappresenta la reazione di riferimento (ρ = 1 per la dissociazione dell'acido benzoico) e σ_X è la costante di Hammett del sostituente, che contiene informazioni sia sugli effetti induttivi sia sugli aspetti di risonanza del sostituente. Valori di σ positivi indicano sostituenti elettron-attrattori, mentre valori di σ negativi indicano sostituenti elettron-donatori.
Per valori positivi di ρ in una reazione reversibile, valori positivi di σ indicano una reazione che procede verso destra, mentre valori negativi di σ indicano una reazione che procede verso sinistra.
In particolare, sono stati determinati i valori per moltissimi sostituenti in posizione *orto* ($\blacklozenge_o$)**,** *meta* ($\blacklozenge_m$) e *para* ($\blacklozenge_p$) per anelli aromatici. Oltre a questi, sono state definite numerose altre costanti ♦ e, tra queste, $\blacklozenge^+$ e $\blacklozenge^-$ che indicano, rispettivamente, il grado di delocalizzazione di carica positiva e negativa del legame ♦.

🗀 sigma-star di Taft (σ^*)

Per catene alifatiche gli effetti elettronici dei gruppi sostituenti possono essere modellati in accordo alla seguente espressione:

$$\sigma^* = \frac{log\left(k/k_0\right)_B - log\left(k/k_0\right)_A}{2.48}$$

dove σ^* è il valore del parametro per il sostituente R in RCOOR', k e k_0 sono le costanti di velocità di idrolisi di RCOOR' e MeCOOR', rispettivamente, e A e B indicano l'idrolisi acida e basica. Il fattore 2.48 consente di scalare i valori nella scala di σ.

🗀 costante sterica di Taft (E_S)

Parametro sterico ancora molto utilizzato in *QSAR*, riferito a gruppi sostituenti e calcolato dal rapporto tra le costanti di idrolisi acida catalizzata di esteri:

$$E_s = log\left(k_R / k_{Me}\right)$$

dove R indica il gruppo sostituente e Me il gruppo metile, assunto come riferimento ($E_s(metile) = 0$). Più grande è il gruppo sostituente, minore (più negativo) è il valore di E_S. In questa scala risulta $E_s(idrogeno) = 1.24$. Per questa ragione tutti i valori di E_S sono stati riscalati in modo tale che $E_s(idrogeno) = 0$.
Questo descrittore è correlato sia con i raggi di van der Waals sia con termini elettronici.
Per gli alcani, al fine di eliminare gli effetti elettronici, è stata proposta l'espressione:

$$E_S^0 = E_S + 0.306 \cdot (n_H - 3)$$

dove n_H è il numero di idrogeni nella posizione α del sostituente (n_H = 0,1,2).

🗀 costante sterica di Charton (υ_x)

E' un parametro direttamente correlato ai raggi di van der Waals dei sostituenti ed è definito come

$$\upsilon_x = r_{vdw}(X) - r_{vdw}(H) = r_{vdw}(X) - 1.20$$

🗀 volume caratteristico di McGowan (V_x)

E' un parametro che fornisce una misura del volume di un frammento molecolare, calcolato come somma dei contributi dei singoli atomi e legami.

🗀 rifrazione molare di gruppo (MR_x)

E' la rifrazione molare precedentemente definita, calcolata da Hansch come contributo dei singoli gruppi sostituenti.

🗀 costanti di Swain-Lupton (*F* e *R*)

La costante σ dell'equazione di Hammett contiene informazioni sia sugli effetti induttivi sia sugli aspetti di risonanza del sostituente. Swain e Lupton hanno definito σ come il contributo dei due effetti, induttivo e di risonanza, secondo l'espressione:

$$\sigma = \sigma_I + \sigma_R = aF + bR$$

dove *F* è un *effetto di campo* e *R* un *effetto di risonanza*.
Sono stati pubblicati valori di *F* e di *R* per molti sostituenti.

🗀 termine dipolare (π*)

Il termine dipolare e i due successivi parametri, sono noti come *parametri solvatocromici* in quanto sono stati proposti da Kamlet e Taft nell'ambito degli studi sull'effetto dei solventi sugli spettri elettronici.

🗀 potere elettron-donatore del legame idrogeno (α)

E' una misura dell'acidità di un frammento, cioè della sua propensione a cedere elettroni. Viene sovente utilizzato come somma di tutti i contributi presenti nella molecola.

🗀 potere elettron-accettore del legame idrogeno (β)

E' una misura della basicità di un frammento, cioè della sua propensione ad attrarre elettroni. Viene sovente utilizzato come somma di tutti i contributi presenti nella molecola.

🗀 parametri Sterimol di Verloop (L, B_1, B_2, B_3, B_4, B_5)

I parametri Sterimol di Verloop sono 6 parametri sterici definiti per le dimensioni di un sostituente lungo 5 direzioni. La lunghezza L è determinata lungo l'asse principale del sostituente, in una situazione di minima energia. In direzioni perpendicolari a L vengono calcolati, dai raggi di van der Waals, B_1 e B_4 , che sono rispettivamente la lunghezza minima e massima; B_2 e B_3 sono valori intermedi calcolati perpendicolarmente a B_1 - B_4.

Un caratteristica dei parametri Sterimol è la loro natura direzionale: ad esempio, il rapporto L/B_1 è un'indicazione della deviazione di un sostituente dalla forma sferica.
Verloop ha pubblicato i valori di questi parametri per 243 sostituenti.

🗀 costante idrofobica di Rekker (*f*)

Nell'approccio di Rekker, il logP di un composto viene calcolato come somma dei contributi dei diversi frammenti del composto:

$$\log P = \sum_i a_i f_i$$

dove a_i indica l'occorrenza del frammento *i*-esimo nella struttura del composto.
Alla precedente espressione è stato successivamente un fattore correttivo, multiplo di una costante pari a 0.289:

$$log\ P = \sum_i a_i f_i + \sum_j n_j \cdot 0.289$$

La somma delle costanti di Rekker f_i dei frammenti presenti in una molecola viene utilizzato come indice di idrofobicità.

🗀 costante idrofobica di Leo-Hansch (*f'*)

Basata sulla stessa idea di Rekker, la costante idrofobica di Leo-Hansch si basa sull'aggiunta di fattori di correzione che tengono della complessità molecolare. La relazione fondamentale è

$$log\ P = \sum_i a_i f_i' + \sum_j b_j \cdot F_j$$

dove f'_i sono i contributi dei diversi frammenti, a_i indica l'occorrenza del frammento *i*-esimo nella struttura del composto, *F* i fattori di correzione e b_i i loro coefficienti moltiplicativi.

gruppo	π	σ_m	σ_p	σ^*	E_S	υ_x	*MR*	*F*	*R*
H	0.00	0.00	0.00	0.49	0.00	0.00	0.10	0.00	0.00
CH_3	0.56	-0.07	-0.17	0.00	-1.24	1.00	0.56	0.01	-0.18
C_2H_5	1.02	-0.07	-0.15	-0.10	-1.31	2.00	1.03	0.00	-0.15
OH	-0.67	0.12	-0.37	1.37	-0.55	1.00	0.28	0.33	-0.70
SH	0.39	0.25	0.15	1.68	-1.07	1.20	0.92	0.30	-0.15
NH_2	-1.23	-0.16	-0.66	0.62	-0.61	1.00	0.54	0.08	-0.74
NO	-0.12	0.62	0.91	2.08	-	2.00	0.52	0.49	0.42
NO_2	-0.28	0.71	0.78	4.66	-2.52	3.00	0.74	0.65	0.13
F	0.14	0.34	0.06	3.19	-0.55	0.80	0.09	0.45	-0.39
Cl	0.71	0.37	0.23	2.94	-0.97	1.20	0.60	0.42	-019
Br	0.86	0.39	0.23	2.80	-1.16	1.30	0.89	0.45	-0.22
I	1.12	0.35	0.18	2.22	-1.62	1.70	1.39	0.42	-0.24
CN	-0.57	0.56	0.66	3.64	-0.51	2.00	0.63	0.51	0.15
CHO	-0.65	0.35	0.42	2.15	-	2.00	0.69	0.33	0.09
COOH	-0.32	0.37	0.45	2.94	-	3.00	0.69	0.34	0.11
$COCH_3$	-0.55	0.38	0.50	1.65		3.00	1.12	0.33	0.17
$CONH_2$	-1.49	0.28	0.36	1.66	-	3.00	0.98	0.26	0.10
$CH=CH_2$	0.82	0.06	-0.04	0.52	-3.19	2.00	1.10	0.13	-0.17
C≡CH	0.40	0.21	0.23	2.15	-	2.00	0-95	0.22	0.01

TAB. 13-1

🗀 raggio di van der Waals (r_V)

E' il raggio di van der Waals di un atomo o di un sostituente.

🗀 trasferimento di carica (K_{ap})

E' il contributo di ciascun gruppo al trasferimento di carica.

🗀 **dipolo di gruppo (DP_x)**

E' il contributo di ciascun gruppo al momento dipolare.

🗀 **elettronegatività di gruppo (ELN_x)**

E' l'elettronegativtà di ciascun gruppo.

Esempio

Nella Tab.13-1 sono riportati i valori caratteristici di alcuni descrittori definiti per i più comuni gruppi sostituenti.

13.6 - Descrittori globali geometrici

Sono parametri geometrici di dimensione, volume, superficie e forma relativi all'intera molecola e tengono generalmente conto anche degli aspetti conformazionali della molecola. Questi descrittori rivestono un ruolo di grande importanza per la determinazione di molte proprietà fisiche, termodinamiche, biologiche, tossicologiche, di trasporto, ecc. Altri descrittori di questo tipo sono i descrittori WHIM, diversi descrittori chimico-fisici, descrittori topologici.

🗀 **momenti principali d'inerzia (I_a, I_b, I_c)**

I momenti di inerzia di una molecola dipendono dalle masse di ciascun atomo e dalla geometria molecolare e rappresentano gli aspetti dinamico-rotazionali di un composto. I momenti di inerzia sono definiti da una tripla di valori (uno per ciascun asse di rotazione). Il generico momento di inerzia *I* è definito come:

$$I = \sum_i m_i r_i^2$$

dove m_i sono le masse atomiche e r_i la distanza dell'*i*-esimo atomo perpendicolarmente all'asse di rotazione. Per convenzione i momenti di inerzia sono indicati in modo tale che valga la relazione $I_c \geq I_b \geq I_a$.

🗀 volume molecolare totale (TMV)

Questo descrittore rappresenta il volume molecolare totale di un composto e ha come unità di misura $Å^3$.

🗀 volume di van der Waals (V_{VDW})

E' il volume molecolare calcolato additivamente dai contributi dei volumi di van der Waals degli atomi che compongono la molecola.

🗀 volume molecolare CPK (V_{CPK})

E' il volume molecolare ottenuto direttamente dai modelli CPK di una molecola.

🗀 area superficiale totale (TSA)

Questo descrittore rappresenta l'area superficiale totale di un composto e ha come unità di misura $Å^2$. Esistono molte diverse tecniche per la stima dell'area superficiale che utilizzano, ad esempio, contributi di gruppo, la somma dei quadrati dei raggi atomici covalenti, relazioni con l'energia libera di solvatazione, ecc.

🗀 raggio di rotazione ($\overline{R}$, radius of gyration)

E' il raggio geometrico medio di rotazione di una molecola ed è correlato alla sua dimensione ed alla sua forma. Viene calcolato dai momenti principali di inerzia; per una molecola planare è definito come:

$$\overline{R} = \sqrt{\frac{(I_a \cdot I_b)^{1/2}}{MW}}$$

Per una molecola tridimensionale è definito come:

$$\overline{R} = \sqrt{\frac{2\pi \cdot (I_a \cdot I_b \cdot I_c)^{1/3}}{MW}}$$

dove *MW* è il peso molecolare.

🗀 fattore di forma inerziale (S)

Fattore di forma definito dai momenti principali di inerzia della molecola:

$$S = \frac{I_b}{I_a \cdot I_c}$$

🗀 ovalità (O)

E' un fattore di forma calcolabile una volta noti la superficie ed il volume della molecola. Infatti, dal volume molecolare è possibile calcolare il raggio e da questo la superficie corrispondente al volume, superficie che è la minima superficie possibile a parità di volume. Il rapporto tra la superficie *S* e la minima superficie S_0 viene chiamato *ovalità* ed è quindi definito come:

$$O = \frac{S}{S_0} = \frac{S}{4\pi r^2} = \frac{S}{4\pi \cdot \left(\frac{3V}{4\pi}\right)^{2/3}} \qquad O \geq 1$$

dove il raggio *r* viene ricavato dal volume noto.

🗀 indice di flessibilità di Kier (F)

Indice della flessibilità conformazionale della molecola ed è funzione della più lunga catena lineare (*n*) presente nella molecola e del numero di possibili cammini di lunghezza 4 (^3N_p) nel grafo molecolare (v. indici topologici semplici di Kier-Hall). E' definito come:

$$F = \frac{n}{1 - \frac{1}{^3N_p}}$$

13.7 - Descrittori locali geometrici

Questi descrittori sono definiti dalla geometria tridimensionale della molecola, derivano cioè dalle coordinate cartesiane (x,y,z) degli atomi che la compongono.
I descrittori geometrici presuppongono quindi la conoscenza, da dati sperimentali o da calcoli teorici, della struttura 3D del composto. Per la loro immediatezza sia a livello di disponibilità che a livello di interpretabilità, questi descrittori sono molto utilizzati nei problemi *QSAR*. Tuttavia, i loro limiti evidenti sono che (a) la grandezza selezionata deve essere comune a tutti i composti considerati e che (b) l'informazione portata da ciascuno di questi descrittori resta pur sempre un'informazione locale e parziale dell'intera molecola.

🗀 distanze di legame e interatomiche, angoli di legame e angoli torsionali

Per ogni struttura molecolare 3D, cioè rappresentata dalle coordinate spaziali (x,y,z) di tutti gli atomi che compongono la molecola, è possibile definire delle *coordinate molecolari interne*, cioè le **distanze** tra coppie di atomi (distanze atomiche e interatomiche), gli **angoli piani** tra triple di atomi connessi (*angoli di legame*), gli **angoli diedri** tra quadruple di atomi connessi (*angoli di torsione*).
Questi valori possono essere utilizzati come descrittori locali di un aspetto della struttura molecolare comune a tutti i composti studiati.

🗀 indice di Taillander (∑D)

E' un indice esplicitamente definito per i benzeni sostituiti come somma delle distanze che uniscono i vertici di un poligono, cioè il perimetro del poligono stesso. Ad esempio, in benzeni variamente sostituiti da atomi di Cl, l'indice di Taillander è definito dalla somma delle distanze tra le proiezioni nel piano del benzene dei vertici degli atomi di Cl legati ai carboni benzenici. L'indice di Taillander rappresenta la sezione efficace della molecola nel piano del ciclo aromatico.

13.8 - Descrittori binari

I descrittori binari sono i descrittori più semplici per i quali il descrittore (la variabile) può assumere solo due valori (generalmente 0/1 o -1/+1).
Questi descrittori sono quindi del tipo si/no e indicano generalmente la presenza o l'assenza di una caratteristica definita nel composto. Poiché si tratta di variabili molto grezze, il loro contenuto di informazione è generalmente limitato.
Alcuni esempi di descrittori di questo tipo sono definiti qui di seguito.

🗀 **presenza/assenza di una proprietà definita**

Per ogni tipo di proprietà, predicabile per tutti i composti considerati, la presenza o l'assenza della proprietà viene identificata da un descrittore con valori 1/0.

🗀 **presenza/assenza di un sostituente in un sito**

Per ogni sito di sostituzione, comune a tutti i composti considerati, la presenza o l'assenza di un generico sostituente viene identificata dai valori 1/0. La descrizione del sistema è quindi data da un numero di colonne (le variabili) pari al prodotto del numero di siti di sostituzione per il numero dei sostituenti considerati. Il metodo di Free Wilson utilizza questa descrizione delle molecole per ricercare la correlazione, mediante i metodi di regressione, tra molecole e risposta sperimentale.

🗀 **presenza/assenza di un gruppo funzionale nella molecola**

Per ogni molecola, la presenza o l'assenza di un determinato gruppo funzionale viene identificata dai valori 1/0.

🗀 **conformeri cis/trans**

Per ogni isomero conformazionale, la distinzione tra cis/trans viene definita con un descrittore i cui valori sono generalmente -1/+1 .

🗀 **isomeri destro/levo**

Per ogni composto che presenta isomeri ottici, la distinzione del loro potere rotatorio viene definita da un descrittore che prende i valori -1/+1.

13.9 Descrittori di punteggio

I descrittori di punteggio codificano particolari proprietà di sostanze o molecole con una sequenza di punteggi (spesso numeri interi). Questi descrittori presuppongono che la proprietà descritta sia almeno ordinabile e sono spesso un'estensione dei descrittori binari.

🗀 **indicatori di proprietà**

Descrittori di punteggio sono le categorizzazioni di proprietà molecolari (attività farmacologica, tossicità) o di qualità di prodotti, riconducibili a categorie come *bassa*, *media*, *alta*, descrivibili con punteggi come 1, 2 e 3.

🗀 **punteggi di proprietà**

Descrittori di punteggio sono anche, ad esempio, i punteggi assegnati da giudici selezionati nella valutazione di caratteristiche sensoriali di prodotti alimentari o di prodotti cosmetici mediante *panel test*. I punteggi sono assegnati all'interno di una scala predefinita e possono essere anche numeri non interi.

13.10 - Descrittori enumerativi

I descrittori enumerativi sono descrittori monodimensionali definiti mediante un semplice conteggio (generalmente di atomi o di legami). A dispetto della loro semplicità, questi descrittori hanno per molti problemi un ruolo modellante di un certo interesse in quanto rappresentano modalità diverse per rappresentare un'informazione dimensionale della molecola. La maggior parte di essi è direttamente calcolabile dalla formula bruta o dal grafo molecolare del composto.
Uno sviluppo di questo tipo di informazione si può ritrovare in molti indici topologici.
I più comuni tra questi descrittori sono:

🗀 **numero di atomi totali (*N*)**

E' il numero di atomi totali della molecola ed è un indice dimensionale. Spesso questo indice viene sostituito dal numero di atomi totali esclusi gli atomi di idrogeno.

🗀 **numero di legami totali (*B*)**

E' il numero di legami totali della molecola ed è un indice dimensionale. Spesso questo indice viene sostituito dal numero di legami esclusi i legami con gli atomi di idrogeno.

🗀 **numero di atomi o di sostituenti di un determinato tipo (n_X)**

E' il numero di atomi di una specie chimica o di gruppi funzionali di un determinato tipo. Viene utilizzato in serie omologhe ove è spesso sostitutivo del peso molecolare.

🗀 **numero di legami insaturi (*UB*)**

E' il numero di legami insaturi presenti in in una molecola ed è una misura della reattività della molecola.

🗀 **numero di legami idrogeno (n_H)**

E' il numero di legami idrogeno che una molecola può formare ed puo essere una misura delle forze di interazione intermolecolari.

🗀 **numero di molecole d'acqua della sfera di solvatazione (n_{H_2O})**

E' il numero di molecole d'acqua della sfera di solvatazione di una molecola ed è una misura del suo grado di solvatazione.

13.11 - Descrittori matriciali di connettività

Dalle informazioni contenute nel *grafo molecolare* di una molecola costituita da *N* atomi è possibile definire una matrice di dimensioni *N* x *N* che rappresenta il grafo stesso.
In generale, è possibile definire diversi tipi di *matrici di connettività*, quali le *matrici di adiacenza* o le *matrici delle distanze* (*topologiche*): questi descrittori si presentano quindi in forma di matrici e non presuppongono la conoscenza di alcun dato sperimentale e della geometria 3D della molecola.
Da queste matrici è possibile derivare alcune loro rappresentazioni matematiche quali, ad esempio, il loro **polinomio caratteristico** e il **determinante**.

