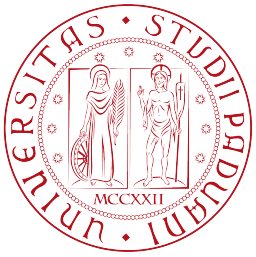
Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica, Economia e Finanza



Relazione Finale

Analisi chemiometriche tramite spettrofotometria

Relatore:
Prof. Bruno Scarpa

Laureando:
Alberto Bianchi
Matricola 1010874

Anno accademico 2013/2014

Indice

Introduzione **5**

1 La chemiometria **7**
1.1 Origini 7
1.2 Chemometrics 9
1.3 L'attività di stage 12

2 I metodi **15**
2.1 Analisi delle Componenti Principali 18
2.2 I modelli di regressione: *PCR* e *PLS* 21
2.2.1 Il pretrattamento dei dati 22
2.2.2 Adattamento dei modelli 23
2.2.3 Outlier 26
2.3 *Principal Component Regression* (*PCR*) 27
2.4 *Partial Least Square Regression* (*PLSR*) 29

3 Applicazione **31**
3.1 I dati 31
3.2 Analisi esplorative 32
3.3 Calibrazione 35
3.3.1 Modelli di regressione *PCR* 37
3.3.2 Modelli di regressione *PLS* 42

4 Conclusioni **47**

Bibliografia **49**

A Codice R **51**

Introduzione

La chemiometria è la scienza che applica la matematica e la statistica ai dati chimici. Questa disciplina ha visto l'apice del suo sviluppo in tempi molto recenti, basti pensare che la coniazione del termine *chemometrics*, avvenne meno di cinquant'anni fa. Tale recente sviluppo è dato da un lato dalle necessità, in termini di quantità e rapidità di controlli, e dall'altro dalle possibilità, in termini di velocità di elaborazione numerica.

In questa relazione finale di stage, si vedrà l'applicazione dei metodi chemiometrici alla spettrofotometria. Quest'ultima è volta ad estrarre, tramite radiazione luminosa, informazioni da una determinata sostanza. In modo specifico, si farà uso dell'informazione spettrometrica nel vicino infrarosso. É questo il principale campo applicativo in cui opera IOR3, la società per la quale è stato creato un *software* di previsione chemiometrica il cui scopo è l'integrazione nello spettrometro poliSPEC^{NIR}, di loro fabbricazione. L'informazione viene elaborata come discretizzazione spettrale ed è modellata tramite modelli statistici per determinare la composizione di un campione sottoposto ad analisi.

Nel primo capitolo si vedrà il lungo processo che portò alla definizione della chemiometria. L'atomismo rappresenta un primo tentativo, ad opera dei filosofi greci, di spiegare il funzionamento della materia. Grazie alle influenze arabe ed egizie, si giunse poi ad una nuova dottrina, l'alchimia. Successivamente, la rivoluzione scientifica portò tale corrente di pensiero a scindersi e a definire una nuova scienza, la chimica. Ai giorni nostri, le conoscenze chimiche sviluppate possono essere raggruppate in tre macro settori: la chimica teoria, la chimica sintetica e la chimica analitica, all'interno della quale si colloca la chemiometria. Saranno in seguito presentati alcuni campi applicativi per mettere in luce una disciplina che spesso rimane in ombra ma il cui uso è ormai molto diffuso ed ancora in

espansione. Si concluderà presentando l'attività di stage presso la società IOR3 per la quale è stato creato il *software* oggetto della relazione.

Nel secondo capitolo si entrerà in maggior dettaglio sull'acquisizione del dato spettrale e sul motivo per cui questo sia relazionabile alla composizione chimica. Saranno presentati alcuni aspetti teorici alla base delle tecniche di ottimizzazione, utilizzate per la soluzione di tipici problemi chemiometrici. Sono infatti oggetto di studio dataset con un gran numero di variabili, di cui non si conosce a priori la loro rilevanza o la presenza di collinearità. Per risolvere i problemi legati a questa matrice di dati verrà presentata l'analisi delle componenti principali (*PCA*), per facilitare la comprensione dei successivi modelli statistici: la regressione in componenti principali (*PCR*) e la regressione su strutture latenti (*PLSR*).