🗀 matrice di adiacenza (A)

Le matrici di adiacenza sono matrici le cui celle non-diagonali contengono 0 se i due atomi non sono tra loro legati oppure 1 se gli atomi sono legati. E' la più semplice matrice di connettività.

🗀 matrice delle distanze (D)

Le matrici delle distanze contengono in ogni cella non-diagonale un numero intero che rappresenta il numero di passi (i legami) con cui si connettono due atomi, seguendo il percorso più breve. Queste matrici hanno un contenuto di informazione maggiore di quelle di adiacenza in quanto il valore uno rappresenta una coppia di atomi legati, mentre valori maggiori di uno forniscono un'informazione sulla struttura topologica della molecola. Da questo tipo di matrici vengono calcolati molti indici topologici, ma non è possibile distinguere legami multipli poiché ad ogni coppia di atomi legati viene assegnato comunque il valore uno, indipendentemente dall'orine di legame.

🗀 matrice delle distanze multigrafo (M)

Sono una variante delle matrici delle distanze, ove, per gli atomi legati, si rappresenta il legame multiplo con il corrispondente numero di legami, preceduto dal segno meno, consentendo quindi di distinguere coppie di atomi legati dalla distanza di legame tra coppie di atomi non legati. Queste matrici non vengono in generale trattate indipendentemente, ma vengono solo utilizzate per il calcolo di indici topologici.

13.12 - Descrittori topologici

I descrittori topologici sono largamente utilizzati da molti anni nello studio delle relazioni attività-struttura e proprietà-struttura. Essi vengono per lo più calcolati dalle matrici di connettività mediante semplici algoritmi che elaborano l'informazione in esse contenuta e si possono quindi considerare descrittori 2D.
Quasi tutti fondati sui concetti matematici di entropia e contenuto di informazione, gli indici topologici hanno la proprietà di essere invarianti al grafo molecolare, cioè il loro valore è indipendente da come vengono numerati i vertici del grafo (gli atomi). Con pochissime eccezioni, gli indici topologici vengono calcolati, per convenzione, sul grafo molecolare in cui non vengono riportati gli atomi di idrogeno (*hydrogen depleted graphs*).
La loro semplicità di calcolo fa degli indici topologici uno strumento facilmente disponibile per la ricerca di modelli. I loro limiti più evidenti sono legati (a) alla totale mancanza di informazione riguardo alla geometria 3D e agli aspetti conformazionali del composto e (b) alla loro non sempre chiara interpretabilità chimica. A dispetto di questi limiti, in numerosissimi problemi chimici si sono mostrati estremamente efficaci per il loro alto potere modellante.
Gli indici topologici più comuni sono elencati qui di seguito, in accordo con la simbologia di Tab.13-2.
Il grado del vertice di un atomo δ_i è definito come il numero di legami (idrogeni esclusi) che l'atomo forma con gli altri. Il raggio di un atomo r_i è il valore della distanza topologica massima tra l'atomo considerato e gli altri atomi della molecola. La somma delle distanze di un atomo s_i è la somma di tutte le distanze topologiche dell'atomo da tutti gli altri atomi che compongono la molecola. La frequenza di una distanza topologica g_{d} è il numero di volte che un distanza *d* compare nella matrice delle distanze, divisa per due. Le frequenze del grado dei vertici $g_{\underline{\delta}}$ e del raggio g_{r} sono, rispettivamente, le frequenze di ciascun grado e ciascun raggio, calcolati dalla matrice delle distanze **D**.

Simbolo	*Definizione*
D	matrice delle distanze topologiche
B	numero di legami nella molecola
C	numero di cicli nella molecola
N	numero totale di atomi della molecola (idrogeni esclusi)
N'	numero di atomi dello scheletro molecolare (idrogeni inclusi)
n_e	numero di atomi della classe di equivalenza e
p_e	probabilità di atomi della classe di equivalenza e
d_{ij}	distanza topologica tra gli atomi i e j in D
t_e	numero di atomi nella classe di equivalenza topologica e
δ_i	grado del vertice relativo all'atomo i
r_i	raggio del vertice relativo all'atomo i
s_i	somma delle distanze (grado della distanza) relative all'atomo i
g_d	frequenza della distanza topologica d in D
$g_{\underline{\Omega}}$	frequenza del grado dei vertici $\underline{\Omega}$ calcolato da D
g_s	frequenza della somma delle distanze s calcolata da D
g_r	frequenza del raggio r calcolato da D

TAB. 13-2

🗀 indice di legame topologico (K_i)

E' il numero di legami del grafo molecolare con distanza topologica *i*.

🗀 **indice di informazione sulla dimensione (ISIZ)**

Indice puramente dimensionale, definito come

$$ISIZ = N' \log_2 N'$$

🗀 **indice di informazione sulla composizione atomica totale (I_{AC})**

Indice di complessità sui diversi tipi di atomi presenti nella molecola, calcolato sul grafo completo (cioè, idrogeni compresi)

$$I_{AC} = N' \log_2 N' - \sum_e n_e \log_2 n_e$$

🗀 **indice di informazione sulla composizione atomica media ($\bar{I}_{AC}$)**

Indice di complessità media, derivato dall'indice di informazione sulla composizione totale I_{AC}, quindi calcolato dal grafo completo

$$\bar{I}_{AC} = -\sum_e p_e \log_2 p_e$$

🗀 **indice di complessità molecolare di Bertz (MIC)**

Indice di complessità sui diversi tipi di atomi presenti nella molecola, calcolato sul grafo senza gli atomi di idrogeno e definito come

$$MIC = N \log_2 N - \sum_e n_e \log_2 n_e$$

Se tutti gli atomi sono dello stesso tipo (ad esempio, gli idrocarburi), MIC = 0.

🗀 **indici di Zagabria sul grado dei vertici (M_1 e M_2)**

Sono due indici topologici basati sul grado dei vertici, definiti come:

$$M_1 = \sum_i \delta_i^2$$

$$M_2 = \sum_k \delta_i \cdot \delta_j$$

dove gli indici *i* e *j* scorrono sugli *N* atomi che compongono la molecola, mentre l'indice *k* scorre su tutti i *B* legami *i-j* della molecola.

🗀 indice di Wiener (W)

E' la somma delle distanze in un grafo molecolare:

$$W = \frac{1}{2} \cdot \sum_i \sum_j d_{ij}$$

Il valore massimo si ottiene per gli *n*-alcani lineari, mentre i valori minimi si ottengono per le molecole più compatte, ramificate e cicliche.
Il valore medio dell'indice di Wiener totale, è

$$\overline{W} = \frac{2W}{N(N-1)}$$

🗀 indice di Rouvray (*I*)

E' la somma di tutti gli elementi della matrice delle distanze e vale quindi la seguente relazione con l'indice di Wiener: $I = 2W$.

🗀 indice di connettività di Randic (χ)

E' inversamente proporzionale alla somma dei prodotti tra i gradi dei vertici che costituiscono un legame in un grafo molecolare, cioè:

$$\chi = \sum_k \left(\delta_i \cdot \delta_j\right)^{-0.5}$$

dove *k* è l'insieme di tutti i *B* legami *i-j* della molecola. Un'estensione di questo indice è data dagli indici topologici di Kier-Hall. Il valore medio dell'indice di Randic totale, è definito come

$$\overline{\chi} = \chi / B$$

dove *B* è il numero di legami totale.

🗀 indice di distanza topologica media quadratica di Balaban (*D*)

E' un indice della distribuzione media quadratica delle distanza topologiche ed è definito come:

$$D = \sqrt{\frac{\sum_i g_i i^2}{\sum_i g_i}}$$

dove g_i è la frequenza della distanza topologica i.

🗀 indice di connettività di Balaban (J)

Un indice topologico estremamente discriminante, definito come:

$$J = \frac{B}{C+1} \sum_k \left(s_i s_j \right)^{-0.5}$$

dove k scorre su tutti i B legami i-j della molecola, s_i è la somma delle distanze dell'i-esimo atomo da tutti gli altri, B è il numero di legami totale e C è il numero totale di cicli.

🗀 indice centrico di Balaban (C)

E' un indice legato alla forma di molecole acicliche, calcolato mediante una procedura iterativa dove, ad ogni passo, vengono cancellati tutti i vertici terminali. L'indice viene calcolato dalla espressione:

$$C = \sum_k a_k^2$$

dove l'indice k scorre su tutte le iterazioni e a_k è il numero di vertici eliminati ad ogni passo. L'ultimo valore di a_k contiene il numero di vertici restanti.

🗀 indice di informazione centrica ($^{v}\bar{I}_{C,L}$)

E' uno sviluppo dell'indice centrico di Balaban nei termini di indice di informazione.
E' definito come:

$$^{v}\bar{I}_{C,L} = -\sum_k \frac{a_k}{N} log_2 \frac{a_k}{N}$$

dove l'indice *k* scorre su tutte le iterazioni e a_k è il numero di vertici eliminati ad ogni passo. L'ultimo valore di a_k contiene il numero di vertici restanti.

🗀 indice di informazione centrica radiale ($^{v}I_{C,R}$)

E' un indice centrico, proposto da Bonchev, basato sulla partizione dei vertici in classi di equivalenza in funzione dei loro diversi raggi. E' definito come:

$$^{v}\bar{I}_{C,R} = \sum_r \frac{n_r}{N} log_2 \frac{n_r}{N}$$

dove n_r è il numero di atomi di raggio *r*.

🗀 indice di Hosoya (Z)

Detto anche numero di non-adiacenza e correlato alla ramificazione molecolare, l'indice di Hosoya è il numero di modi in cui *k* lati di un grafo molecolare possono essere scelti in modo che nessuna coppia sia adiacente. Viene calcolato secondo l'espressione:

$$Z = \sum_{e=0}^{N/2} n_e$$

dove *N*/2 è il numero intero superiore. I primi due termini della sommatoria sono, per definizione:

$$n_0 = 1 \qquad n_1 = B$$

🗀 indice di informazione di Hosoya ($\bar{I}_Z$)

E' l'indice di informazione derivato dall'indice di Hosoya Z, definito come:

$$\bar{I}_Z = -\sum_{e=0}^{N/2} \frac{n_e}{Z} log_2 \frac{n_e}{Z}$$

🗀 indice delle radici caratteristiche (*Characteristic Root Index*, CRI)

Questo indice è definito come somma degli autovalori positivi (*radici caratteristiche*) ottenuti dalla matrice i cui elementi sono definiti come

$$w_{ij} = \left(\delta_i \delta_j\right)^{-1/2}$$

dove δ_i è il grado del vertice dell'*i*-esimo atomo, non considerando gli atomi di idrogeno.

🗀 indice di Lovasz-Pelikan

E' l'autovalore massimo del polinomio caratteristico ottenuto dalla matrice delle distanze ed è correlato alla ramificazione della molecola.

🗀 indice di Platt (F)

E' il numero di atomi di carbonio che hanno una distanza topologica uguale a tre legami. Indica il grado di ramificazione nella molecola e si calcola determinando per ciascun legame il numero di legami adiacenti e sommando il risultato per tutti i legami.

🗀 indice di Gordon-Scatleburry (N_2)

E' il numero di modi distinti in cui un frammento C-C-C può essere sovrapposto in un grafo molecolare (senza idrogeni). Si dimostra che valgono le seguenti relazioni con l'indice di Platt F e col primo indice di Zagabria M_1:

$$N_2 = F/2 = \left(M_1 - 2B\right)/2$$

dove *B* è il numero totale di legami.

🗀 indice di ramificazione di Austel (S_b)

E' un indice che misura la ramificazione, calcolato dalla struttura molecolare secondo l'espressione:

$$S_b = \sum_i a_i k_i n_i$$

dove l'indice *i* scorre su tutti gli atomi, esclusi gli idrogeni, a_i descrive il contributo della ramificazione all'effetto sterico, k_i è un fattore che tiene conto della dimensione degli atomi legati all'atomo *i* e n_i è il numero dei legami con altri atomi diversi dall'idrogeno.

🗀 indice di informazione totale sull'equivalenza delle distanze (I_D^E)

E' un indice di informazione sulla distribuzione delle distanze, definito come

$$I_D^E = \frac{N(N-1))}{2} log_2 \frac{N(N-1))}{2} - \sum_i g_i \, log_2 \, g_i$$

dove g_i è la frequenza con cui la distanza topologica *i* compare nella matrice delle distanze.

🗀 indice di informazione media sull'equivalenza delle distanze ($\bar{I}_D^E$)

E' l'indice di informazione media sulla distribuzione delle distanze, definito come

$$\bar{I}_D^E = -\sum_i \frac{2g_i}{N(N-1)} log_2 \frac{2g_i}{N(N-1)} = \frac{I_D^E}{\left(\frac{N(N-1)}{2} \right)}$$

dove g_i è la frequenza con cui la distanza topologica *i* compare nella matrice delle distanze.

🗀 indice di informazione totale sulla grandezza delle distanze (I_D^W)

E' un indice di informazione sulla distribuzione della grandezza delle distanze topologiche, definito come

$$I_D^W = W \, log_2 W - \sum_i g_i i \, log_2 \, i$$

dove *W* è l'indice di Wiener e g_i è la frequenza (incidenza) della *i*-esima distanza topologica nel grafo molecolare e *i* la distanza topologica di valore *i*.

🗀 **indice di informazione media sulla grandezza delle distanze ($\bar{I}_D^W$)**

E' l'indice di informazione media sulla distribuzione della grandezza delle distanze topologiche, definito come

$$\bar{I}_D^W = -\sum_i g_i \frac{i}{W} log_2 \frac{i}{W}$$

dove W è l'indice di Wiener e g_i è la frequenza (incidenza) della i-esima distanza topologica nel grafo molecolare e i la distanza topologica di valore i.

🗀 **indice di informazione media sull'equivalenza del grado delle distanze ($\bar{I}_{D,deg}^E$)**

E' il contenuto di informazione media relativo alla distribuzione delle somme delle distanze topologiche ed è definito come

$$\bar{I}_{D,deg}^E = -\sum_s \frac{g_s}{N} log_2 \frac{g_s}{N}$$

dove g_S è la frequenza della somma s delle distanze topologiche.

🗀 **indice di informazione media sulla grandezza del grado delle distanze ($\bar{I}_{D,deg}^W$)**

E' il contenuto di informazione media relativo ai valori delle somme delle distanze topologiche ed è definito come

$$\bar{I}_{D,deg}^W = -\sum_k \frac{s_k}{2W} log_2 \frac{s_k}{2W}$$

dove W è l'indice di Wiener, k scorre su tutti gli atomi del grafo molecolare e s_k è il valore della k-esima somma delle distanze topologiche.

- **indice di informazione media sull'equivalenza del grado dei vertici (${}^{v}\bar{I}^{E}_{adj,deg}$)**

E' il contenuto di informazione media relativo alla distribuzione del grado dei vertici ed è definito come

$$ {}^{v}\bar{I}^{E}_{adj,deg} = -\sum_{\delta} \frac{g_{\delta}}{N} log_2 \frac{g_{\delta}}{N} $$

dove g_{δ} è la frequenza del grado dei vertici.

- **indice di informazione media sulla grandezza del grado dei vertici (${}^{v}\bar{I}^{W}_{adj,deg}$)**

E' il contenuto di informazione media relativo ai valori dei gradi dei vertici ed è definito come

$$ {}^{v}\bar{I}^{W}_{adj,deg} = -\sum_{k} \frac{\delta_k}{2B} log_2 \frac{\delta_k}{2B} $$

dove B è il numero di legami, k scorre su tutti gli atomi del grafo molecolare e δ_k è il valore del k-esimo grado dei vertici.

- **indice di informazione topologica ($\bar{I}_{top}$)**

Indice basato sulla definizione di entropia di Shannon, ove le classi di equivalenza sono derivate dal multigrafo molecolare ripartito in g sottoinsiemi disgiunti, ciascuno costituito dal numero t_e di atomi topologicamente equivalenti:

$$ \bar{I}_{top} = -\sum_{e} \frac{t_e}{N} log_2 \frac{t_e}{N} $$

- **indice di contenuto di informazione sul multigrafo (IC)**

Indice basato sulla definizione di entropia di Shannon, ove le classi di equivalenza sono derivate dal multigrafo molecolare ripartito in g sottoinsiemi disgiunti, ciascuno contenente n_g elementi degli n elementi totali. Ciascuna

ripartizione in classi di equivalenza può essere effettuata considerando contemporaneamente $k + 1$ elementi contigui del grafo molecolare, essendo k l'ordine della partizione ed essendo il suo valore massimo pari al raggio del grafo (la distanza massima).

$$IC_k = -\sum_g p_g \log_2 p_g$$

dove $p_g = n_g / n$.
Per $k = 0$, le classi di equivalenza sono costruite considerando i singoli vertici del grafo (gli atomi); per $k = 1$, le classi di equivalenza sono costruite considerando i singoli legami; per $k \geq 2$, le classi di equivalenza sono costruite considerando i diversi cammini di ordine k.

🗀 **indice di contenuto di informazione totale (TIC)**

Indice topologico che fornisce il contenuto di informazione totale, ricavato dal contenuto di informazione:

$$TIC_k = n \cdot IC_k$$

dove n è la cardinalità della ripartizione del grafo molecolare.

🗀 **indice di contenuto di informazione strutturale (SIC)**

Indice topologico normalizzato ricavato dal contenuto di informazione:

$$SIC_k = \frac{IC_k}{\log_2 n}$$

dove n è la cardinalità della ripartizione del grafo molecolare.

🗀 **indice di contenuto di informazione complementare (CIC)**

Indice topologico ricavato come complemento del contenuto di informazione:

$$CIC_k = \log_2 n - IC_k$$

dove n è la cardinalità della ripartizione del grafo molecolare.

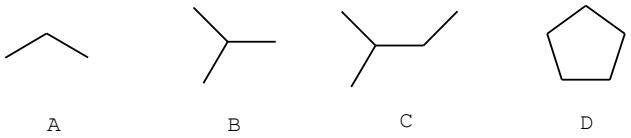
🗀 indice di contenuto di informazione di legame (BIC)

Indice topologico normalizzato sui legami, ricavato dal contenuto di informazione:

$$BIC_k = \frac{IC_k}{log_2 B}$$

🗀 indici topologici semplici di Kier e Hall ($^m\chi_q$)

Molto utilizzati in *QSAR*, sono indici topologici basati sul grado dei vertici δ e sono una generalizzazione dell'indice di Randic, ove *m* (ordine dell'indice) indica il grado di dissezione della topologia molecolare. *q* indica il tipo di frammenti strutturali considerati per il calcolo e può essere *path* (*p*), *cluster* (*c*), *path-cluster* (*pc*) o *chain* (*ch*). Ad esempio,



A - *path* di 2° ordine **B** - *cluster* di 3° ordine
C - *path-cluster* di 4° ordine **D** - *chain* di 5° ordine

L'espressione è definita come.

$$^m\chi_q = \sum_A \left(\delta_i \cdot \delta_j \cdot \delta_k \cdot \ldots \cdot \delta_{m+1}\right)^{-0.5}$$

dove *A* è l'insieme di tutte le sequenze *i-j-k-...-m+1* di lunghezza *m+1*. I frammenti *path* sono definiti anche per ordini tra 0 e 2, mentre, per $m > 2$, il frammento considerato può essere *cluster* o *chain*; per $m > 3$ può essere anche *path-cluster* .
L'indice di Randic coincide con l'indice di Kier-Hall $^1\chi_p$. Ad esempio, gli indici di *path* di ordine 0, 1 e 2 sono:

$$^{0}\chi_p = \sum_i \delta_i^{-0.5}$$
$$^{1}\chi_p = \sum_A \left(\delta_i \delta_j\right)^{-0.5}$$
$$^{2}\chi_p = \sum_A \left(\delta_i \delta_j \delta_k\right)^{-0.5}$$

dove l'indice *i* scorre su tutti gli atomi e *A* è l'insieme di tutte le coppie di legame e di tutte le triple di atomi legati, rispettivamente.