Nel terzo capitolo, infine, si vedrà l'applicazione dei metodi presentati ed il funzionamento dello script attraverso l'uso di un dataset di esempio, contenente l'acquisizione spettrale di campioni di soia. I modelli di regressione saranno infine confrontati tra loro per decidere quale sia la miglior strategia da utilizzare a seconda della quantità da prevedere.

Capitolo 1

La chemiometria

1.1 Origini

La chimica è una delle conoscenze più antiche dell'umanità. I primi tentativi per spiegare il funzionamento della materia risalgono agli antichi filosofi greci. Gli insegnamenti di Leucippo, Democrito e successivamente Epicuro, fecero sviluppare nel VII secolo a.C. l'atomismo, una corrente di pensiero, che già dagli albori teorizzava il mondo naturale come congiunzione di due parti: gli atomi invisibili e il vuoto. L'atomismo, fra le prime teorie a separare la sfera religiosa da quella scientifica, subì poi gli influssi della cultura araba ed egizia che portarono alla nascita dell'alchimia, affermatasi poi più come corrente filosofica esoterica che come scienza.

L'alchimia, che comprendeva al suo interno discipline come l'astrologia, la fisica, la medicina e la chimica, aveva tre obiettivi distinti ma collegati fra loro: la trasformazione di metalli vili in nobili attraverso la cosiddetta pietra filosofale, la creazione di un elisir di lunga vita capace di curare qualsiasi malattia, e la crescita spirituale dell'alchimista legata a un esperienza mistica e di illuminazione.

Quando la rivoluzione scientifica del XVII secolo d.C. segnò un cambiamento di pensiero in tutte le dottrine dell'uomo, anche l'alchimia prese parte a questo rinnovamento: si abbandonò quel misticismo che la caratterizzava a favore di un analisi scientifica della materia e delle sue trasformazioni, mirata alla formulazione di ipotesi ed esperimenti rivolti a confermare o confutare le teorie.

I primi testi che trattano di chimica vengono attribuiti al francese Jean

Béguin[1], che nel 1610 pubblicò *Tyrocinium Chymicum*. L'accelerazione della scissione fra le due materie e una più profonda comprensione della chimica avvenne, però, grazie alle scoperte dell'italiano Evangelista Torricelli[2] il quale elaborò un modo per misurare la pressione atmosferica e formulò il concetto di vuoto, dando un forte impulso allo studio dei gas.

Successivamente, attraverso l'uso sempre più scrupoloso del metodo sperimentale, l'inglese Robert Boyle[3] riprese lo studio dei gas pubblicando nel 1661 le sue scoperte insieme alle prime imprecise definizioni di composto chimico. *The Sceptical Chymist*, che diede alle stampe nello stesso anno, è considerato simbolicamente il testo spartiacque fra chimica e alchimia.

Nel 1700 vi fu la necessità di riunire tutte le conoscenze sviluppate fino a quel momento nel campo dei gas e fu Antoine Lavoisier[4] che adempì a tal compito. Oltre ad aver determinato la corretta composizione dell'aria, scoprì la presenza di "mattoncini" base (idrogeno, ossigeno e carbonio) presenti nelle sostanze organiche. Enunciò anche la famosa "legge di conservazione"[5] demolendo così la "teoria del flogisto"[6] e segnando il definitivo superamento dell'alchimia.

Il vero salto di qualità però avvenne dapprima a Londra, quando fu fondato nel 1845 il *Royal College of Chemistry*, e successivamente in Germania dove nel 1860, si riunirono studiosi da tutta Europa con l'intento di ridefinire i concetti basilari della chimica e creare un'unica nomenclatura universale.

Fino al XX secolo, la chimica presentava pochi punti in comune con la fisica tant'è che nel 1830 Comte affermava che tentare di utilizzare metodi matematici allo studio della chimica era «profondamente irritante» e ad-

[1] Chimico francese del XVI secolo.