🗀 indici topologici di valenza di Kier e Hall ($^{m}\chi_q^{v}$)

Questi indici sono ottenuti come sviluppo dell'indice di Randic e degli indici topologici semplici di Kier-Hall utilizzando un diverso criterio per la definizione del grado dei vertici δ_i . Il normale grado del vertice δ_i dell'*i*-esimo atomo viene rimpiazzato per gli atomi della prima serie da:

$$\delta_i^{v} = Z_i^{v} - h_i$$

dove Z_i^{v} è il numero di elettroni di valenza dell'*i*-esimo atomo e *h* il numero di atomi di idrogeno legati all'*i*-esimo atomo.
Per gli altri atomi il grado del vertice δ_i dell'*i*-esimo atomo viene rimpiazzato da:

$$\delta_i^{v} = \left(Z_i^{v} - h_i\right) \cdot \left(Z_i - Z_i^{v} - 1\right)$$

dove Z_i è il numero atomico dell'*i*-esimo atomo.
Come per gli indici topologici semplici, *m* (ordine dell'indice) indica il grado di dissezione della topologia molecolare. *q* indica il tipo di frammenti strutturale considerati per il calcolo e può essere *path* (*p*), *cluster* (*c*), *path-cluster* (*pc*) o *chain* (*ch*). I frammenti *path* sono definiti anche per ordini tra 0 e 2, mentre, per $m > 2$, il frammento considerato può essere *path*, *cluster* o *chain*; per $m > 3$ può essere anche *path-cluster*.

$$^{m}\chi_q^{v} = \sum_A \left(\delta_i^{v} \cdot \delta_j^{v} \cdot \delta_k^{v} \cdot \ldots \cdot \delta_{m+1}^{v}\right)^{-0.5}$$

dove *A* è l'insieme di tutte le sequenze *i-j-k-...-m+1* di lunghezza *m+1*.

🗀 indici di stato elettrotopologico di Kier (S)

Sono indici atomici che mimano lo stato elettronico e topologico di ciascun atomo nella molecola. Sono definiti come

$$S_i = I_i + \Delta I_i = I_i + \sum_j \frac{I_i - I_j}{d_{ij}^2}$$

dove d_{ii} è la distanza topologica tra gli atomi *i* e *j*.
L'indice atomico elettrotopologico è definito in funzione del grado del vertice e del grado di valenza del vertice:

$$I_i = \frac{\delta_i^v + 1}{\delta_i}$$

Per gli atomi delle serie superiori a 2 (numeri quantici principali n = 3,4,5), l'espressione è invece:

$$I_i = \frac{(2/n)^2 \delta_i^v + 1}{\delta_i}$$

🗀 parametro di forma di Kier ($^2\kappa$)

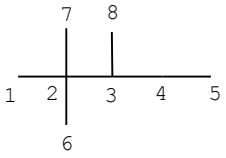
Un parametro di forma, definibile per una serie isomerica, derivato dal grafo molecolare senza tenere conto degli atomi di idrogeno. Esso è basato sul rapporto tra il numero di frammenti a 2 legami in un grafo (^2p) in rapporto al massimo numero possibile di grafi isomerici a stella (massima ramificazione, $^2p_{max}$) e il minimo numero grafi isomerici lineari (nessuna ramificazione, $^2p_{min}$).

$$^2\kappa = 2\frac{^2p_{max}\,^2p_{min}}{(^2p)^2}$$

L'indice può essere normalizzato sul numero totale di atomi nella molecola.
Questo indice può essere utilizzato sia per rappresentare l'intera molecola che per suoi frammenti o per i sostituenti.

Esempio

Consideriamo il composto 2,2,3-trimetil-pentano, la cui formula bruta è: C_8H_{18} e il cui grafo molecolare è il seguente:



Per questa molecola calcoliamo alcuni degli indici precedentemente definiti.

$N' = 26$; $N = 8$; $B = 7$

Matrice delle distanze

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**δ**	**s**	**r**
1	-	1	2	3	4	2	2	3	**1**	**17**	**4**
2	1	-	1	2	3	1	1	2	**4**	**11**	**3**
3	2	1	-	1	2	2	2	1	**3**	**11**	**2**
4	3	2	1	-	1	3	3	2	**2**	**15**	**3**
5	4	3	2	1	-	4	4	3	**1**	**21**	**4**
6	2	1	3	3	4	-	2	3	**1**	**17**	**4**
7	2	1	3	3	4	2	-	3	**1**	**17**	**4**
8	3	2	1	2	3	3	3	-	**1**	**17**	**3**

distanze	*1*	*2*	*3*	*4*
frequenza delle distanze	7	10	8	3
frequenza del grado dei vertici	5	1	1	1

grado delle distanze	*11*	*15*	*17*	*21*
frequenza del grado delle distanze	2	1	4	1

$$ISIZ = N' log_2 N' = 3.3219 \cdot 26 \cdot log\, 26 = 122.21$$

$$I_{AC} = N' \log_2 N' - \sum_e n_e \log_2 n_e = 3.3219 \cdot \ 26 \log 26 - 8 \log 8 - 18 \log 18 \ = 23.153$$

$$\bar{I}_{AC} = -\sum_e p_e \log_2 p_e = -3.3219 \cdot \left[\frac{8}{26} \log \frac{8}{26} + \frac{18}{26} \log \frac{18}{26} \right] = 0.890$$

$$MIC = N \log_2 N - \sum_e n_e \log_2 n_e = 3.3219 \cdot \ 8 \log 8 - 8 \log 8 \ = 0$$

$$\mathrm{M}_1 = \sum_i \delta_i^2 = 5 \cdot 1^2 + 2^2 + 3^2 + 4^2 = 34$$

$$\mathrm{M}_2 = \sum_k \delta_i \cdot \delta_j = 3 \cdot (1 \cdot 4) + 4 \cdot 3 + 3 \cdot 2 + 2 \cdot 1 + 3 \cdot 1 = 35$$

$$W = \frac{1}{2} \cdot \sum_i \sum_j d_{ij} = \frac{4 \cdot 17 + 2 \cdot 11 + 15 + 21}{2} = 63$$

$$\overline{W} = \frac{2W}{N(N-1)} = \frac{2 \cdot 63}{8 \cdot 7} = 2.250$$

$$\chi = \sum_k \left(\delta_i \cdot \delta_j \right)^{-0.5} = 1/\sqrt{1 \cdot 2} + 1/\sqrt{1 \cdot 3} + 3/\sqrt{1 \cdot 4} + 1/\sqrt{2 \cdot 3} + 1/\sqrt{3 \cdot 4} = 3.481$$

$$\overline{\chi} = \chi / B = 3.481 / 7 = 0.497$$

$$D = \sqrt{\frac{\sum_i g_i i^2}{\sum_i g_i}} = \sqrt{\frac{4 \cdot 1^2 + 2^2 + 3^2 + 4^2}{4 + 1 + 1 + 1}} = 2.171$$

$$J = \frac{B}{C+1} \sum_k \left(s_i s_j \right)^{-1/2} =$$

$$= \frac{7}{1} \cdot \ 4 \cdot (17 \cdot 11)^{-0.5} + (11 \cdot 11)^{-0.5} + (11 \cdot 15)^{-0.5} + (15 \cdot 21)^{-0.5} \ = 3.623$$

$$C = \sum_k a_k^2 = 5^2 + 2^2 + 1 = 30$$

13.13 - Descrittori di correlazione strutturale

Descrittori di questo tipo si ottengono utilizzando gli **autocorrelogrammi** (o **funzioni di autocorrelazione**), cioè sequenze di numeri che ottenuti dividendo in classi discrete (*bins*) una descrizione molecolare di tipo 2D (topologica) o 3D.

🗀 autocorrelogrammi 2D

Autocorrelogramma calcolato dalla matrice topologica delle distanze, secondo la formula:

$$bin(n) = \sum_{inj} w_i w_j$$

dove *bin(n)* è il valore assunto dalla classe *n* e l'indice *inj* indica tutte le coppie di atomi *i-j* separati da *n* atomi (a distanza topologica *n*).
w_i sono dei pesi, uguali ad uno nel caso di un semplice conteggio oppure uguali alla carica, al volume di van der Waals, all'elettronegatività, alla valenza dell'atomo, ecc.

🗀 autocorrelogrammi 3D

Autocorrelogramma calcolato dalla matrice delle distanze interatomiche, in sostituzione delle distanze topologiche, secondo la formula:

$$bin(r) = \sum_{irj} w_i w_j$$

dove *bin(r)* è il valore assunto dalla classe *r* e l'indice *irj* indica tutte le coppie di atomi *i-j* separati da distanze comprese tra *r-1* e *r*.
w_i sono dei pesi, uguali ad uno nel caso di un semplice conteggio oppure uguali alla carica, al volume di van der Waals, all'elettronegatività, alla valenza dell'atomo, ecc., nei casi in cui si vogliano pesare diversamente le distanze.

Esempio

Calcoliamo i valori dei primi 3 termini della funzione di autocorrelazione 2D per la molecola 2,2,3-trimetil-pentano. Il peso viene assunto unitario.

bin(0) = 5 (1x1) + (4x4) + (3x3) + (2x2) = 34

bin(1) = 3 (1x4) + (4x3) + (3x1) + (3x2) + (1x2) = 35

Si osservi che bin(0) e bin(1) coincidono rispettivamente con gli indici topologici M_1 e M_2.
Il calcolo di bin(2) si esegue effettuando la somma di tutti i prodotti del grado dei vertici tra le coppie di atomi separati da un solo atomo, cioè tra le seguenti coppie di atomi: 1-3, 1-6, 1-7, 2-4, 2-8, 3-5, 3-6, 3-7, 4-8, 6-7.

bin(2) = (1x3) + (1x1) + (1x1) + (4x2) + (4x1) + (3x1) + (3x1) + (3x1) + (2x1) + (1x1) = 29

Il vettore di ordine 3 che descrive il composto è quindi il seguente: [34, 35, 29].

13.14 - Descrittori WHIM

I descrittori WHIM (*Weighted Holistic Invariant Molecular descriptors*) sono descrittori che rappresentano la struttura tridimensionale della molecola, cioè tengono conto della sua configurazione o conformazione nello spazio. Per il calcolo di questi descrittori è quindi necessario conoscere le coordinate spaziali x-y-z degli atomi. Il calcolo prevede la centratura della molecola nel suo baricentro e, mediante l'analisi delle componenti principali (PCA), la determinazione degli assi principali. La rappresentazione della molecola avviene mediante una proiezione degli atomi sui tre assi principali (1,2,3) e sul calcolo di alcuni parametri statistici (gli indici WHIM) per ciascuno di questi assi.
Il calcolo di questi descrittori prevede la definizione di **pesi** per ciascun atomo: a secondo del peso utilizzato, viene estratta dalla configurazione spaziale della molecola un tipo di informazione diversa. Attualmente, sono stati proposti 6 schemi di pesatura:

a) **pesi unitari** (**U**), cioè tutti i pesi sono uguali ad uno e non si ha quindi alcuna distinzione tra atomi chimicamente diversi: l'informazione estratta è di carattere puramente geometrico.

b) **masse atomiche** (**M**), cioè ciascun atomo è pesato dalla propria massa atomica e le direzioni principali ottenute coincidono con quelle degli assi principali d'inerzia. L'informazione estratta è relativa al comportamento inerziale della molecola.

c) **volumi di van der Waals** (**V**), cioè ciascun atomo è pesato dal proprio volume di van der Waals e l'informazione chimica viene ricercata lungo assi principali volumici.

d) **elettronegatività atomica di Mulliken** (**E**), cioè ciascun atomo è pesato dal proprio valore di elettronegatività e l'informazione chimica viene ricercata lungo assi principali di polarizzazione. Con questo tipo di pesi, si assume che le cariche atomiche siano fisse, cioè indipendenti dall'intorno chimico di ciascun atomo.

e) **indici di polarizzabilità atomica** (**P**), cioè ciascun atomo è pesato dalla sua polarizzabilità.

f) **indici elettrotopologici di Kier** (**S**), cioè ciascun atomo è pesato dalla carica calcolata come indice elettrotopologico di Kier (v. descrittori topologici), cioè le cariche atomiche variano con l'intorno chimico di ciascun atomo. Per trasformare in pesi statistici gli indici atomici elettrotopologici originali *S* (che in alcuni casi presentano anche valori leggermente negativi) viene adottata la seguente trasformazione lineare:

$$S_i' = S_i + 7 \qquad 0 < S_i'$$

Nella Tab.13-3 sono riportati i valori delle grandezze atomiche utilizzate (casi b - e) e i corrispondenti valori dei pesi, ottenuti scalando i valori originali rispetto all'atomo di carbonio.
Per ciascun tipo di peso, vengono calcolati i seguenti parametri statistici sulle proiezioni degli atomi (gli *scores* $\mathbf{t}_m$ con m = 1,2,3) lungo ciascuna delle 3 componenti principali.

Si ottengono così i **descrittori WHIM direzionali**, suddivisi nei seguenti gruppi:

🗀 **descrittori di dimensione (λ_1 λ_2 λ_3)**

Sono gli autovalori ottenuti dall'analisi delle componenti principali (sono i momenti centrali del secondo ordine e coincidono con la varianza di ciascuna componente principale). Sono in relazione con le dimensioni molecolari (**size factor**).

🗀 descrittori di forma (ϑ_1 ϑ_2 ϑ_3)

Ciascuno di questi descrittori è definito come il rapporto tra l'autovalore relativo ad una direzione e la somma degli autovalori:

$$\vartheta_m = \frac{\lambda_m}{\sum_m \lambda_m} \qquad m = 1,2,3$$

La loro somma è ovviamente uguale a uno: per questa ragione, nella costruzione di modelli matematici vengono presi in considerazione solo ϑ_1 e ϑ_2.
Sono in relazione con la forma della molecola (**shape factor**): infatti $\vartheta_3 = 0$ indica, ad esempio, una molecola planare ($\lambda_3 = 0$), mentre $\vartheta_1 = 0.5$ e $\vartheta_2 = 0.5$ indicano una molecola planare simmetrica come il benzene.

ID	*Massa*		*Volume VdW*		*Elettroneg.*		*Polarizzab.*	
	M	*M/C*	*V*	*V/C*	*E*	*E/C*	*P*	*P/C*
H	1.01	0.084	6.709	0.299	2.592	0.944	0.667	0.379
B	10.81	0.900	17.875	0.796	2.275	0.828	3.030	1.722
C	12.01	1.000	22.449	1.000	2.746	1.000	1.760	1.000
N	14.01	1.166	15.599	0.695	3.194	1.163	1.100	0.625
O	16.00	1.332	11.494	0.512	3.654	1.331	0.802	0.456
F	19.00	1.582	9.203	0.410	4.000	1.457	0.557	0.316
Al	26.98	2.246	36.511	1.626	1.714	0.624	6.800	3.864
Si	28.09	2.339	31.976	1.424	2.138	0.779	5.380	3.057
P	30.97	2.579	26.522	1.181	2.515	0.916	3.630	2.063
S	32.07	2.670	24.429	1.088	2.957	1.077	2.900	1.648
Cl	35.45	2.952	23.228	1.035	3.475	1.265	2.180	1.239
Fe	55.85	4.650	41.052	1.829	2.000	0.728	8.400	4.773
Co	58.93	4.907	35.041	1.561	2.000	0.728	7.500	4.261
Ni	58.69	4.887	17.157	0.764	2.000	0.728	6.800	3.864
Cu	63.55	5.291	11.494	0.512	2.033	0.740	6.100	3.466
Zn	65.39	5.445	38.351	1.708	2.223	0.810	7.100	4.034
Br	79.90	6.653	31.059	1.384	3.219	1.172	3.050	1.733
Sn	118.71	9.884	45.830	2.042	2.298	0.837	7.700	4.375
I	126.90	10.566	38.792	1.728	2.778	1.012	5.350	3.040
Gd	157.25	13.093	72.776	3.242	2.000	0.728	23.500	13.352

TAB. 13-3

🗀 **descrittori di simmetria (**γ_1 γ_2 γ_3**)**

I descrittori WHIM di simmetria sono calcolati da una funzione inversa di un indice di informazione di simmetria lungo ciascuna componente, secondo l'espressione

$$\gamma'_m = -\left[\frac{n_s}{n}\cdot log_2\frac{n_s}{n} + n_a\cdot\left(\frac{1}{n}\cdot log_2\frac{1}{n}\right)\right] \qquad \gamma_m = \frac{1}{1+\gamma'_m} \qquad 0 < \gamma_m \le 1$$

dove n_s è la somma di tutti i gruppi di atomi con gli stessi valori opposti degli scores lungo la *m*-esima componente o un valore dello score uguale a zero e n_a è il numero di atomi con un valore dello score senza un corrispondente valore opposto lungo la *m*-esima componente. *n* è il numero totale di atomi. Questi descrittori rappresentano la simmetria molecolare (**symmetry factor**) lungo ciascuna componente; un valore di γ uguale a uno corrisponde a una proiezione in cui tutti gli atomi sono disposti simmetricamente rispetto al centro di un asse principale $n_s = n$ e $n_a = n$.

🗀 **descrittori di distribuzione o densità (**η_1 η_2 η_3**)**

Questo gruppo di descrittori viene ricavato dalla definizione statistica di curtosi (funzione dei momenti centrali del quarto ordine) e è definito come l'inverso della curtosi:

$$\kappa_m = \frac{\sum_i t_i^4}{s^4} \qquad \eta_m = \frac{1}{\kappa_m} \qquad 0 \le \eta_m \le 1 \qquad m = 1,2,3$$

Poiché κ rappresenta un indice che assume il valore minimo (uno) per distribuzioni bimodali ed il valore massimo per distribuzioni unimodali a picco (infinito), η_m rappresenta la proiezione lungo la componente *m*-esima della quantità di spazio molecolare interno non occupato per atomo (**emptiness factor**).

🗀 **descrittori WHIM non-direzionali (**T, A, V, *G*, ω, *K*, *D***)**

Per ogni schema di pesatura, da ciascun gruppo dei descrittori direzionali precedenti è possibile definire degli indici globali, cioè non-direzionali.
Dai descrittori dimensionali (λ_m) vengono definiti tre descrittori globali non-direzionali per ogni schema di pesatura:

$$T = \lambda_1 + \lambda_2 + \lambda_3 \qquad \textit{termine dimensionale lineare}$$

$$A = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 \qquad \textit{termine dimensionale quadratico}$$

$$V = \prod_{m=1}^{3}(1+\lambda_m) - 1 = T + A + \lambda_1\lambda_2\lambda_3 \qquad \textit{volume totale}$$

Un primo fattore di forma è il *fattore acentrico* (**acentric factor**) definito come differenza tra i fattori di forma lungo la prima e la terza componente principale:

$$\omega = \vartheta_1 - \vartheta_3 \qquad 0 \le \omega \le 1$$

Il fattore acentrico è quindi uguale a uno per molecole lineari e uguale a zero per molecole di forma sferica. Si noti che per molecole planari $\omega = \vartheta_1$, essendo $\vartheta_3 = 0$.
Il fattore acentrico è stato recentemente sostituito da un fattore di forma di carattere più generale e definito dalla seguente espressione:

$$K = \frac{\sum_m \left| \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{3} \right|}{4/3}$$

Questo *descrittore globale di forma* viene utilizzato come misura della quantità di correlazione nei dati (capitolo 3). *K* assume valore zero per una molecola sferica, uno per una molecola ideale lineare, 0.5 per molecola planare in cui la seconda dimensione è maggiore o uguale alla metà della prima dimensione.
Dai descrittori di simmetria direzionali viene definito un *descrittore di simmetria globale* come

$$G = (\gamma_1 \cdot \gamma_2 \cdot \gamma_3)^{1/3}$$

dove *G* è la media geometrica delle simmetrie direzionali.

Come descrittore derivato dall'inverso delle curtosi, viene definito un *indice globale di distribuzione* come

$$D = \left(\eta_1 + \eta_2 + \eta_3\right)/3$$

13.15 - Descrittori differenziali

Quando la relazione specifica tra farmaco e recettore è di fondamentale importanza e sia nota la struttura del sito recettoriale (la cavità) oppure un composto di riferimento che abbia la massima attività farmacologica, è possibile calcolare dei descrittori che valutino la minima differenza tra il composto di riferimento e gli altri composti omologhi della serie studiata.

🗀 Minima differenza sterica (MSD)

Secondo questo metodo ogni struttura viene sovrapposta alla struttura di riferimento, trascurando le piccole diversità geometriche e conformazionali e considerando soltanto gli atomi non sovrapponibili. Il valore di *MSD* per ogni composto viene calcolato dal numero di atomi non sovrapponibili, pesando diversamente gli atomi delle diverse righe della tavola degli elementi.

🗀 Minima differenza topologica (MTD)

E' una versione modificata del metodo precedente ove le sovrapposizione ottimale di tutte le molecole a quella di riferimento genera una supermolecola. Questa supermolecola origina un struttura topologica che viene assunta come riferimento.
Per ogni molecola è quindi possibile costruire un vettore x_{ii} di lunghezza pari al numero N di vertici della struttura topologica di rierimento, i cui elementi sono uguali ad uno per tutti gli atomi della molecola considerata che coincidono con i vertici della struttura di riferimento. Diversamente gli elementi del vettore sono uguali a zero.
Il valore del descrittore *MTD* per ogni molecola viene calcolato dalla funzione:

$$MTD_i = s + \sum_{j=1}^{N} \varepsilon_j x_{ij}$$

dove *s* è il numero totale di vertici della molecola di riferimento. I coefficienti ε_j sono uguali a -1 se i vertici corrispondono a quelli della molecola di riferimento, sono uguali a +1 in caso contrario. Questi ultimi possono essere eventualmente

posti uguali a zero, anzichè uguali ad uno, se si ritiene che i corrsipondenti atomi non influiscano sulla risposta biologica.

🗀 Analisi della forma molecolare (MSA)

Simile ai due approcci precedenti, il tipo di descrittore calcolato mediante l'analisi della forma molecolare si basa sul calcolo, effettuato con una procedura abbastanza complessa, del volume totale V_0(**S**, *i*) definito dalla sovrapposizione del composto di riferimento **S** con ciascun *i*-esimo composto della serie studiata. Come descrittori di forma molecolare vengono definite, in alternativa, le due funzioni:

$$S_0 = V_0^{2/3} \text{ oppure } L_0 = V_0^{1/3}$$

13.16 - Descrittori cromatografici

I metodi cromatografici si suddividono in diverse tecniche, quali, ad esempio, la cromatografia su strato sottile (TLC) e su carta (PC), la gascromatografia (GC), la cromatografia liquida ad alte prestazioni (HPLC). Qui di seguito si riportano alcuni dei descrittori più comuni relativi a queste tecniche cromatografiche.

🗀 indice di ritenzione R_f

L'indice di ritenzione $\mathbf{R_f}$ ottenuto da cromatografia a strato sottile (TLC) o su carta (PC) è definito come il rapporto tra la distanza percorsa dall'*i*-esimo componente d_i e la distanza percorsa dal fronte dell'eluente d_s :

$$R_f = \frac{d_i}{d_s}$$

🗀 indice di ritenzione di Bate-Smith e Westall (R_M)

Un parametro di ritenzione derivato da $\mathbf{R_f}$, è definito come

$$R_M = log\left(\frac{1}{R_f} - 1\right)$$

🗀 tempo di ritenzione (t_R)

Il tempo di ritenzione viene calcolato in gascromatografia ed è definito come l'intervallo di tempo intercorrente tra l'istante di introduzione del campione e quello della comparsa del massimo del picco corrispondente al campione stesso. Per una serie di composti omologhi, è possibile definire un tempo di ritenzione relativo.

🗀 fattore di risposta (RF)

Il fattore di risposta è una misura quantitativa utilizzata in gascromatografia e definita come

$$RF = \frac{w_S \cdot A_C}{w_C \cdot A_S}$$

dove w_S e w_C sono rispettivamente i pesi dello standard di taratura e del composto di cui si vuole calcolare *RF*. A_S e A_C sono rispettivamente le aree dei picchi cromatografici dello standard e del composto. Il valore di RF dello standard viene assunto uguale a 1; un tipico composto utilizzato come standard è il *n-eptano*.
Il peso incognito w_X del composto X viene quindi determinato mediante l'espressione

$$w_X = \frac{w_S \cdot A_X}{A_S \cdot RF_X}$$

dove A_X è l'area misurata per il composto *X* di cui è stato precedentemente determinato il valore di *RF*.

🗀 indice di ritenzione di Kovats (I_k)

L'*indice di ritenzione di Kovats* è un indice gas-cromatografico relativo alla serie omologa delle n-paraffine, funzione dei logaritmi dei tempi di ritenzione e definito come:

$$I_k(i) = 100 \times \frac{log\, t_{R(i)} - log\, t_{R(z)}}{log\, t_{R(z+1)} - log\, t_{R(z)}} + 100 \times z$$

dove $t_{R(i)}$ è il tempo di ritenzione della sostanza studiata $t_{R(z)} < t_{R(i)} < t_{R(z+1)}$, $t_{R(z)}$ e $t_{R(z+1)}$ sono i tempi di ritenzione corretti degli idrocarburi lineari con z e z + 1 atomi di carbonio. L'indice di Kovats permette di valutare il comportamento cromatografico di una sostanza che corrisponde convenzionalmente ad un idrocarburo lineare con un numero di atomi di carbonio compreso tra z e z + 1.

🗀 **fattore di capacità (*k'*)**

Grandezza definita in HPLC come

$$k' = \frac{t_R - t_M}{t_M} = \frac{V_R - V_M}{V_M}$$

dove t_R e V_R sono rispettivamente il tempo ed il volume di ritenzione del soluto cromatografato e t_M e V_M sono rispettivamente il tempo ed il volume di ritenzione del soluto non trattenuto dalla colonna (generalmente il solvente). Il logaritmo del fattore di capacità *k'* può essere considerato come una quantità analoga al parametro R_M utilizzato in TLC.

13.17 - Descrittori spettroscopici

Descrittori di questo tipo possono essere singoli componenti di uno spettro o le parti più significative dell'intero spettro. Mentre nel primo caso, abbiamo a che fare con singoli descrittori specifici, nel secondo caso, cui si perviene mediante la digitalizzazione dello spettro, otteniamo descrittori vettoriali ad alta dimensionalità.

🗀 **spostamento chimico (*chemical shift,* δ)**

E' una grandezza misurata in Risonanza Nucleare Magnetica (NMR), è applicabile ai nuclei ^{1}H, ^{13}C, ^{19}F, ^{31}P e fornisce informazioni sugli effetti elettronici quantitativi.

E' lo spostamento del segnale di risonanza di un composto immerso in un campo magnetico rispetto ad una sostanza di riferimento ed è definito come:

$$\delta = \frac{\nu_{ref} - \nu}{\nu_{ref}} \times 10^6$$

Lo spostamento chimico è di norma espresso in ppm ed è generalmente compreso tra 0 e 10. La relazione che lega i ppm ai cicli-per-secondo (cps) è $ppm = cps / Mc$, dove Mc sono i megacicli a cui lavora lo strumento.
Una misura derivata è definita come il complemento a 10 di δ, cioè $\tau = 10 - \delta$.

🗀 **relassività (R)**

E' una proprietà intensiva caratteristica di ogni sostanza paramagnetica. Misura la capacità di modificare la velocità del processo di diseccitazine (*rilassamento*) di un nucleo atomico dopo che questo è stato immerso in un campo magnetico ed eccitato con una opportuna radiofrequenza. Unità di misura sono $mM^{-1}s^{-1}$.

🗀 **coefficiente di estinzione molare (ε_λ)**

Dalla legge di Lambert-Beer, il coefficiente di estinzione molare è una misura della tendenza di un composto ad assorbire la radiazione di lunghezza d'onda λ ed è definito come:

$$\varepsilon_\lambda = \frac{D_\lambda}{c \cdot l}$$

dove D_λ, c e l sono rispettivamente l'assorbanza, la concentrazione della sostanza e lo spessore del campione. L'assorbanza è il logaritmo del rapporto tra l'intensità della luce trasmessa e quella della luce assorbita; è quindi un numero puro. Essendo, di norma, la concentrazione espressa in moli/litro e lo spessore in cm, il coefficiente di estinzione molare ha dimensioni $l / mole\ cm$.

🗀 **frequenza di massimo assorbimento/emissione (ν)**

Corrisponde alla frequenza di massimo assorbimento (generalmente di spettri IR) o di massima emissione (generalmente di spettri UV) di un composto.

🗀 **spettri digitalizzati**

Sono descrittori ad alta dimensionalità definiti dalle intensità di assorbimento in una regione selezionata dello spettro di un composto. Lo spettro del composto viene digitalizzato e normalmente rappresentato da molte centinaia di intensità dei segnali ottenuti.

13.18 - Descrittori di interazione a campi scalari

Sono descrittori ad alta dimensionalità ottenuti mediante una procedura di *immersione* della molecola in una *griglia* (*grid*) tridimensionale. Una volta effettuata questa operazione, in ogni punto della griglia (spesso molte migliaia di punti) vengono calcolate delle *energie di interazione* tra un test, definito *probe* (ad esempio, un atomo di idrogeno, un metile, una molecola d'acqua, una coppia di elettroni, etc.), e la molecola considerata. Il risultato finale consiste nell'associare ad ogni punto della griglia uno scalare che definisce il particolare valore del potenziale di interazione calcolato in quel punto, ottenendo alla fine un campo di scalari.
Il problema più evidente nell'utilizzare questo tipo di descrittori ai fini *QSAR* è il loro elevatissimo numero. Si pongono quindi problemi di selezione di sottoinsiemi informativi di questi potenziali che nello stesso tempo siano confrontabili per tutte le molecole considerate.
Una risposta a questi problemi viene data dai descrittori *G-WHIM* (v. oltre).
A secondo dei potenziali di interazione definiti (e del metodo utilizzato per il loro calcolo), possiamo avere campi scalari diversi, i più comuni dei quali sono:

🗀 **campi scalari elettrostatici**

Campi scalari ottenuti dall'interazione tra la molecola considerata e un *probe* carico, atto a simulare un'interazione puramente elettrostatica. Ad esempio, il *probe* H^+. Le mappe di densità ottenute in questo caso vengono sovente chiamate *mappe di densità elettronica* (MEP).

🗀 **campi scalari sterici**

Campi scalari ottenuti dall'interazione tra la molecola considerata e un *probe* privo di carica, atto a simulare un'interazione puramente sterica.

🗀 **campi scalari di van der Waals**

Campi scalari ottenuti dall'interazione tra la molecola considerata e un *probe* neutro, atto a simulare un'interazione di non-legame. Ad esempio, il *probe* metile.

🗀 **campi scalari di legame idrogeno**

Campi scalari ottenuti dall'interazione tra la molecola considerata e un *probe* in grado di simulare un'interazione dovute alla presenza di legami idrogeno. Ad esempio, il *probe* H_2O.

🗀 **descrittori G-WHIM**
(*Grid-Weighted Holistic Invariant Molecular descriptors*)

Basati sull'approccio con cui vengono calcolati i descrittori WHIM, si differenziano da questi ultimi essenzialmente perchè la molecola è descritta dai punti della griglia e non dagli atomi che la rappresentano.
Infatti, mentre nel calcolo dei descrittori WHIM, l'analisi delle componenti principali viene effettuata su una matrice di covarianza ottenuta pesando le coordinate degli atomi con pesi quali le masse atomiche, i volumi atomici di van der Waals, le polarizzabilità atomiche, ecc., nel calcolo dei descrittori G-WHIM, l'analisi delle componenti principali viene effettuata su una matrice di covarianza ottenuta pesando le coordinate della griglia con i corrispondenti valori dei campi scalari ottenuti (separatamente per potenziali negativi e positivi, per ogni *probe*). I punti della griglia possono essere tutti i punti considerati, un sottoinsieme di essi al di sopra di una determinata energia di interazione, una superficie di potenziale opportunamente definita, o qualsiasi altro criterio chimicamente significativo.
In ogni caso, dall'analisi in componenti principali si ottiene una rappresentazione sintetica e univoca del campo molecolare (rappresentazione indipendente dall'orientazione della molecola nell griglia), definita da una serie di descrittori di dimensione, forma, simmetria e densità (G-WHIM).

13.19 - Descrittori di similarità molecolare

L'analisi di similarità tra un composto di riferimento e altri composti omologhi viene normalmente effettuata valutando la distanza tra il composto di riferimento e gli altri composti e trasformando la misura della distanza in una misura di similarità. La similarità dipende quindi dalla scelta della metrica per

misurare la distanza (v. cap. 4), dai descrittori utilizzati nel rappresentare i composti e dalla *funzione di similarità* prescelta.
Nel caso in cui la descrizione di un composto sia definita da una sola proprietà determinata nei punti di una griglia in cui il composto è immerso (v. descrittori di interazione a campi scalari), è possibile valutare la similarità tra due composti mediante il confronto tra le coppie di valori della proprietà assunti in ogni punto della griglia. Le misure di similarità tra il composto di riferimento *A* e gli altri *n* - 1 composti possono essere ordinate e confrontate con la risposta sperimentale di interesse. Una recente alternativa si propone di mettere in relazione, mediante l'utilizzo di reti neurali, l'intera matrice di similarità tra tutte le coppie di composti (matrice di dimensione *n* x *n*) con la risposta sperimentale.
Le più comuni misure di similarità molecolare quando una proprietà molecolare *P* viene valutata in una griglia di N punti sono definite qui di seguito.

🗀 indice di Carbò (R)

Proposto inizialmente per confrontare la densità elettronica molecolare, è stato esteso al confronto tra qualsiasi proprietà misurata *P* su una griglia di N punti. La similarità tra il composto *A* e il composto *B* è definita dall'espressione:

$$\mathrm{R}_{AB} = \frac{\sum_{k=1}^{\mathrm{N}} P_A P_B}{\left(\sum_{k=1}^{\mathrm{N}} P_A^2\right)^{1/2} \cdot \left(\sum_{k=1}^{\mathrm{N}} P_B^2\right)^{1/2}}$$

Questa misura di similarità è sensibile soprattutto alla forma della distribuzione della proprietà nel campo.

🗀 indice di Hodgkin (H)

L'indice di similarità di Hodgkin è un'alternativa all'indice di Carbò ed è maggiormente legato alle dimensioni della distribuzione della proprietà nel campo su una griglia di N punti. E' definito come

$$\mathrm{H}_{AB} = \frac{2 \cdot \sum_{k=1}^{\mathrm{N}} P_A P_B}{\sum_{k=1}^{\mathrm{N}} P_A^2 + \sum_{k=1}^{\mathrm{N}} P_B^2} = 1 - \frac{d_{AB}^2}{\sum_{k=1}^{\mathrm{N}} P_A^2 + \sum_{k=1}^{\mathrm{N}} P_B^2}$$

dove d_{AB}^2 è il quadrato della distanza euclidea tra gli oggetti *A* e *B*.

🗀 indice di similarità lineare (L)

E' un indice di similarità utilizzato soprattutto per valutare la similarità di campi di interazione elettrostatici in una griglia di N punti ed è definito come

$$L_{AB} = \frac{\sum_{k=1}^{N}\left(1 - \frac{|P_A - P_B|}{max(|P_A|,|P_B|)}\right)}{N}$$

dove la proprietà *P* viene valutata in ogni *k*-esimo punto per i composti *A* e *B*.

🗀 indice di similarità esponenziale (E)

E' un indice di similarità utilizzato soprattutto per valutare la similarità di campi di interazione elettrostatici in una griglia di N punti, in cui l'andamento della similarità varia in modo esponenziale, ed è definito come

$$E_{AB} = \frac{\sum_{k=1}^{N} exp^{-\frac{|P_A - P_B|}{max(|P_A|,|P_B|)}}}{N}$$

dove la proprietà *P* viene valutata in ogni *k*-esimo punto per i composti *A* e *B*. Tuttavia, essendo la distanza *d* utilizzata normalizzata tra 0 e 1, la funzione proposta non va a zero quando la distanza è massima (*d* = 1), ma converge a exp(-1) = 0.368. Per questo motivo, è corretto utilizzare la funzione seguente:

$$E_{AB} = \frac{\sum_{k=1}^{N}(exp^{-d} - d \cdot exp^{-1})}{N} \qquad d = \frac{|P_A - P_B|}{max(|P_A|,|P_B|)}$$

13.20 - Descrittori di reattività chimica e processo

Questa classe di descrittori viene utilizzata nei problemi riguardanti lo studio di reazioni chimiche. Il loro scopo è quello, ad esempio, di rendere conto della resa di un prodotto selezionato rispetto alle condizioni di processo, al catalizzatore utilizzato, ai sottoprodotti ottenuti, ecc. oppure è quello di spiegare la diversa reattività di una classe di composti rispetto alle loro proprietà strutturali, ai prodotti ottenuti, ecc.
Alcuni di questi descrittori sono elencati qui di seguito.

🗀 **conversione percentuale**

E' la percentuale di reagente scomparso nella reazione.

🗀 **resa percentuale**

E' la percentuale di prodotto ottenuto nella reazione. In presenza di sottoprodotti, possono essere importanti anche le loro percentuali .

🗀 **costante di velocità**

Le costanti di velocità forniscono informazioni riguardo gli aspetti cinetici di un processo chimico. Esse rappresentano una misura della reattività di un composto per una reazione data e sotto condizioni specificate. Particolarmente utilizzate sono le costanti di velocità di reazioni di idrolisi, dissociazione acida o basica, di fotolisi in acqua, di biodegradazione, ecc.

🗀 **specificità**

E' la percentuale del prodotto desiderato rispetto a tutti i prodotti della reazione.

🗀 **selettività**

E' il potere discriminante di un reagente nell'attacco competitivo di due o più substrati o su due o più posizioni nello stesso substrato. La selettività è quantitativamente espressa dal rapporto tra le costanti di velocità delle reazioni in competizione.

🗀 **catalizzatore del processo**

E' la quantità di catalizzatore utilizzato nella reazione oppure il tipo di catalizzatore. In questo secondo caso, la variabile che descrive i diversi tipi di catalizzatori è una variabile categorica.

🗀 **condizioni di processo**

Per condizioni di processo si intendono quelle variabili che descrivono le condizioni in cui avviene il processo o la reazione. Tra questi, i più comuni sono la *temperatura*, la *pressione*, il *pH*, la *quantità di ossigeno* disponibile in una combustione, il *tempo* di semitrasformazione.

13.21 - Descrittori di attività biologica

Come affermò Paracelso, "Tutte le sostanze sono veleni; non esiste alcuna sostanza che non sia un veleno. La giusta dose differenzia un veleno da un rimedio."
Così l'azione farmacologica o tossica di un composto è basata sulla relazione tra la **dose** somministrata ad un soggetto e la risposta, cioè l'effetto del composto sul soggetto. L'interpretazione della relazione dose-risposta è principalmente basata sulle seguenti assunzioni:

a) la risposta è proporzionale alla dose del composto nel sito obiettivo (recettore);
b) la concentrazione nel sito obiettivo è correlata alla dose;
c) la risposta è causalmente correlata al composto.