[2] Fisico e matematico italiano del XVII secolo; amico, discepolo e collaboratore di Galilei.

[3] Scienziato eclettico del XVII secolo, fra l'altro fisico e chimico, ipotizzò un modello della materia molto simile a quello oggi accertato dalla comunità scientifica.

[4] Universalmente riconosciuto come il padre della chimica, fu un chimico e biologo del XVIII secolo.

[5] La "legge della conservazione della massa" fu formulata da Lavoisier, il quale osservò che, in una reazione chimica, la massa complessiva dei reagenti è uguale alla massa complessiva dei prodotti, da cui venne poi parafrasato il famoso "nulla si crea, nulla si distrugge, tutto si trasforma".

[6] Un misterioso principio di infiammabilità introdotto per spiegare l'ossidazione e la combustione.

dirittura «contrario allo spirito della chimica». Quando venne scoperta e studiata la radioattività degli elementi, attorno al 1800, si capì che l'atomo non era una particella indivisibile e gli scienziati cambiarono il loro punto di vista.

1.2 Chemometrics

Data la versatilità delle conoscenze chimiche non stupisce come esse abbiano dato vita, oltre alla classica chimica organica ed inorganica, ad una varietà di applicazioni come quella industriale, medica, ambientale e molte altre, tant'è che oggi viene vista come scienza centrale.

Tutte queste divisioni possono essere raggruppate in tre grandi settori: la chimica teorica, la chimica sintetica e la chimica analitica. La prima è la branca che ha come scopo la definizione dei fondamenti teorici, facendo uso di principi matematici e fisici. La seconda ha lo scopo di ottenere un dato composto tramite reazione chimica. L'ultima, infine, ha come finalità la determinazione qualitativa e quantitativa dei componenti di un determinato campione attraverso l'uso di metodi matematici e statistici.

La chimica analitica, la più giovane delle tre, ha visto il suo massimo sviluppo nel 1950, grazie alla modernizzazione della strumentazione: la bilancia e il polarografo, gli apparecchi più classici, furono soppiantati da strumenti con componenti elettroniche. Le innovazioni tecnologiche portarono ad una cultura nuova all'interno della fisica. Le classiche conoscenze teoriche si fusero con aspetti prettamente strumentali quali l'elettronica e la "teoria dei segnali"[7]. Questo si è reso necessario poiché non si misura direttamente la quantità chimica in analisi, ma le caratteristiche fisiche del campione sottoposto ad eccitazione, osservandone l'agitazione molecolare data dall'esposizione del campione a radiazione luminosa.

I nuovi strumenti non solo godono di maggior precisione (che porta a determinare sia le quantità in maniera più precisa, sia a cogliere le concentrazioni particolarmente basse) ma anche di costi minori rispetto alle classiche analisi chimiche in laboratorio, ancora usate per calibrare i mo-

[7] L'oggetto dello studio è il segnale, definito come funzione matematica del tempo. Esso è una variazione nel tempo dello stato fisico della materia utile, in questo contesto, a raccogliere informazione a distanza.

delli statistici, ma non più l'unica via per determinare la composizione oggetto di studio.

Gradualmente la chimica cominciò a sfruttare la statistica e l'informatica, riorganizzandosi in una nuova dottrina, la chemiometria. Fra il 1970 e il 1980 si affermò come disciplina indipendente all'interno della chimica analitica, interessando l'opinione pubblica a tal punto da far apparire, in alcune riviste scientifiche, specifiche sezioni dedicate. Poco più tardi nacquero riviste interamente dedicate alla chemiometria come *Journal of Chemometrics* e *Chemometrics and Intelligent Laboratory Systems*.