Più in particolare, il punto c) significa che l'attività biologica è proporzionale all'affinità di legame (*binding affinity*) tra composto e recettore.
I descrittori che rappresentano le dosi sono sempre definiti da alcune condizioni particolari della relazione dose-risposta.

🗀 **affinità di legame (*binding affinity*, K_b)**

E' l'affinità di legame tra il composto (spesso farmaco) e il recettore ed è definita come il $-\Delta G$ (variazione di energia libera) della corrispondente reazione.

🗀 **costante di inibizione di Michaelis (K_i)**

Per reazioni catalizzate da enzimi, la costante di inibizione di Michaelis è la costante della velocità di reazione determinata dalle modalità di interazione tra farmaco e recettore.

🗀 **dose efficace (*Effective Dose*, ED_{50})**

E' la dose minima che causa l'effetto atteso nel 50% dei soggetti.
Il confronto della dose letale o della dose tossica con la dose efficace fornisce il cosiddetto **indice terapeutico**:

$$I.T.=\frac{LD_{50}}{ED_{50}} \qquad \text{o} \qquad I.T.=\frac{TD_{50}}{ED_{50}}$$

dove maggiore è il rapporto più grande è il margine di sicurezza per l'uso del composto.
Analogamente, è possibile definire un indice di maggior sicurezza, chiamato **margine di sicurezza**, definito come

$$M.S.=\frac{LD_{1}}{ED_{99}} \qquad \text{o} \qquad M.S.=\frac{TD_{1}}{ED_{99}}$$

dove vengono confrontate le dosi letali (tossiche) per l'1% dei soggetti contro le dosi terapeutiche per il 99% dei soggetti.

🗀 **concentrazione efficace (*Effective Concentration*, EC_{50})**

E' la concentrazione minima che risulta efficacie sul 50% dei soggetti. Ad esempio, l'effetto narcotico, l'effetto terapeutico, l'effetto inibitore.

🗀 **dose letale (*Letal Dose*, LD_{50})**

E' la dose minima che causa il decesso del 50% dei soggetti.

🗀 **concentrazione letale (*Letal Concentration*, LC_{50})**

E' la concentrazione minima che causa il decesso del 50% dei soggetti.

🗀 **concentrazione di soglia (*threshold concentration*, NOEC)**

La concentrazione al di sotto della quale non viene osservato nessun effetto o risposta, cioè livello di nessun effetto osservato (*No Observed Effect Concentration*).

🗀 **dose di soglia (*threshold dose*, NOEL)**

La dose al di sotto della quale non viene osservato nessun effetto o risposta, cioè livello di nessun effetto osservato (*No Observed Effect Level*). NOEL è una grandezza importante per definire gli effetti all'esposizione.
Ad esempio, la **dose accettabile di assunzione giornaliera** (*Acceptable Daily Intake, ADI*) può essere definita come

$$ADI = \text{NOEL (mg/kg/day)} / 100$$

e un fattore opportuno di sicurezza può essere definito per valori fino a 100.

🗀 **dose tossica (*Toxic Dose*, TD_{50})**

La dose minima che mostra un effetto tossico sul 50% dei soggetti.

13.22 - Descrittori chemo-ambientali

Sono descrittori che si riferiscono alla stabilità e reattività di sostanze chimiche nei diversi comparti ambientali. I parametri più significativi sono riportati qui di seguito.

🗀 **domanda biochimica di ossigeno (*Biochemical Oxygen Demand, BOD*)**

E' la quantità di ossigeno consumata in 5 giorni (BOD_5) o in 20 giorni (BOD_{20}) per l'ossidazione biochimica delle sostanze organiche presenti in un campione.
Questo descrittore indica quindi la quantità di sostanza organica biodegradabile presente. Ad ogni sostanza organica pura può essere associato un valore di BOD, generalmente espresso in mg di ossigeno/g di composto.

🗀 domanda chimica di ossigeno (*Chemical Oxygen Demand, COD*)

E' la quantità di ossigeno richiesta per l'ossidazione chimica di una sostanza organica o inorganica. La misura standard di COD è data dal valore determinato utilizzando come ossidante il bicromato di potassio in presenza di acido solforico concentrato.

🗀 carbonio organico totale (*Total Organic Carbon, TOC*)

E' la quantità di carbonio organico totale contenuto nel campione, generalmente espressa in ppm.

🗀 domanda totale di ossigeno (*Total Oxygen Demand, TOD*)

E' la quantità di carbonio totale contenuto nel campione, corrispondente alla sua completa trasformazione in acqua e biossido di carbonio.

🗀 carbonio organico teorico (*Theoretical Organic Carbon, ThOC*)

E' la quantità teorica di carbonio organico totale presente nella sostanza considerata.

🗀 domanda teorica di ossigeno (*Theoretical Oxygen Demand, ThOD*)

E' la quantità di ossigeno teoricamente necessaria per la completa trasformazione ossidativa della sostanza in esame.

🗀 biodegradabilità (*biodegradability*)

La biodegradazione è uno dei processi ambientali più importanti per la diminuzione delle concentrazioni di composti organici nell'ambiente.
Un indice della biodegradabilità di un composto è, ad esempio, la costante di velocità della reazione di biodegradazione del composto. Il BOD, ad esempio, è ritenuto rappresentativo del processo di biodegradazione.

🗀 **fattore di bioconcentrazione (*bioconcentration factor, BCF*)**

Il fattore di bioconcentrazione indica il grado in cui un composto chimico può accumularsi in organismi acquatici, relativamente alla concentrazione media del composto nell'acqua ed è definito come:

$$BCF_i = \frac{C_i(organismo)}{\overline{C_i}(acqua)}$$

dove le due concentrazioni dell'*i*-esimo composto sono rispettivamente quella nell'organismo e la concentrazione media nell'acqua.
Le unità di misura al numeratore e al denominatore devono essere le stesse, cioè *BCF* è una grandezza adimensionale. Comunemente *BCF* è compreso tra circa 1 e oltre 1 milione.

🗀 **tempo di residenza in atmosfera (*atmospheric residence time*, τ)**

E' il tempo di residenza di un composto chimico in un determinato comparto ambientale (atmosfera in generale, la troposfera, la stratosfera). Esso è ben definito solo in condizioni di stazionarietà, cioè quando la massa totale e la sua distribuzione statistica non variano nel tempo. In questi casi, il tempo di residenza viene definito come:

$$\tau = \frac{Q}{\mathrm{E}} = \frac{Q}{\mathrm{R}}$$

dove *Q* è la massa totale nel comparto ambientale e E o R sono, rispettivamente, la velocità di emissione o la velocità di rimozione totali. Quindi la quantità E è il contributo complessivo di tutte le emissioni (di terre, acque e qualsiasi altro contributo) all'atmosfera; la quantità R è, analogamente, il contributo totale di tutte le modalità di eliminazione del composto dal comparto considerato, compresa la degradazione *in-situ*.
Il tempo di residenza in atmosfera non può essere tuttavia calcolato direttamente da queste definizioni, ma deve essere inferito da modelli semplificati del comparto ambientale studiato.

13.23 - Le proprietà principali

Sono un caso particolare di variabili che derivano dall'analisi delle componenti principali (PCA) come combinazioni lineari delle variabili originali considerate. Esse, se opportunamente interpretate, possono rappresentare delle macroproprietà del sistema studiato, cioè sue proprietà emergenti o globali (v. par. 3-6).
Gli *scores* di una componente principale, ottenuti dall'analisi delle componenti principali, sono i valori della combinazione lineare considerata e rappresentano i valori della macroproprietà per ogni oggetto. Questo tipo di variabile è particolarmente utile quando si voglia condensare in un'unica variabile l'informazione contenuta in più singole variabili tra loro correlate.

🕮 Bibliografia

A.L. HORVATH (1992): *Molecular Design* - Elsevier, Amsterdam.

B. TESTA (ED.) (1984): *Advances in Drug Research* - Academic Press, vol.13 - Orlando, FL.

C. HANSCH E A. LEO (1995): *Exploring QSAR.* - American Chemical Society, Washington, DC.

W.J. LYMAN, W.F. REEHL E D.H. ROSENBLATT (1982): *Handbook of Chemical Properties Estimation Methods.* - American Chemical Society, Washington, DC.

R.TODESCHINI AND P.GRAMATICA (1977): *3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of the WHIM descriptors.* Quant. Struct.-Act. Relat., **16**, 113-119.

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Appendice A

ALGEBRA DELLE MATRICI

Una **matrice** è una tabella bidimensionale di numeri costituita da *n* righe e *p* colonne. Le matrici sono generalmente denotate da lettere maiuscole in grassetto: **A, B, C, X, Y**.
Un **vettore** è una sequenza di numeri organizzati lungo una fila; in algebra, se non diversamente specificato, il vettore è sempre inteso come **vettore colonna**. I vettori sono denotati da lettere minuscole in grassetto: **a, b, c, x, y, z**.
Una matrice è quindi costituita da *n* righe e da *p* colonne. Il generico elemento della matrice viene rappresentato con due indici relativi ad ogni elemento della matrice; questi indici rappresentano il numero della riga e il numero della colonna, rispettivamente;

$$x_{ij} \quad i = 1,2,\ldots,n \quad j = 1,2,\ldots,p$$

$$\mathbf{X} \equiv \begin{vmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{vmatrix} \equiv \begin{vmatrix} \mathbf{x}_1^{\mathbf{T}} \\ \mathbf{x}_2^{\mathbf{T}} \\ \ldots \\ \ldots \\ \mathbf{x}_n^{\mathbf{T}} \end{vmatrix}$$

L'operazione di trasposizione degli elementi di una matrice consiste nello scambio degli indici di riga e colonna ed è indicata col suffisso "T". Quindi, mediante questa operazione, l'*i*-mo vettore colonna $\mathbf{x_i}$ viene scambiato nel corrispondente vettore riga $\mathbf{x}_i^{\mathbf{T}}$. Analogamente, la matrice $\mathbf{X^T}$ corrisponde alla matrice **X** ove per tutti gli elementi sono state scambiate le righe con le colonne e viceversa.

🗀 Spazio vettoriale

Un *insieme di p vettori* $\mathbf{x}_j$ della stessa dimensione *n* è **linearmente indipendente** se l'espressione:

$$\sum_j c_j \mathbf{x}_j = 0$$

vale solo quando tutti i coefficienti c_j sono nulli. Diversamente, i *p* vettori sono **linearmente dipendenti**.
Un insieme di *p* vettori linearmente indipendenti della stessa dimensione *n* viene detto **base** dello spazio vettoriale $\mathbf{V^n}$. Questo spazio include anche il vettore nullo **0**, i cui elementi sono tutti uguali a zero.
Uno **spazio vettoriale** $\mathbf{V^n}$ definito da *p* vettori $\mathbf{x}_j$ della stessa dimensione *n* è l'insieme di tutti i vettori che sono ottenuti da combinazioni lineari dei *p* vettori della base.
Gli *n* valori che definiscono ciascun vettore dello spazio vettoriale sono detti **coordinate del vettore** nella base $\left\{\mathbf{x}_1,\ldots,\mathbf{x}_p\right\}$.

🗀 Operazioni e relazioni vettoriali importanti

Il prodotto tra un vettore colonna **a** di *n* elementi e un vettore riga $\mathbf{b^T}$ di *p* elementi è una matrice di dimensione *n* x *p*. Questo prodotto tra vettori prende il nome di **prodotto vettoriale** o **prodotto esterno** ed è definito come segue:

$$\mathbf{ab^T} = \begin{bmatrix} a_1 b_1 & \ldots & a_1 b_p \\ \ldots & \ldots & \ldots \\ a_n b_1 & \ldots & a_n b_p \end{bmatrix}$$

Il prodotto tra un vettore riga $\mathbf{a^T}$ di *p* elementi e un vettore colonna **b** di *p* elementi è uno scalare. Questo prodotto tra vettori prende il nome di **prodotto scalare** o **prodotto interno** ed è definito come segue:

$$\mathbf{a^T b} = \sum_{j=1}^{p} a_j b_j = \mathbf{b^T a}$$

Dalle operazioni vettoriali ora definite è possibile ricavare alcune grandezze algebriche fondamentali.
Siano **x** e **y** due vettori colonna, ciascuno di *n* elementi e **1** un vettore unitario di *n* elementi.
Analizziamo le seguenti espressioni.

$$x_1^2 + x_2^2 + \ldots + x_n^2 = \sum_i x_i^2 = \mathbf{x}^\mathbf{T}\mathbf{x} = \left\|\mathbf{x}\right\|^2$$

La radice quadrata di questa quantità corrisponde alla distanza euclidea del punto di coordinate **x** dall'origine dello spazio (dal vettore nullo **0**); la quantità $\left\|\mathbf{x}\right\|$ viene detta **norma** o **lunghezza** del vettore **x**.
Analogamente, il quadrato della **distanza tra due punti** definiti dai vettori **x** e **y** è dato dall'espressione:

$$\left(\mathbf{x}-\mathbf{y}\right)^\mathbf{T} \cdot \left(\mathbf{x}-\mathbf{y}\right) = \sum_i \left(x_i - y_i\right)^2 = \left\|\mathbf{x}-\mathbf{y}\right\|^2$$

dove $\left\|\mathbf{x}-\mathbf{y}\right\|$ è la norma del vettore **x - y**.
Viene detta **distanza angolare** o **angolo** tra due vettori **x** e **y**, vista dall'origine dello spazio, la seguente espressione:

$$x_1y_1 + x_2y_2 + \mathsf{K} + x_ny_n = \sum_i x_iy_i = \mathbf{x}^\mathbf{T}\mathbf{y} = \mathbf{y}^\mathbf{T}\mathbf{x} = \left\|\mathbf{x}\right\| \cdot \left\|\mathbf{y}\right\| \cdot \cos\vartheta$$

dove ϑ rappresenta l'angolo tra i due vettori.
Dalla relazione precedente è possibile definire l'**ortogonalità di due vettori**; infatti ne consegue che due vettori **x** e **y** sono tra loro **ortogonali** se il loro prodotto scalare è zero:

$$\mathbf{x}^\mathbf{T}\mathbf{y} = \mathbf{0}$$

Ciò è possibile se $\cos\vartheta = 0$ o, equivalentemente, se $\vartheta = 90°$.
Un vettore è **normale** (o *normalizzato*) se la sua norma è uguale a 1:

$$\left\|\mathbf{x}\right\| = 1 \quad \text{o} \quad \mathbf{x}^\mathbf{T}\mathbf{x} = 1$$

Due vettori normali **x** e **y** che siano anche ortogonali tra loro vengono detti **ortonormali.**

Poichè la somma degli elementi di un vettore può essere scritta come la somma dei prodotti dei suoi elementi, ciascuno moltiplicato per uno,

$$x_1 + x_2 + \mathsf{K} + x_n = \sum_i x_i = \mathbf{1}^{\mathbf{T}}\mathbf{x} = \mathbf{x}^{\mathbf{T}}\mathbf{1}$$

è possibile calcolare il **valore medio di un vettore x** come:

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{\mathbf{1}^{\mathbf{T}}\mathbf{x}}{n}$$

Analogamente, una **combinazione lineare di un vettore x** viene definita come:

$$y_i = c_1 x_{i1} + c_2 x_{i2} + \mathsf{K} + c_p x_{ip} = \mathbf{c}^{\mathbf{T}}\mathbf{x}_{(i)} = \mathbf{x}_i^{\mathbf{T}}\mathbf{c}$$

dove **c** è il vettore dei coefficienti della combinazione lineare.
Il **valore medio di una combinazione lineare** è $\bar{y} = \mathbf{c}^{\mathbf{T}}\bar{\mathbf{x}}$; si può dimostrare inoltre che la **varianza di una combinazione lineare** è:

$$V(\mathbf{y}) = s_y^2 = \mathbf{c}^{\mathbf{T}}\mathbf{S}\mathbf{c} = \sum_j \sum_k c_j c_k s_{jk}$$

dove **S** è la matrice di covarianza e **c** il vettore dei coefficienti della combinazione lineare.
La **covarianza tra due combinazioni lineari** y_1 e y_2, ciascuna definita dai coefficienti **c** e **d**, rispettivamente, è definita da.

$$C(y_1, y_2) = \mathbf{c}^{\mathbf{T}}\mathbf{S}\mathbf{d} = \sum_j \sum_k c_j d_k s_{jk}$$

🗀 Definizioni ed esempi di matrici

Qui di seguito sono riportate le definizioni fondamentali delle matrici, fornite direttamente mediante un esempio o mediante le proprietà che le definiscono.

matrice rettangolare $\Rightarrow \quad n \neq p \quad \Rightarrow \quad \begin{vmatrix} 2 & 5 & -1 \\ 1 & 2 & 4 \end{vmatrix}$

matrice quadrata $\Rightarrow \quad n = p \quad \Rightarrow \quad \begin{vmatrix} 2 & -6 & 18 \\ 5 & 11 & 7 \\ 3 & 9 & -3 \end{vmatrix}$

matrice simmetrica $\Rightarrow \quad x_{ij} = x_{ji}$ ovvero $\mathbf{X} = \mathbf{X}^{\mathbf{T}} \quad \Rightarrow \quad \begin{vmatrix} 2 & 5 & -1 \\ 5 & 15 & 24 \\ -1 & 24 & 7 \end{vmatrix}$

matrice diagonale $\Rightarrow$ diag(**A**) $\quad \Rightarrow \quad \begin{vmatrix} 32 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{vmatrix}$

matrice unità J $\quad \Rightarrow \quad \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix}$

matrice identità I $\Rightarrow \quad$ **I** = diag(**J**) $\quad \Rightarrow \quad \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}$

matrice ortogonale A $\Rightarrow \mathbf{A} \cdot \mathbf{A}^{\mathbf{T}} = \mathbf{D_n}$ **,**

dove $\mathbf{D_n}$ è una matrice diagonale di dimensione *n*.

matrice ortogonale A $\Rightarrow \mathbf{A}^{\mathbf{T}} \cdot \mathbf{A} = \mathbf{D_p}$

dove $\mathbf{D_p}$ è una matrice diagonale di dimensione *p*.

matrice ortonormale A $\Rightarrow \mathbf{A} \cdot \mathbf{A}^{\mathbf{T}} = \mathbf{I_n}$

dove $\mathbf{I_n}$ è una matrice identità di dimensione *n*.

matrice ortonormale A $\Rightarrow \mathbf{A}^{\mathbf{T}} \cdot \mathbf{A} = \mathbf{I_p}$

dove $\mathbf{I_p}$ è una matrice identità di dimensione *p*.

🗀 Operazioni matriciali fondamentali

Qui di seguito sono riportate le operazioni fondamentali per le matrici, fornite direttamente mediante un esempio o mediante le proprietà che definiscono il risultato.