Il termine "chemometrics" venne coniato all'inizio del 1970 da Svante Wold, docente all'Università di Omeö in Svezia, e da Bruce Kowalski, dell'Università di Seattle in America. I due nel 1974 fondarono l'*International Chemometrics Society* definendo la chemiometria come:

> Chemometrics is the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments; and (b) to provide maximum chemical information by analyzing chemical data.
>
> *(Bruce Kowalski, in a formal CPAC presentation, December 1997)*

Da questa definizione possiamo ricavare le argomentazioni chiave della chemiometria: l'estrazione della massima informazione utile dal campione, l'ottimizzazione dei metodi e la convalida dei risultati. L'utilizzo sistematico di tecniche di validazione, nell'ambito della convalida dei risultati, è una peculiarità della disciplina nonché una caratteristica qualificante dei modelli statistici utilizzati. L'attenzione di tali modelli, infatti, non è posta sulla capacità degli stessi di adattarsi ai dati o di descriverne il principio generatore, bensì sulla capacità che questi hanno di prevedere la quantità oggetto di studio.

L'attenzione rivolta ai dati è di centrale importanza durante tutto il corso dell'analisi, tant'è che diversi importanti esponenti della chemiometria iniziarono i loro studi nel campo chimico per poi spostare l'attenzione verso la sfera statistica. Fra questi si ricorda William Gosset, meglio conosciuto come Student, e George Box.

Il contributo apportato dalla matematica e dalla statistica alla chemiometria risulta evidente nello sviluppo dei metodi per l'analisi di questi

dati. Buona parte delle tecniche che sono state escogitate per rispondere alle nuove domande poste dalla chemiometria, però, non avrebbero trovato né risposta né così ampia applicabilità, senza lo sviluppo parallelo dell'informatica, che ha permesso di soddisfare le crescenti richieste di controllo sui prodotti della società, sia in termini di numero di analisi sia in termini di velocità.

La stessa Unione Europea si avvale di tecniche chemiometriche nell'ambito dell'analisi ambientale ed in quello alimentare per verifiche di qualità, di origine e di tossicità per la loro versatilità: basti pensare che nel 2006 la sola analisi sulle carni per la presenza di antibattericidi portò ad esaminare più di 230 000 campioni (Forina 2010).

A testimonianza di come la chemiometria abbia superato il contesto puramente chimico da cui è nata, basta dare uno sguardo alle varie applicazioni in cui attualmente viene impiegata. Grazie anche all'utilizzo di innovazioni elettroniche ed ai computer, i metodi associati alla chemiometria risultano essere una valida soluzione, se non a volte l'unica strada percorribile, per estrarre informazioni volte alla risoluzione di problemi di alta complessità, per i quali non esistono delle strategie teoriche ben fondate o procedure predeterminate. Attraverso tali metodi le conoscenze della chimica, della farmacologia, delle scienze ambientali e di molti altri settori, sono state verificate, modificate e rielaborate, spesso presentando anche soluzioni innovative.

Molte industrie europee ed americane adottano ed investono in queste nuove metodologie per l'elevato rapporto fra benefici e costi. Il crescente interesse economico e pratico ha fatto sì che le grandi industrie abbiano preceduto sia i centri di ricerca che le università, nell'utilizzo di queste tecniche. Per svolgere un'analisi, infatti, tutto ciò che occorre è un buon calcolatore, alcuni testi teorici di riferimento, degli specifici pacchetti *software* e del personale specializzato che, fra gli altri requisiti, è quello più dispendioso in termini di istruzione. Per la formazione del personale in Europa, infatti, è nato il progetto *Eurochemometrics*, una rete di scuole di chemiometria coordinata da universitari e non, che da qualche anno dà voce a questa materia, tenendo sia corsi professionalizzanti per settore sia seminari introduttivi.

I tre centri universitari italiani in cui vengono tenuti corsi specifici di

chimica analitica, fanno del nostro paese uno degli importanti poli europei di sviluppo della chemiometria. L'obiettivo in ambito accademico è la formazione di una nuova figura professionale in grado, non solo di far fronte alla rapidità dei cambiamenti ed alla continua innovazione, ma anche di saper proporre nuove strategie e nuove metodologie per rivalutare i grandi database di aziende ed enti pubblici. Il fine è quello di estrarre nuova informazione, mirata alla soluzione di nuovi obbiettivi, nel campo sociale e nel controllo della salute e dell'ambiente, ambito in cui diverse regioni italiane hanno già riposto il loro crescente interesse.

A partire dal 1990, i ricercatori di vari settori iniziarono ad applicare le conoscenze chemiometriche ai campi più disparati, virando verso applicazioni sempre più connesse con il settore industriale.

Uno dei motivi centrali per cui la chemiometria ha trovato una così ampia applicabilità in più settori è proprio il dato chimico in sé, poiché possiede delle peculiarità che lo rendono unico. La composizione chimica, cioè la costituzione della materia e la decomposizione nei suoi costituenti chimici puri, trascende l'ambito puramente chimico ed è conoscenza generale di tutto l'ambito scientifico. Altresì la struttura molecolare, importante mediatore fra la composizione atomica e le sue proprietà fisiche, ha permesso di applicare questa giovane disciplina ad una vasta gamma di settori. Alcuni degli usi in cui ritroviamo la chemiometria concernono il campo clinico, il campo geologico minerario, il campo delle analisi ambientali, il campo agrario e molti altri.

1.3 L'attività di stage

Lo sviluppo del *software* che verrà presentato in questa relazione finale, è stato creato per la società IOR3 di Casalserugo. Lo scopo dello stage è stato quello di replicare il più possibile, tramite uno script sviluppato in ambiente R, il *software* in loro possesso. Il gruppo si occupa della creazione di spettrometri che sfruttano per l'acquisizione di dati per una successiva elaborazione tramite analisi chemiometriche. La società analizza prevalentemente le lunghezze d'onda del vicino infrarosso, lavorando nell'ambito del controllo della qualità e della composizione di alimenti e mangimi.

Lo scopo finale del *software* è quello di essere integrato nello spettrometro poliSPEC^{NIR} di produzione IOR3. Sebbene vi siano già sul mercato diversi programmi chemiometrici, lo scopo della società è quello di far raggiungere un nuovo livello alla versatilità dell'analisi, portando direttamente nello strumento, la visualizzazione della risposta dell'analisi, senza dover passare necessariamente per un calcolatore (a meno del processo di calibrazione).

Questa gamma di mercato, del tutto rivoluzionaria, è mirata ai piccoli e medi imprenditori che hanno bisogno di utilizzare questo tipo di metodologie nel modo più semplice possibile. L'intento nella creazione, sia del *software* vero e proprio sia dell'interfaccia per il suo utilizzo, è quello di rendere l'apparecchio facilmente utilizzabile da tutti.

Il *software* su cui i tecnici della società attualmente lavorano è WinISI, creato dalla FOSS[8] attorno al 1990, ed è sostanzialmente rimasto uguale nel tempo dal punto di vista dell'elaborazione dei dati, dando la possibilità ad altri concorrenti di lanciarsi in questo mercato.

Dato il segreto sul codice e un tetro alone di mistero che avvolge WinISI, per la sua replicazione è stato usato un procedimento quasi di *reverse engineering*, confrontando il risultato di ogni singolo step del codice col rispettivo procedimento del *software* commerciale. La replicazione tramite R resterà invece in versione libera e sarà disponibile per la consultazione e la modifica da parte di coloro che si scontreranno con problemi di spettrometria simili. Alcune scelte effettuate nella creazione dello script, infatti, sono state fatte seguendo WinISI, sia per una confrontabilità dei risultati, sia per volontà della stessa società.

Lo script utilizzerà il pacchetto `chemometrics` (Garcia e Filzmoser 2011), una sorta di "coltellino svizzero" per analisi chemiometriche. Esso infatti, oltre a contenere librerie esterne utili in tale campo, ha al suo interno procedimenti specifici creati appositamente dagli autori del pacchetto che torneranno molto utili.

La replicazione del *software*, tuttavia, non vuole essere un punto d'arrivo bensì il punto di partenza verso lo sviluppo di metodologie più performanti ed affidabili.

[8]Compagnia danese, fondata nel 1956 da Nils Foss con lo scopo di migliorare l'efficacia e l'efficienza di metodi di analisi altrimenti lunghi e complessi.