Somma di matrici $\Rightarrow \quad \mathbf{A} + \mathbf{B} = \mathbf{C}$

dimensioni : (2, 3) (2, 3) = (2, 3)

Ogni singolo elemento c_{ij} della matrice somma viene calcolato come $a_{ij} + b_{ij}$.

$$\begin{vmatrix} 2 & 5 & -1 \\ 1 & 2 & 4 \end{vmatrix} + \begin{vmatrix} 0 & 1 & 3 \\ 4 & -3 & 1 \end{vmatrix} = \begin{vmatrix} 2 & 6 & 2 \\ 5 & -1 & 5 \end{vmatrix}$$

Prodotto di matrici $\Rightarrow$ **A B = C**

dimensioni : (2, 3) (3, 3) = (2, 3)

Ogni singolo elemento della matrice prodotto **C** è uno scalare che viene calcolato come:

$$\mathbf{a^T b} = \sum_k a_{ik} b_{kj} = c_{ij}$$

dove l'indice *k* è detto *indice interno* e corrisponde alle colonne del vettore **a** e alle righe del vettore **b**.
Ad esempio,

$$\mathbf{B} = \begin{vmatrix} 3 & 0 & 3 \\ 7 & 3 & -1 \\ 1 & 2 & 5 \end{vmatrix}$$

$$\mathbf{A} = \begin{vmatrix} 2 & 5 & -1 \\ 1 & 2 & 4 \end{vmatrix} \qquad \mathbf{C} = \begin{vmatrix} 40 & 13 & -4 \\ 21 & 14 & 21 \end{vmatrix}$$

Ad esempio, l'elemento c_{11} viene calcolato nel seguente modo:

$$a_{11} \cdot b_{11} + a_{12} \cdot b_{21} + a_{13} \cdot b_{31} = 2 \cdot 3 + 5 \cdot 7 + (-1) \cdot 1 = 40$$

Traccia di una matrice

La traccia di una matrice quadrata di dimensione *n* è la somma dei suoi elementi diagonali:

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Determinante di una matrice

Il determinante di una matrice quadrata di dimensione *n* è la somma dei prodotti di tutte le permutazioni dei suoi elementi, secondo la formula:

$$det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+1} a_{1i} det(\mathbf{A}_{1i})$$

dove $\mathbf{A}_{1i}$ è la matrice di dimensione (*n*-1) x (*n*-1) ottenuta eliminando la prima riga e la *i*-ma colonna di **A**.

Inversione di una matrice

La **matrice inversa di una matrice diagonale** si calcola semplicemente dall'inverso dei suoi elementi diagonali:

$$\text{matrice diagonale } \mathbf{A} \Rightarrow \begin{vmatrix} 32 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{vmatrix} \qquad \mathbf{A^{-1}} = \begin{vmatrix} 1/32 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/9 \end{vmatrix}$$

Negli altri casi è necessario ricorrere ad opportuni algoritmi di inversione (qui non trattati).

🗀 Proprietà fondamentali delle matrici

Una volta definite le operazioni fondamentali per vettori e matrici, è possibile dimostrare che valgono le seguenti proprietà (α e β sono coefficienti).

$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

$\mathbf{AB} \neq \mathbf{BA}$

$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

$1 \cdot \mathbf{A} = \mathbf{A}$

$0 \cdot \mathbf{A} = \mathbf{0}$

$(\alpha + \beta) \cdot \mathbf{A} = \alpha \cdot \mathbf{A} + \beta \cdot \mathbf{A}$

$\alpha \cdot (\mathbf{A} + \mathbf{B}) = \alpha \cdot \mathbf{A} + \alpha \cdot \mathbf{B}$

$(\mathbf{A} \cdot \mathbf{B})^{\mathbf{T}} = \mathbf{B}^{\mathbf{T}} \cdot \mathbf{A}^{\mathbf{T}}$

$det(\mathbf{AB}) = det(\mathbf{A})\, det(\mathbf{B})$

$det(\mathbf{A}) = det(\mathbf{A}^{\mathbf{T}})$

$det(\alpha\mathbf{A}) = \alpha^n\, det(\mathbf{A}) \qquad con\ \mathbf{A}\ (n\,x\,n)$

🗀 Quantità algebriche importanti

Si definisce **centroide delle colonne di una matrice** il vettore costituito dalle medie di ciascuna colonna:

$$\bar{\mathbf{x}}^{\mathbf{T}} = \frac{\mathbf{1}^{\mathbf{T}}\mathbf{X}}{n}$$

In temini matriciali, la **matrice di covarianza** di una matrice **X** (*n*, *p*) è definita come:

$$\mathbf{S} = \frac{\mathbf{X}^{\mathbf{T}}\mathbf{X}}{n-1} - \bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathbf{T}} = \frac{\mathbf{X}_C^{\mathbf{T}}\mathbf{X}_C}{n-1}$$

dove **S** è una matrice quadrata simmetrica di dimensione (*p*, *p*).
Anche i diversi tipi di scalatura dei dati possono essere rappresentati in forma matriciale. Ad esempio, la **centratura delle colonne di una matrice** viene effettuata con l'espressione:

$$\mathbf{X}_C = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\mathbf{T}}$$

La **autoscalatura delle colonne di una matrice** viene effettuata secondo l'espressione:

$$\mathbf{X}_A = \frac{\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\mathbf{T}}}{\mathbf{S}^{\mathbf{1/2}}} = \mathbf{X}_C\mathbf{S}^{\mathbf{-1/2}}$$

dove **S** è la matrice di covarianza di **X**.

🗀 Interpretazione geometrica dei prodotti matriciali

Il prodotto tra una matrice **X** (*n*,*p*) ed un vettore **v** (*p*,1) può essere interpretato geometricamente come la **proiezione** di un insieme di punti (le righe di **X**) su un asse definito dal vettore **v**:

$$\mathbf{s} = \mathbf{X} \cdot \mathbf{v}$$

ottenendo un vettore *n*-dimensionale **s**.
Ogni *i*-esimo elemento del vettore **s** è calcolato dall'espressione:

$$s_i = \mathbf{x}_i^{\mathrm{T}} \cdot \mathbf{v}$$

Nel caso più generale, è possibile proiettare i punti definiti in uno spazio *p*-dimensionale rappresentati dalla matrice **X** (*n*,*p*) in uno spazio **S** (*n*,*m*) definito da *m* nuovi assi rappresentati dalle colonne della matrice **V**, secondo la relazione:

$$\mathbf{S} = \mathbf{X} \cdot \mathbf{V}$$

I punti rappresentati dalla matrice **S** nel nuovo spazio sono detti *scores* e la matrice **S** viene detta *matrice degli scores*.
Nel caso in cui le dimensioni del nuovo spazio (*M*) coincidano con le dimensioni dello spazio originale (*p*), la proiezione viene chiamata **rotazione** e **V** viene chiamata *matrice di rotazione*.

🗀 Autovalori e autovettori di una matrice

Sia **A** una matrice simmetrica non-singolare (*p* x *p*) e sia **v** un vettore non nullo.

Se esiste un numero λ tale che valga la seguente relazione:

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

v viene detto **autovettore** di **A** e λ il corrispondente **autovalore** di **A**.
Gli autovalori λ di una matrice **A** simmetrica non-singolare (p x p) sono le p radici del suo **polinomio caratteristico**:

$$\mathrm{p}(\lambda) = det(\mathbf{A} - \lambda\mathbf{I})$$

La matrice **A** possiede quindi p autovalori - le soluzioni dell'equazione caratteristica - e il loro insieme è chiamato **spettro** di **A** (λ(**A**)) .
Valgono quindi le due relazioni importanti:

$$det(\mathbf{A}) = \lambda_1\lambda_2 \mathrm{K}\ \lambda_p \qquad tr(\mathbf{A}) = \lambda_1 + \lambda_2 + \mathrm{K}\ + \lambda_p$$

A ciascuno degli autovalori λ possiamo associare un autovettore **v**.
I p autovettori sono ortonormali. Questo sistema ha una soluzione non banale $\mathbf{v} \neq \mathbf{0}$ se e solo se, per ogni j = 1,.., p , $det(\mathbf{A} - \lambda_j \mathbf{I}) = 0$

Nota. Gli autovalori di una matrice reale simmetrica sono numeri reali ≥ 0.

Dalla prima relazione, segue che:

$$\mathbf{v}^{\mathbf{T}} \cdot \mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}^{\mathbf{T}} \cdot \mathbf{v} = \lambda$$

dalla condizione di ortonormalità.
La **decomposizione in autovettori** (*EVD, eigenvector decomposition*) di una matrice simmetrica non-singolare **A** può quindi essere rappresentata dalla seguente relazione:

$$\mathbf{V}^{\mathbf{T}} \cdot \mathbf{A} \cdot \mathbf{V} = \Lambda$$

con $\mathbf{V}^{\mathbf{T}} \cdot \mathbf{V} = \mathbf{V} \cdot \mathbf{V}^{\mathbf{T}} = \mathbf{I_p}$ e $\mathbf{\Lambda}$ una matrice diagonale (p,p).

Poichè $\mathbf{\Lambda}$ è una matrice diagonale, questa decomposizione viene anche chiamata **diagonalizzazione** di **A**. Sfruttando le condizioni di ortonormalità possiamo riscrivere l'espressione precedente come:

$$\mathbf{A} = \mathbf{V} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}}$$

Questa relazione viene chiamata **decomposizione spettrale** di **A** e può essere scritta come:

$$\mathbf{A} = \lambda_1 \mathbf{v}_1 \cdot \mathbf{v}_1^{\mathbf{T}} + \mathrm{K} + \lambda_p \mathbf{v}_p \cdot \mathbf{v}_p^{\mathbf{T}}$$

La stessa matrice **A** può anche essere decomposta in altro modo, secondo la **decomposizione di Young-Householder**:

$$\mathbf{A} = \mathbf{V} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}} = \mathbf{V} \cdot \Lambda^{1/2} \cdot \Lambda^{1/2} \cdot \mathbf{V}^{\mathbf{T}} = \mathbf{S} \cdot \mathbf{S}^{\mathbf{T}}$$

dove $\mathbf{S} = \mathbf{V} \cdot \mathbf{\Lambda}^{1/2}$ e $\mathbf{\Lambda}^{1/2}$ è la matrice diagonale i cui valori sono le radici quadrate degli autovalori.
La decomposizione spettrale viene utilizzata come procedura per calcolare le *potenze di una matrice simmetrica*.
Ad esempio, la matrice quadrata della matrice simmetrica **A** può essere calcolata dalla relazione:

$$\mathbf{A}^2 = \mathbf{A} \cdot \mathbf{A}^{\mathbf{T}} = \left(\mathbf{V} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}}\right) \cdot \left(\mathbf{V} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}}\right) = \mathbf{V} \cdot \Lambda^2 \cdot \mathbf{V}^{\mathbf{T}}$$

Vale quindi la relazione generale:

$$\mathbf{A}^k = \mathbf{V} \cdot \Lambda^k \cdot \mathbf{V}^{\mathbf{T}}$$

per ogni esponente reale *k*. In particolare, questa relazione può essere utilizzata per calcolare la matrice inversa di **A**:

$$\mathbf{A}^{-1} = \mathbf{V} \cdot \Lambda^{-1} \cdot \mathbf{V}^{\mathbf{T}}$$

Il calcolo diretto degli autovalori ed autovettori di una matrice rettangolare **X** (*n*,*p*) viene effettuato mediante una procedura nota col nome di **decomposizione a valore singolo.** Infatti, un teorema dell'algebra, noto col nome di *singular value decomposition* (*SVD*), stabilisce che una qualsiasi matrice **X** (*n* x *p*) può essere definita come prodotto di tre termini **U**, **Λ** e **V**:

$$\mathbf{X} = \mathbf{U} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}}$$

dove **U** è una *matrice ortonormale n* x *q*, **V** è una *matrice ortonormale p* x *q* e $\mathbf{\Lambda}$ è una *matrice diagonale r* x *r*.
La dimensione *q* può essere al più uguale alla dimensione più piccola tra *n* e *p*. Gli elementi della diagonale di $\mathbf{\Lambda}$ possono assumere solo valori positivi, mentre gli elementi fuori dalla diagonale principale sono nulli.
La condizione di ortonormalità implica:

$$\mathbf{U}^{\mathbf{T}} \cdot \mathbf{U} = \mathbf{V}^{\mathbf{T}} \cdot \mathbf{V} = \mathbf{I}_r$$

dove $\mathbf{I}_r$ è la *matrice identità r* x *r*.

In generale, $r \leq p < n$, e *r* viene detto *rango* di **X**. Esso rappresenta il numero di misure indipendenti in **X**. Questo significa che le colonne di **U** e **V** costituiscono dei vettori normalizzati, cioè la cui somma dei quadrati è uguale ad uno e i loro prodotti incrociati sono nulli. Si può anche dimostrare che questa decomposizione è sempre possibile e che la soluzione è unica.

Nota. Gli autovettori **U** e **V** di **X** sono identici agli autovettori di $\mathbf{X}^{\mathbf{T}} \cdot \mathbf{X}$ e $\mathbf{X} \cdot \mathbf{X}^{\mathbf{T}}$, rispettivamente. La matrice **U** è la matrice degli **autovettori di riga**, la matrice **V** è la matrice degli **autovettori di colonna** e Λ^2 è la matrice diagonale degli **autovalori** associati.

Dalle espressioni precedenti, si ricavano immediatamente le seguenti espressioni:

$$\mathbf{C}_n = \mathbf{X} \cdot \mathbf{X}^{\mathbf{T}} = \mathbf{U} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}} \cdot \mathbf{V} \cdot \Lambda \cdot \mathbf{U}^{\mathbf{T}} = \mathbf{U} \cdot \Lambda^2 \cdot \mathbf{U}^{\mathbf{T}}$$

$$\mathbf{C}_p = \mathbf{X}^{\mathbf{T}} \cdot \mathbf{X} = \mathbf{V} \cdot \Lambda \cdot \mathbf{U}^{\mathbf{T}} \cdot \mathbf{U} \cdot \Lambda \cdot \mathbf{V}^{\mathbf{T}} = \mathbf{V} \cdot \Lambda^2 \cdot \mathbf{V}^{\mathbf{T}}$$

La matrice $\mathbf{C}_p$ è direttamente proporzionale alla *matrice di covarianza* di **X**.

Queste equazioni definiscono la cosiddetta *decomposizione ad autovettori* (*Eigenvector Decomposition*). La matrice dei *loadings* **L** è definita come $\mathbf{L} = \mathbf{V}\Lambda$; possiamo quindi definire la seguente relazione:

$$\mathbf{C}_p = \mathbf{X}^{\mathrm{T}} \cdot \mathbf{X} = \mathbf{V} \cdot \Lambda^2 \cdot \mathbf{V}^{\mathrm{T}} = \mathbf{L} \cdot \Lambda \cdot \mathbf{L}^{\mathbf{T}}$$

Nota. La *diagonalizzazione* è l'operazione mediante la quale una matrice simmetrica viene trasformata in una matrice diagonale. In particolare, la matrice quadrata simmetrica $\mathbf{C}_p$ viene diagonalizzata dalla matrice ortonormale **V** nella matrice diagonale Λ^2.

🗀 Tabelle riassuntive

Nelle tre tabelle seguenti vengono riassunte sinteticamente le più importanti definizioni e operazioni utilizzate per vettori e matrici. In particolare, nella Tab.A-1 sono riportate le definizioni, nella Tab.A-2 le operazioni fondamentali, nella Tab.A-3 le definizioni di alcune matrici particolari.

Nome	*Condizioni / Proprietà*	*Notazione*
scalare	$n = p = 1$	*a*, *b*, *x*, *y*, *k*, *m*
vettore colonna	$p = 1$; n = dimensione del vettore	**x**, **y**, **t**, **b**
vettore riga	$n = 1$; p = dimensione della riga	$\mathbf{x}^{\mathbf{T}}$, $\mathbf{y}^{\mathbf{T}}$
M. rettangolare	rig (M) ¹ col (M)	**X**(n,p)
M. quadrata	rig (M) = col (M)	**X**(n,n) , **X**(p,p)
M. diagonale	$m_{ii} \neq 0$; $m_{ij} = 0$ se $i \neq j$	diag(**X**)
M. identità	$m_{ij} = 1$ *se* $i = j$; $m_{ij} = 0$ *se* $i \neq j$	**I**
M. unità	$m_{ij} = 1$ per ogni i e j	**J**
M. nulla	$m_{ij} = 0$ per ogni i e j	**0**
M. triangolare sup.	$m_{ij} = 0$ *se* $i > j$	Δ' o **U**
M. triangolare inf.	$m_{ij} = 0$ *se* $j > i$	Δ o **L**
M. simmetrica	M. quadrata e $m_{ij} = m_{ji}$	

TAB. A-1

Operazione	*Simbolo*	*Definizione*	*Condizioni*	*Dim.*
trasposta	$\mathbf{A}^{\mathbf{T}}$	$c_{ij} = a_{ji}$		c_A, r_A
somma	A + B	$c_{ij} = a_{ij} + b_{ij}$	dim (A) = dim (B)	r_A, c_A
sottrazione	A - B	$c_{ij} = a_{ij} - b_{ij}$	dim (A) = dim (B)	r_A, c_A
prodotto matriciale	AB	$c_{ij} = \sum_k a_{ik} b_{kj} = \mathbf{a}_i^{\mathbf{T}} \mathbf{b}_j$	col (A) = rig (B)	r_A, c_B
prodotto scalare	k A	$c_{ij} = k \cdot a_{ij}$		r_A, c_A
traccia	tr(A)	$tr(\mathbf{A}) = \sum_i a_{ii}$	A quadrata	scalare
determinante	$\lvert\mathbf{A}\rvert$		A quadrata	scalare
inversa	$\mathbf{A}^{-1}$	$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$	A quadrata; $\lvert\mathbf{A}\rvert \neq 0$	r_A, c_A
inversa generalizzata	$\mathbf{A}^{-}$	$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$		r_A, c_A
diagonalizzazione	diag(A)	$\mathbf{V}^{\mathbf{T}}\mathbf{A}\mathbf{V} = \Lambda$	A quadrata simm.	r_A, c_A

TAB. A-2

Nome	*Definizione*
M. non-singolare	M. quadrata; $\left\|\mathbf{A}\right\| \neq 0$
M. singolare	M. quadrata; $\left\|\mathbf{A}\right\| = 0$
M. ortogonale	M. quadrata; $\mathbf{A}\mathbf{A}^{\mathbf{T}} = \mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{I}$
M. ortonormale	M. ortogonale
M. idempotente	$\mathbf{A}^{\mathbf{2}} = \mathbf{A}$
M. definita positiva	$\mathbf{a}^{\mathbf{T}}\mathbf{A}\mathbf{a} > \mathbf{0} \quad \forall \mathbf{a} \neq \mathbf{0}$
M. definita semi-positiva	$\mathbf{a}^{\mathbf{T}}\mathbf{A}\mathbf{a} \geq \mathbf{0} \quad \forall \mathbf{a} \neq \mathbf{0}$

TAB. A-3

🕮 Bibliografia

K.V. MARDIA, J.T. KENT E J.M. BIBBY (1988): *Multivariate Analysis*. Academic Press, London, U.K.

G.H. GOLUB E C.F. VAN LOAN (1983): *Matrix Computations*. John Hopkins Univ. Press, Baltimore, MD, USA.

Appendice B

ELEMENTI DI STATISTICA

In questa appendice sono riportate le definizioni matematiche dei più comuni indici statistici, suddivisi in *momenti, indici di tendenza centrale, indici di dispersione*, *indici di simmetria, altri indici, indici bivariati*. Inoltre sono riportate le espressioni della diverse matrici di covarianza, alcuni cenni sulle distribuzioni statistiche e sui test statistici delle ipotesi.

Nota. Le formule utilizzate si riferiscono sempre ad un **insieme campionario di dati** e non all'intera popolazione.

Momenti

I momenti sono in statistica quantità fondamentali per il calcolo di numerosi altri indici statistici.

🗀 Momenti semplici

I momenti semplici di ordine *k* sono definiti dalla seguente espressione:

$$m_k = \sum_i w_i x_i^k$$

dove w_i sono i pesi (le probabilità) associati a ciascun oggetto ($\sum_i w_i = 1$).
Per definizione: $m_0 = 1$, sempre, e $m_1 = \overline{x}$, la media aritmetica pesata.

Momenti centrali

I momenti centrali di ordine *k* sono definiti dalla seguente espressione:

$$m'_k = \sum_i w_i \left(x_i - \bar{x}\right)^k$$

dove w_i sono i pesi (le probabilità) associati a ciascun oggetto ($\sum_i w_i = 1$) e $\bar{x}$ è la media aritmetica pesata.
Per definizione: $m'_0 = 1$, sempre; $m'_1 = 0$, sempre; $m'_2 = s^2$, la varianza campionaria.

Indici di tendenza centrale

Gli indici di tendenza centrale stimano il valore centrale di un insieme di dati.

media aritmetica

$$\bar{x}_j = \frac{\sum_i x_{ij}}{n}$$

media pesata

$$\bar{x}_j = \frac{\sum_i w_i x_{ij}}{\sum_i w_i}$$

dove w_i è il peso dell'*i*-mo dato.

mediana

E' il valore centrale di una distribuzione.

$$med_j = x_{(n+1)/2} \qquad \text{se } n \text{ è dispari}$$

$$med_j = \frac{x_{n/2} + x_{(n/2)+1}}{2} \qquad \text{se } n \text{ è pari}$$

🗀 **moda**

La moda è il valore corrispondente alla massima frequenza nei dati. Una distribuzione uniforme non ha quindi moda; distribuzioni ove la frequenza massima viene conseguita per due valori diversi della variabile si chiamano bimodali. Nel caso una distribuzione normale, la moda coincide con la media (e con la mediana).

🗀 **media tagliata (*trimmed mean*)**

Questo tipo di media viene calcolata dalla media aritmetica ottenuta eliminando una data percentuale di punti dall'inizio e dalla fine dei dati, una volta che questi sono stati ordinati. Le due medie tagliate più comuni sono al 5% e al 10%. Ciò significa, ad esempio, che nel secondo caso vengono eliminati il 10% dei punti inferiori ed il 10% dei punti superiori e la media aritmetica viene calcolata sull'80% dei punti. In questo modo, la media aritmetica non viene influenzata dai valori estremi.

🗀 **media geometrica**

E' la versione moltiplicativa della media aritmetica. E' definita come

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

ed applicabile solo se per tutti i valori vale la condizione $x_i \geq 0$.
Caratteristica principale della media geometrica è che è sufficiente che un solo valore x_i sia nullo per rendere nulla la media geometrica.

Indici di dispersione

Gli indici di dispersione misurano la dispersione di una distribuzione di dati intorno ad un valore centrale.

varianza

$$s_j^2 = \frac{\sum_i \left(x_{ij} - \bar{x}_j \right)^2}{n-1}$$

varianza pesata

$$s_j^2 = \frac{\sum_i \ w_i \left(x_{ij} - \bar{x}_j \right)^2}{n-1}$$

dove la media è la media pesata dei dati e w_i è il peso dell'*i*-mo dato.

deviazione standard

$$s_j = \sqrt{s_j^2}$$

scarto medio

$$\Delta x = \frac{\sum_i \left| x_i - \bar{x} \right|}{n}$$

errore standard della stima

$$\bar{s}_j = \sqrt{\frac{s_j^2}{n}}$$

🗀 **coefficiente di variazione**

Il coefficiente di variazione è un indice di dispersione standardizzato con il valor medio, in modo da rendere confrontabili deviazioni standard di dati con medie differenti.

$$CV\% = \frac{s}{\bar{x}} \times 100$$

Ovviamente questo indice è calcolabile solo per medie diverse da zero.

Indici di simmetria

Gli indici di simmetria misurano la simmetria di una distribuzione di dati rispetto ad un centro (ad esempio, la media $\bar{x}$).

🗀 **primo indice di Pearson**

$$\gamma_1 = \frac{\sum_i \left(x_i - \bar{x}\right)^3}{n \cdot s^3}$$

🗀 **skewness**

$$Sk = \frac{\bar{x} - \text{moda}}{s}$$

Nel caso di distribuzioni simmetriche (ad esempio, le distribuzioni normale ed uniforme), entrambi gli indici di simmetria assumono valore zero. Nel caso in cui la distribuzione dei dati abbia una coda verso destra (valori superiori al valore centrale), gli indici di simmetria assumono valori positivi, mentre, nel caso contrario, assumono valori negativi.

Altri indici

Vengono qui riportati alcuni indici di diversa natura, quali l'indice di curtosi e gli indici di entropia.

🗀 **indice di curvatura: curtosi**

$$\gamma_2 = \frac{\sum_i \left(x_i - \bar{x}\right)^4}{n \cdot s^4}$$

Valori caratteristici dell'indice di curtosi sono: $\gamma_2 = 3$ per distribuzioni normali; $\gamma_2 = 1.8$ per distribuzioni uniformi. Il valore minimo è uguale a 1 (con due soli oggetti); il valore massimo è infinito, quando tutti i campioni hanno lo stesso valore, cioè formano un picco. La curtosi è un indice del *carattere di bimodalità* della distribuzione dei dati, cioè se tutti i dati sono equamente distribuiti ai due estremi (il minimo ed il massimo), la distribuzione è strettamente bimodale (nessun dato cade tra i due estremi) e la curtosi assume valore minimo (uno). Se tutti i dati sono concentrati nel valore centrale (la distribuzione è un picco), la distribuzione è strettamente unimodale e la curtosi vale infinito.

🗀 **indice di diversità di Shannon**

$$H = -\sum_i p_i \log_2 p_i$$

dove p_i è la probabilità dell'*i*-mo evento, calcolata di norma come il rapporto tra il numero di eventi dell'*i*-ma classe e il numero di eventi totali. Questo indice, detto anche *entropia* o *indice di variabilità*, può essere normalizzato con la seguente espressione:

$$H_N = \frac{-\sum_i p_i \log_2 p_i}{\log_2 n}$$

Questo indice vale 0 in caso di variabilità nulla e 1 nel caso di variabilità massima, cioè quando tutte le probabilità sono date da $p_i = 1/n$.

🗀 **indice di diversità di Gini**

$$G.I. = \sum_{k,k'} p_k p_{k'} \quad k \neq k'$$

dove p_k è la probabilità del *k*-mo evento, calcolata di norma come il rapporto tra il numero di eventi della *k*-ma classe e il numero di eventi totali.

Matrici di dispersione, di covarianza e di correlazione

Quando vengono prese in considerazione molte variabili contemporaneamente, la matrice di varianza e covarianza viene utilizzata per raccogliere in modo organizzato tutte le informazioni riguardanti la dispersione delle variabili. Data l'importanza che questo tipo di informazione ha nella maggior parte delle tecniche multivariate, riportiamo qui di seguito le diverse tipologie di matrici di dispersione.

🗀 **matrice di dispersione (*scatter matrix*)**

E' una matrice quadrata simmetrica che descrive la dispersione di dati mutivariati intorno alla media:

$$t_{jk} = \sum_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$$

Nel caso di variabili centrate, la matrice di dispersione si chiama **matrice di informazione**. La matrice di dispersione **T** è legata alla matrice di covarianza dalla relazione:

$$\mathbf{S} = \frac{\mathbf{T}}{n-1}$$

Nel caso in cui i dati siano suddivisi in gruppi (classi), la matrice di dispersione può essere decomposta in una matrice di dispersione all'interno dei gruppi (**W**) e in una matrice di dispersione tra i gruppi (**B**):

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

🗀 matrice di covarianza (*covariance matrix*)

La matrice di covarianza, di dimensione p x p, è l'estensione naturale del concetto di varianza. I valori della varianza di ciascuna variabile compaiono come elementi della diagonale principale, mentre fuori dalla diagonale compaiono i valori delle covarianza, cioè di come "co-variano" o variano congiuntamente coppie di variabili.

$$s_{jk}^2 = \frac{\sum_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{n-1}$$

$$\mathbf{S} = \begin{vmatrix} s_{11}^2 & s_{12}^2 & \cdots & \cdots & s_{1p}^2 \\ s_{21}^2 & s_{22}^2 & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{p1}^2 & & & & s_{pp}^2 \end{vmatrix}$$

e, in termini matriciali:

$$\mathbf{S} = \frac{\mathbf{X}_C^{\mathrm{T}}\mathbf{X}_C}{n-1}$$

Gli elementi della **matrice di covarianza pesata** sono definiti come:

$$s_{jk}^2 = \frac{\sum_i w_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{\sum w_i}$$

dove w_i rappresentano i pesi assegnati a ciascun *i*-esimo oggetto.

Nota. La covarianza *C* tra due vettori **x** e **y** ha le seguenti proprietà:

1. $C\left[(\mathbf{x}+\alpha),(\mathbf{y}+\beta)\right] = C(\mathbf{x},\mathbf{y})$

2. $C(\alpha\mathbf{x},\beta\mathbf{y}) = \alpha\beta\ C(\mathbf{x},\mathbf{y})$

dove α e β sono due coeffcienti numerici.

🗀 matrice di correlazione (*correlation matrix*)

La matrice di correlazione è la matrice di covarianza standardizzata a varianza unitaria.

$$c_{jk}^2 = \frac{s_{jk}^2}{s_j \cdot s_k}$$

Dalla definizione ne consegue che $c_{jj}^2 \equiv c_j^2 = 1$.
Si noti che la matrice di correlazione coincide con la matrice di covarianza calcolata su dati autoscalati.

🗀 matrice di covarianza intorno all'origine

$$s_{jk}^2 = \frac{\sum_i x_{ij} x_{ik}}{n}$$

e, in termini matriciali,

$$\mathbf{S} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}$$

🗀 **matrice di covarianza di gruppo (*within-group covariance matrix*)**

$$s_{jkg}^2 = \frac{\sum_i \left(x_{ijg} - c_{jg}\right)\left(x_{ikg} - c_{kg}\right)}{n_g - 1}$$

e, in termini matriciali,

$$\mathbf{S}_w = \frac{\mathbf{W}}{n_g - 1}$$

🗀 **matrice di covarianza tra i gruppi (*between-group covariance matrix*)**

E' una misura della dispersione dei gruppi intorno al baricentro complessivo.

$$s_{jkg}^2 = \frac{G}{n} \frac{\sum_i n_g \left(c_{jg} - c_j\right)\left(c_{kg} - c_k\right)}{G - 1}$$

dove *G* è il numero delle classi.

🗀 **matrice di covarianza media (*mean-group covariance matrix*)**

E' la media delle matrici di covarianza di gruppo, definita come:

$$\mathbf{S}' = \frac{\sum_g \mathbf{S}_g}{G}$$

dove *G* è il numero delle classi.

🗀 **matrice di covarianza comune (*pooled covariance matrix*)**

E' la dispersione media pesata dei gruppi, definita come:

$$\mathbf{S}_p = \frac{\sum_g \left(n_g - 1\right)\mathbf{S}_g}{n - G}$$

dove *G* è il numero delle classi.

Distribuzioni di probabilità

Il comportamento di ogni variabile casuale (*aleatoria*) X può essere descritto per mezzo di una funzione *f(x)* chiamata **densità di probabilità (f.d.p)**. Nota la distribuzione statistica, la funzione *f*(*x*) è il valore della probabilità nel punto *x* della distribuzione, cioè *P*(X = *x*).
L'integrale di questa funzione

$$F(x) = \int_{-\infty}^{x} f(u)\mathbf{du}$$

corrisponde alla **funzione di distribuzione cumulativa (c.d.f)**, cioè *P*(X ≤ *x*).
Il valore di *F*(*x*) è il valore della probabilità che un valore a caso di *x* sia compreso tra -∞ e *x* stesso. Ovviamente, dato *x*, la probabilità di ottenere un valore maggiore di *x*, cioè tra *x* e +∞, è data dall'espressione:

$$S(x) = 1 - F(x)$$

dove *S*(*x*) è detta **funzione di sopravvivenza** (***survival function*** o ***reliability function*****)**, cioè *P*(X > *x*)**.**
Per le proprietà della probabilità, qualsiasi sia la funzione di distribuzione, deve essere:

$$F(x) = \int_{-\infty}^{+\infty} f(u)\mathbf{du} = 1$$

cioè la probabilità che venga estratto un qualsiasi valore di X tra -∞ e +∞ deve essere uguale a uno.
In statistica sono note molte decine di diverse distribuzioni statistiche. Le più note sono le distribuzioni normale (o distribuzione gaussiana), *t* di Student, *F* di Fisher e χ^2 (chi-quadrato).

Test delle ipotesi

I test delle ipotesi sono una parte importante della statistica e sono effettuati per verificare se un'ipotesi statistica deve essere accettata o rifiutata ad un dato livello di probabilità.
Di norma le ipotesi statistiche sono formulate come segue:

H_0: ipotesi nulla o di indifferenza statistica (variazioni esclusivamente casuali)

H_1 : ipotesi di lavoro o alternativa (variazioni significative, non dovute al caso)

Perciò, secondo questo schema, ad ogni ipotesi H_1 è associata la corrispondente ipotesi nulla o di indifferenza statistica.

	D E C I S I O N E	
	H_0 vera	H_0 falsa
H_0 **è** vera	livello di fiducia $1 - \alpha$	errore del 1° tipo α
H_0 **è** falsa	errore del 2° tipo β	livello di significatività $1- \beta$

Se il valore ottenuto da un test statistico ha probabilità uguale o minore di ♋, respingiamo H_0 e accettiamo H_1. Questa decisione viene presa in base al seguente ragionamento: se la probabilità, sotto l'ipotesi H_0, di un valore particolare della variabile casuale è molto piccola, possiamo spiegare in due modi l'evento con questa probabilità o con una probabilità anche più piccola di ♋:

a) l'ipotesi nulla H_0 è falsa (cioè, è vera H_1)

b) si è verificato un evento estremamente raro (restando vera H_0).

Generalmente la decisione tende verso la prima delle due ipotesi, anche se occasionalmente può essere corretta la seconda spiegazione.

Nota. Le **tabelle statistiche** raccolgono i valori di una variabile casuale in corrispondenza dei quali la sua funzione di distribuzione $F(x)$ assume valori prefissati di probabilità (ad esempio, 1%, 5%, 95%).
I valori assunti dalla variabile casuale sono detti **valori critici**.
Le tabelle più utilizzate sono quelle corrispondenti alla distribuzione normale, alla distribuzione t di Student, alla distribuzione F di Fisher e alla distribuzione χ^2.
Se il test statistico prevede una direzione della differenza tra valore calcolato e valore critico della variabile (cioè, $\mu > \nu$) si parla di **test a una coda**, in cui la zona di rifiuto dell'ipotesi nulla basata sulla variabile considerata è localizzata ad un estremo della curva di distribuzione, mentre se il test valuta solo la differenza tra valore calcolato e valore critico della variabile (cioè, $\mu \neq \nu$) si parla di **test a due code**, in cui la zona del rifiuto è localizzata in entrambi gli estremi della distribuzione.

L'effettuazione di un test statistico consiste generalmente di 5 passi:

1. definire l'ipotesi nulla e l'ipotesi alternativa;
2. scegliere il test statistico appropriato;
3. specificare il livello di significatività del test;
4. stabilire la regola di decisione in base al test e al livello di significatività prescelti;
5. calcolare il valore del test, confrontarlo col valore critico ed applicare la regola di decisione.

I tre test statistici più comuni sono elencati qui di seguito.

🗀 test t di Student

Il test *t* di Student viene utilizzato per confrontare tra loro due valori medi. In particolare, possono essere stabilite le seguenti due ipotesi a due code:

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

Ad esempio, ci possiamo chiedere se le medie $\overline{x}_1$ *e* $\overline{x}_2$ calcolate da due campioni provenienti da due popolazioni diverse siano significativamente differenti. Fissata come probabilità a cui vogliamo effettuare il test α = 0.05, il test t di Student è:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

dove s_p^2 è la varianza *pooled* dei due campioni:

$$s_p^2 = \frac{SS_1 + SS_2}{\nu_1 + \nu_2}$$

🗀 test F di Fisher

Il test F di Fisher consente di confrontare tra loro due varianze ed è quindi utilizzato per verificare l'ipotesi che due varianze siano tra loro statisticamente diverse ad un dato livello di significatività. Nel caso di un test a due code possono essere quindi formulate le due ipotesi:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

Il valore del test F viene calcolato dal rapporto tra le due varianze (la più grande generalmente al numeratore) e viene confrontato con il valore di tabella corrispondente ai gradi di libertà al numeratore e al denominatore.

Ci possiamo chiedere quale è la probabilità che le varianze s_1^2 *e* s_2^2 calcolate da due campioni provenienti da due popolazioni siano significativamente diverse.
Ad esempio, se la varianza calcolata dal primo campione di 11 oggetti è 21.87 e la varianza calcolata dal secondo campione di 8 oggetti è di 15.36, il valore di F è viene calcolato dalla formula:

$$F = \frac{s_1^2}{s_2^2} = \frac{21.87}{15.36} = 1.42$$

Se scegliamo come valore di probabilità a cui effettuare il test α = 0.05, il valore critico di F_C , per 10 e 7 gradi di libertà, è:

$$F_c \equiv F_{0.05(2),10,7} = 4.76$$

Quindi, poichè $F < F_c$ alla probabilità del 5%, non possiamo rifiutare l'ipotesi nulla H_0.

🗀 test chi-quadrato (χ^2)

Il test χ^2 consente di confrontare le frequenze osservate per un certo numero di eventi con le frequenze teoriche attese per gli stessi eventi. Quindi, diversamente dai due test precedenti, il test χ^2 si applica a osservazioni che vengono espresse come frequenze. Il test viene calcolato mediante la seguente espressione:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

dove O_i e E_i sono rispettivamente le frequenze osservate e quelle teoriche per l'*i*-mo campione. I gradi di libertà del test sono *n* - 1, cioè il numero di campioni meno uno.
Ad esempio, se in 5 stabilimenti il numero di pezzi non idonei di un processo di produzione sono 23, 10, 15, 3, 4, si vuole verificare l'ipotesi nulla che non vi sia una differenza di affidabilità nel processo di produzione dei 5 stabilimenti.

Le frequenze teoriche dei pezzi non idonei sono 55 / 5 = 11. Il calcolo è descritto in Tab. B-1.
Il valore di tabella per 5 - 1 = 4 gradi di libertà è 9.49, allo 0.05% di significatività: poichè il valore calcolato (24.90) è maggiore del valore di tabella (9.49), l'ipotesi nulla di indifferenza deve essere respinta.

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
23	11	12	13.09
10	11	-1	0.09
15	11	4	1.45
3	11	-8	5.82
4	11	-7	4.45
55	55	0	$\chi^2 = 24.90$

TAB. B-1

🕮 Bibliografia

J.H.Zar (1984). *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, N.J. (USA)

M.R.SPIEGEL (1973). *Statistica*. Etas-Libri, Milano.

Appendice C

ASPETTI MATEMATICI DELLE COMPONENTI PRINCIPALI

In questa appendice riportiamo alcune dimostrazioni matematiche che riguardano l'analisi delle componenti principali e il metodo di regressione PLS.

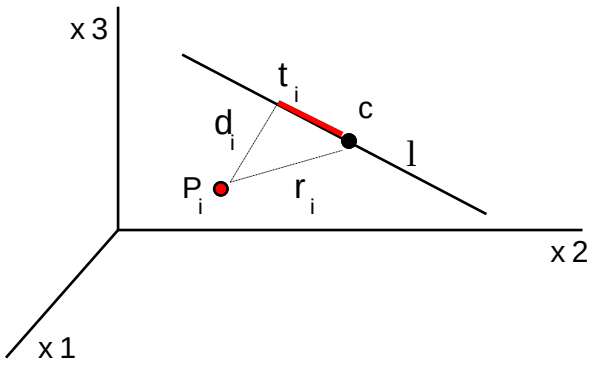
Interpretazione geometrica delle componenti principali

Sia **c** il punto nello spazio dei descrittori che corrisponde al valor medio di ciascun descrittore (il *centroide*):

$$\mathbf{c} = \left(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathrm{K}\ , \bar{\mathbf{x}}_p\right)$$

Sia $\mathbf{P}_i$ il punto corrispondente all'*i*-esimo dato:

$$\mathbf{P}_i = \left(x_{i1}, x_{i2}, \mathrm{K}\ , x_{ip}\right)$$



Sia **l** il vettore unitario con origine in **c**, sia d_i la distanza perpendicolare a **l** tra $\mathbf{P}_i$ e **l** (la proiezione di $\mathbf{P}_i$ su **l**), sia $\mathbf{r}_i$ la distanza tra $\mathbf{P}_i$ e **c**. Sia $\mathbf{t}_i$ la distanza dall'origine **c** del vettore **l** e la proiezione del punto $\mathbf{P}_i$.

Per il teorema di Pitagora, vale la seguente relazione:

$$r_i^2 = d_i^2 + t_i^2$$

Per *n* punti (i dati), la somma dei quadrati delle distanze $\mathbf{r}_i$ è costante:

$$\sum_i r_i^2 = \sum_i d_i^2 + \sum_i t_i^2 = cos\,tan\,te$$

Da quanto detto ne consegue che minimizzare $\sum_i d_i^2$ equivale a massimizzare $\sum_i t_i^2$, cioè ottenere una proiezione di tutti i punti che abbia la massima varianza lungo la direzione definita dal vettore **l**.
Il valore di $\mathbf{t}_i$ è dato dal prodotto scalare tra $\mathbf{r}_i$ e **:**

$$t_i = r_i \cdot \mathbf{l}$$

Sia $y = \sum_i t_i^2$. Si ha allora:

$$y = \left(\mathbf{X} - \overline{\mathbf{X}}\right)\cdot \mathbf{l}^{\mathbf{T}} \cdot \left(\mathbf{X} - \overline{\mathbf{X}}\right)\cdot \mathbf{l} = \mathbf{l}^{\mathbf{T}} \cdot \left(\mathbf{X} - \overline{\mathbf{X}}\right)^{\mathbf{T}} \cdot \left(\mathbf{X} - \overline{\mathbf{X}}\right)\cdot \mathbf{l}$$

Se introduciamo il vincolo che il vettore **l** sia normalizzato a lunghezza unitaria, cioè $\mathbf{l}^{\mathbf{T}}\mathbf{l} = \mathbf{1}$, possiamo scrivere il vincolo

$$C = 1 - \mathbf{l}^{\mathbf{T}}\mathbf{l} = 0$$

Ora, per massimizzare *y* sotto il vincolo *C*, utilizziamo il metodo dei moltiplicatori di Lagrange per determinare il vettore **l** che massimizza la funzione:

$$L = y + \lambda C$$

$$L = \mathbf{l}^{\mathbf{T}} \cdot \left(\mathbf{X} - \overline{\mathbf{X}}\right)^{\mathbf{T}} \cdot \left(\mathbf{X} - \overline{\mathbf{X}}\right)\cdot \mathbf{l} + \lambda \cdot (1 - \mathbf{l}^{\mathbf{T}}\mathbf{l})$$

Ponendo $\partial L/\partial \mathbf{l} = 0$ per la ricerca del massimo, otteniamo:

$$\left(\mathbf{X}-\overline{\mathbf{X}}\right)^{\mathbf{T}}\cdot\left(\mathbf{X}-\overline{\mathbf{X}}\right)\cdot\mathbf{l}=\lambda\cdot\mathbf{l}$$

Vediamo quindi che λ è l'autovalore e **l** è un autovettore della matrice reale simmetrica $\left(\mathbf{X}-\overline{\mathbf{X}}\right)^{\mathbf{T}}\cdot\left(\mathbf{X}-\overline{\mathbf{X}}\right)$, il cui rango è determinato dal numero di autovalori non nulli.
Questa matrice è proporzionale alle matrici di covarianza e di correlazione dei dati **X**.

Per ogni vettore $\mathbf{l_m}$ otteniamo il vettore degli *scores* $\mathbf{t_m}$, secondo la relazione:

$$\mathbf{t}_m=\left(\mathbf{X}-\overline{\mathbf{X}}\right)\cdot\mathbf{l}_m$$

cioè il vettore che contiene tutti i valori degli oggetti proiettati lungo l'*m*-esima componente.

In termini matriciali, questo diventa:

$$\mathbf{T}=\left(\mathbf{X}-\overline{\mathbf{X}}\right)\cdot\mathbf{L}=\mathbf{X}_c\cdot\mathbf{L}$$

dove **T** è la matrice degli scores e **L** è la matrice dei loadings.
Moltiplicando a destra per $\mathbf{L}^{\mathbf{T}}$ e ricordando che **L** è una matrice di autovettori ortogonali ($\mathbf{L}^{\mathbf{T}}=\mathbf{L}^{-\mathbf{1}}$), si ricava:

$$\left(\mathbf{X}-\overline{\mathbf{X}}\right)=\mathbf{X}_c=\mathbf{T}\cdot\mathbf{L}^{\mathbf{T}}$$

Giustificazione dell'algoritmo PLS

Definendo $\mathbf{O_X}$ e $\mathbf{O_y}$ come due matrici di rotazione ortogonali e **S** come la matrice di covarianza, valgono le relazioni:

$$\mathbf{S} = \mathbf{XO_x} \qquad \sum \mathbf{s}_{ij}^2 = tr(\mathbf{S^T S}) = tr(\mathbf{X^T X}) = \sum \mathbf{x}_{ij}^2$$

e

$$\mathbf{Z} = \mathbf{YO_y}$$

Siano $\mathbf{s_i}$ e $\mathbf{z_i}$ sono le colonne di **S** e di **Z,** e supponiamo che *p* sia maggiore di *r*. La distanza tra i due vettori $\mathbf{s_i}$ e $\mathbf{z_i}$ è data da:

$$\mathbf{d} = \sum_{1}^{r} |\mathbf{s}_i - \mathbf{z}_i|^2 + \sum_{r+1}^{p} |\mathbf{s}_i|^2$$

e il valore minimo di **d** è:

$$\min(\mathbf{d}) = tr(\mathbf{X^T X}) + tr(\mathbf{Y^T Y}) - 2\sum\nolimits_m \lambda_m$$

Nel metodo PLS viene selezionata solo una coppia di componenti alla volta: cioè, **d** viene ogni volta ridotta della quantità $2\lambda_1$. La scelta di una sola componente per ogni iterazione è motivata dal fatto che la quantità $2\lambda_2$ (corrispondente alla seconda componente in ogni iterazione) è minore della quantità $2\lambda_1$ all'iterazione successiva.

Appendice D

I DATI UTILIZZATI NEGLI ESEMPI

Dati: REGTEST

Sono dati generici utilizzati per studiare i modelli di regressione. Sono 17 dati relativi a 5 variabili indipendenti e a una risposta.

ID	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1603.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1584.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20106	3655.08	180.5	6.15	3503.93
11	96.00	13313	2912.00	60.9	5.88	3571.89
12	131.42	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34783	12446.33	169.4	10.78	11732.17
16	463.70	39204	14098.4	331.4	7.05	15414.94
17	510.22	86533	15524.00	371.6	6.35	18854.45
media	148.27	18163.2	4480.62	106.32	5.89	4978.48
dev.std.	156.23	20642.8	4760.14	104.73	1.54	5394.51

Dati: WINES

I dati sono relativi a 17 misure chimiche di 38 campioni di vino. Questi campioni sono stati sottoposti ad un panel test per la valutazione dell'aroma: il giudizio medio ottenuto è riportato nell'ultima colonna (Aroma).

ID	*Cd*	*Mo*	*Mn*	*Ni*	*Cu*	*Al*	*Ba*	*Cr*	*Sr*	*Pb*
1	.005	.044	1.51	.122	.83	.982	.387	.029	1.23	.561
2	.055	.16	1.16	.149	.066	1.02	.312	.038	.975	.697
3	.056	.146	1.1	.088	.643	1.29	.308	.035	1.14	.73
4	.063	.191	.959	.38	.133	1.05	.165	.036	.927	.796
5	.011	.363	1.38	.16	.051	1.32	.38	.059	1.13	1.73
6	.05	.106	1.25	.114	.055	1.27	.275	.019	1.05	.491
7	.025	.479	1.07	.168	.753	.715	.164	.062	.823	2.06
8	.024	.234	.906	.466	.102	.811	.271	.044	.963	1.09
9	.009	.058	1.84	.042	.17	1.8	.225	.022	1.13	.048
10	.033	.074	1.28	.098	.053	1.35	.329	.03	1.07	.552
11	.039	.071	1.19	.043	.163	.971	.105	.028	.491	.31
12	.045	.147	2.76	.071	.074	.483	.301	.087	2.14	.546
13	.06	.116	1.15	.055	.18	.912	.166	.041	.578	.518
14	.067	.166	1.53	.041	.043	.512	.132	.026	.229	.699
15	.077	.261	1.65	.073	.285	.596	.078	.063	.156	1.02
16	.064	.191	1.78	.067	.552	.633	.085	.063	.192	.777
17	.025	.009	1.57	.041	.081	.655	.072	.021	.172	.232
18	.02	.027	1.74	.046	.153	1.15	.094	.021	.358	.025
19	.034	.05	1.15	.058	.058	1.35	.294	.006	1.12	.206
20	.013	.03	2.82	.058	.05	.623	.349	.082	2.91	.171
21	.043	.268	2.32	.066	.314	.627	.099	.045	.36	1.28
22	.061	.245	1.61	.07	.172	2.07	.071	.053	.186	1.19
23	.047	.161	1.47	.154	.082	.546	.181	.06	.898	.747
24	.048	.146	1.85	.092	.09	.889	.328	.1	1.32	.604
25	.049	.155	1.73	.051	.158	.653	.081	.037	.164	.767
26	.042	.126	1.7	.112	.21	.508	.299	.054	.995	.686
27	.058	.184	1.28	.095	.058	1.3	.346	.037	1.17	1.28
28	.065	.211	1.65	.102	.055	.308	.206	.028	.72	1.02
29	.065	.129	1.56	.166	.151	.373	.281	.034	.889	.638
30	.068	.166	3.14	.104	.053	.368	.292	.039	1.11	.831
31	.067	.199	1.65	.119	.163	.447	.292	.058	.927	1.02
32	.084	.266	1.28	.087	.071	1.14	.158	.049	.794	1.3
33	.069	.183	1.94	.07	.095	.465	.225	.037	1.19	.915
34	.087	.208	1.76	.061	.099	.683	.087	.042	.168	1.33
35	.074	.142	2.44	.051	.052	.737	.408	.022	1.16	.745
36	.084	.171	1.85	.088	.038	1.21	.263	.072	1.35	.899
37	.106	.307	1.15	.063	.051	.643	.29	.031	.885	1.61
38	.102	.342	4.08	.065	.077	.752	.366	.048	1.08	1.77

ID	*B*	*Mg*	*Si*	*Na*	*Ca*	*P*	*K*	*Aroma*
1	2.63	128	17.3	66.8	80.5	150	1130	3.3
2	6.21	193	19.7	53.3	75	118	1010	4.4
3	3.05	127	15.8	35.4	91	161	1160	3.9
4	2.57	112	13.4	27.5	93.6	120	924	3.9
5	3.07	138	16.7	76.6	84.6	164	1090	5.6
6	6.56	172	18.7	15.7	112	137	1290	4.6
7	4.57	179	17.8	98.5	122	184	1170	4.8
8	3.18	145	14.3	10.5	91.9	187	1020	5.3
9	6.13	113	13	54.4	70.2	158	1240	4.3
10	3.3	140	16.3	70.5	74.7	159	1100	4.3
11	6.56	103	9.47	45.3	67.9	133	1090	5.1
12	3.5	199	9.18	80.4	66.3	212	1470	3.3
13	6.43	111	11.1	59.7	83.8	139	1120	5.9
14	7.27	107	6	55.2	44.9	148	854	7.7
15	5.04	94.6	6.34	10.4	54.9	132	899	7.1
16	5.56	110	6.96	13.6	64.1	167	976	5.5
17	3.79	75.9	6.4	11.6	48.1	132	995	6.3
18	4.24	80.9	7.92	38.9	57.6	136	876	5
19	2.71	120	14.7	68.1	64.8	133	1050	4.6
20	3.54	208	9.32	79.2	66.4	266	1430	3.4
21	5.68	98.4	9.11	19.5	64.3	176	945	6.4
22	4.42	87.6	7.62	11.6	70.6	156	820	5.5
23	8.11	160	19.3	12.5	82.1	218	1220	4.7
24	6.42	134	19.3	125	83.2	173	1810	4.1
25	4.91	86.5	6.46	11.5	53.9	172	1020	6
26	6.94	129	43.6	45	85.9	165	1330	4.3
27	3.29	145	16.7	65.8	72.8	175	1140	3.9
28	6.12	99.3	27.1	20.5	95.2	194	1260	5.1
29	7.28	139	22.2	13.3	84.2	164	1200	3.9
30	4.71	125	17.6	13.9	59.5	141	1030	4.5
31	6.97	131	38.3	42.9	85.9	164	1390	5.2
32	3.77	143	19.7	39.1	128	146	1230	4.2
33	2	123	4.57	7.51	69.4	123	943	3.3
34	5.04	92.9	6.96	12	56.3	157	949	6.8
35	3.94	143	6.75	36.8	67.6	81.9	1170	5
36	2.38	130	6.18	101	64.4	98.6	1070	3.5
37	4.4	151	17.4	7.25	103	177	1100	4.3
38	3.37	145	5.33	33.1	58.3	117	1010	5.2

Dati: GRANULI

I dati riguardano l'ottimizzazione di un processo di rivestimento mediante materiale in grani. Sono state misurate 6 diverse caratteristiche del materiale in 14 prove. Queste prove sono state ottenute variando 3 parametri di processo (temperatura, velocità di spray, pressione dell'aria di atomizzazione) secondo un disegno sperimentale centrato.

ID	*Uniformità dei grani*	*Densità di ingombro*	*Densità utile*	*Dimensione particelle*	*Dissoluzione in acido*	*Quantità di sostanza %*
1	1.35	0.64	0.75	103.6	57.3	12.8
2	1.01	0.62	0.72	113.8	56.6	13.3
3	3.31	0.65	0.77	100.5	59.9	12.8
4	0.98	0.61	0.68	139.4	25.2	13.5
5	1.62	0.67	0.78	106.8	74.1	12.6
6	1.19	0.64	0.70	124.3	37.6	13.2
7	1.30	0.63	0.72	114.6	39.3	13.2
8	1.04	0.61	0.70	127.8	41.1	13.2
9	1.24	0.64	0.75	94.9	60.5	13.7
10	1.86	0.66	0.74	99.2	57.2	13.4
11	1.39	0.64	0.73	114.4	47.6	13.3
12	1.01	0.60	0.72	107.8	40.8	13.4
13	1.30	0.65	0.75	103.9	47.8	13.2
14	1.30	0.60	0.70	130.5	32.8	13.2

Dati: AUTO

I dati riguardano 48 autovetture di piccola cilindrata per le quali sono state selezionati da una rivista specializzata (maggio 1995) i seguenti parametri: cilindrata (cc), potenza (hp), consumo urbano (curb), consumo a 120 km/h (c10), comfort (comf), sicurezza (sicur), costo di esercizio (cese), costo totale (ctot).

ID	*Auto*	*cc*	*hp*	*vmax*	*curb*	*c120*	*comf*	*sicur*	*cese*	*ctot*
1	AUBI-Y10_Junior	1108	55	155	13.7	16.1	4.5	5	156	15300
2	AUBI-Y10_Igloo	1108	55	155	13.7	16.1	5	5	156	16750
3	CITR-AX_I3P	954	50	149	14.9	15.6	5	5	151	14990
4	CITR-AX_I5P	954	50	149	14.9	15.6	5.5	5	151	16000
5	CITR-AX_ITZX_5P	1124	60	167	14.1	15.2	5.5	5	158	19000
6	FIAT-500_700ED	704	30	126	15.4	15.4	4.5	5	132	11300
7	FIAT-500Sporting	1108	54	150	13.3	15.9	4.5	5	153	14800
8	FIAT-Panda_Young	899	39	135	15.2	14.5	4.5	5	150	13250
9	FIAT-Panda:C_Club	1108	50	130	12	10.5	4.5	5	186	20450
10	FIAT-Uno_Start3p	994	48	145	13.2	14.5	5	5	159	14500
11	FIAT-Uno_Start5P	994	48	145	13.2	14.5	5.5	5	159	15500
12	FIAT-Uno_Cond5P	1372	69	165	10.6	13.9	6	5	186	17650
13	FIAT-Punto_6_Speed	1108	54	150	14.3	14.5	5.5	5	160	18050
14	FIAT-Punto_S55_5P	1108	54	150	12.7	15.4	6	5	159	17050
15	FIAT-Punto_60_Sel5P	1242	60	150	13.3	13.5	6.5	5	169	21850
16	FORD-Fiesta_Caym3P	1118	50	143	13.9	15.2	5.5	5.5	157	16000
17	FORD-Fiesta_Caym5P	1118	50	143	13.9	15.2	6	5.5	157	17000
18	FORD-Fiesta_Wind5P	1297	60	150	12.5	15	6	5.5	175	18500
19	HOND-Civic_EX3P	1343	75	170	12.5	13.2	6.5	5	191	23500
20	HYUN-Accent_LS3P	1341	84	174	14	16.3	5.5	5	174	16200
21	HYUN-Accent_HS5P	1341	84	174	14	16.3	7	6	174	24050
22	INNO-Mille_3P	994	48	145	12	13.7	5	5	166	13400
23	INNO-Mille_5P	994	48	145	12	13.7	5.5	5	166	14400
24	MAZD-LX4P	1324	54	145	13.2	14.1	6.5	5	174	18802
25	MAZD-Cabrio_LX4P	1324	73	150	13.5	13.2	6.5	5	176	21777
26	NISS-Micra_3P	998	55	149	16.7	14.9	5.5	5	151	15750
27	NISS-Micra_SLX5P	1275	75	170	15.2	14.9	6	5	169	19610
28	OPEL-Corsa_City3P	1195	45	145	13.9	16.1	5.5	5	154	15760
29	OPEL-Corsa_City5P	1195	45	145	13.9	16.1	6	5	154	16660
30	OPEL-Corsa_Swing5P	1195	45	145	13.2	14.9	6.5	5	165	18820
31	PEUG-106_Holl3P	954	50	150	14.1	14.9	5	5	157	14990
32	PEUG-106_XN5P	954	50	150	14.1	14.9	5.5	5	157	16125
33	PEUG-106_XT5P	1124	60	165	13.5	14.9	5.5	5	163	21400

34	RENL-Clio_RL3P	1171	55	150	13.3	16.7	5.5	5	159	15950

ID	*Auto*	*cc*	*hp*	*vmax*	*curb*	*c120*	*comf*	*sicur*	*cese*	*ctot*
35	RENL-Clio_RL5P	1171	55	150	13.3	16.7	6	5	159	16950
36	RENL-Clio_RTI5P	1171	55	150	13.3	16.7	6.5	5.5	159	20800
37	RENL-Twingo	1239	55	150	13.5	14.3	5.5	5	164	14950
38	RENL-Twingo_Spring	1239	55	150	13.5	14.3	5.5	5	164	16750
39	ROVE-Mini_B_Open	1275	50	138	13.2	11.9	4.5	5	181	15295
40	ROVE-Mini_Cooper	1275	63	152	13	11.9	4.5	5	187	15788
41	ROVE-111_Si3P	1119	60	153	14.3	15.6	5	5	155	15303
42	ROVE-111_Si5P	1119	60	153	14.3	15.6	5.5	5	155	16030
43	SEAT-Ibiza_Cli3P	1043	45	135	12.7	13.9	5.5	5	167	15721
44	SKOD-Felicia_LX	1289	54	145	12.5	13	5.5	5	184	12990
45	SKOD-Felicia_GLXi	1289	68	150	12.7	13.5	5.5	5	180	14560
46	VOLK-Polo_1.0_3P	1043	45	145	13.3	14.1	5.5	5	165	15996
47	VOLK-Polo_1.0_5P	1043	45	145	13.3	14.1	6	5	165	16870
48	VOLK-Polo_1.3_5P	1296	55	156	12.7	14.7	6	5	174	19482