Capitolo 2

I metodi

Per analizzare la composizione di una sostanza tramite analisi chemiometrica, il primo passo è la preparazione del campione da sottoporre ad uno spettrofotometro, fino ad ottenere un numero rappresentativo di ripetizioni. Tale apparecchio è costituito da una sorgente luminosa che irradia perpendicolarmente di luce la sostanza. Quando la luce colpisce una qualsiasi superficie essa ripartisce la propria energia in tre parti: riflettanza, trasmittanza ed assorbanza. Il fascio luminoso cioè, scontrandosi con un ostacolo, può in parte riflettersi, in parte attraversare ed in parte essere assorbito dal campione. La somma di tali misure, dato il principio di conservazione dell'energia, sarà sempre pari ad 1 ma, data la diversità della materia, le tre misure non saranno costanti ma varieranno da sostanza a sostanza.

Dopo aver "eccitato" tramite radiazioni luminose il campione, un rilevatore interno allo spettrometro, è incaricato di analizzare lo spettro luminoso riflesso dal campione. I sensori del rilevatore, posti a distanza di pochi nanometri (nm), analizzeranno specifiche lunghezze d'onda, al fine di cogliere i cambiamenti intervenuti nei legami molecolari. Il rapporto fra la quantità di energia inizialmente emessa e quella registrata al ritorno, viene definita riflettanza, ed è, per costruzione, un numero compreso fra 0 e 1 privo di unità di misura.

La singola rilevazione, quindi, ci fornisce il valore della riflettanza registrata da ogni sensore, cioè alle diverse lunghezze d'onda proprie del rilevatore dello strumento. Rapportando il valore di tale misura ottenuta dal campione e quello registrato su un "bianco", ossia una superficie di

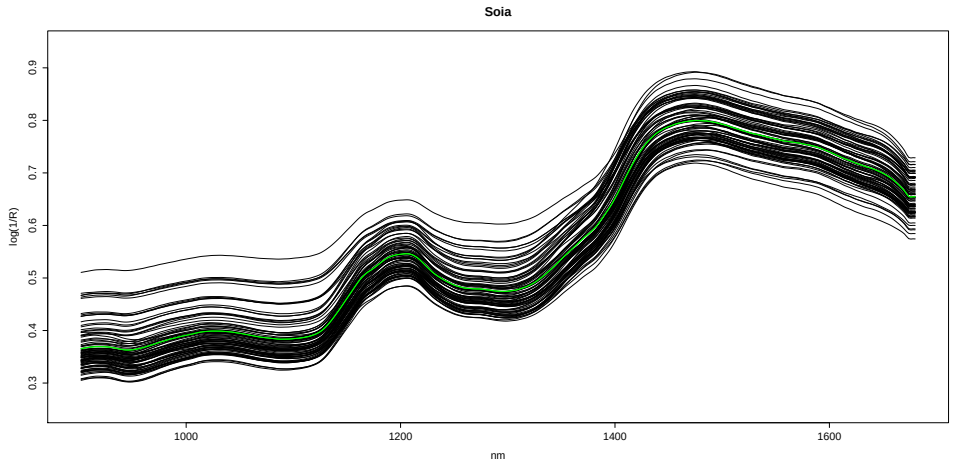


Figura 2.1: Rappresentazione di 81 spettri ottenuti dalla soia. In ordinata abbiamo il valore dell'assorbanza relativamente alle varie lunghezze d'onda, rappresentate in ascissa. In verde è evidenziato lo spettro medio.

cui si conosce per certo il valore della riflettanza (o a cui si attribuisce la totalità del valore), è possibile determinare l'assorbanza delle misure. Poiché vi è una relazione tra l'assorbanza e la composizione del campione, la cui determinazione è il fine ultimo dell'analisi, si è soliti utilizzare quest'ultima grandezza, ricavabile dalla riflettanza come $A = log(1/R)$, in cui A rappresenta l'assorbanza e R la riflettanza.

La quantità appena definita può essere rappresentata visivamente su un grafico cartesiano: in ascissa poniamo, in relazione alle varie lunghezze d'onda indagate, tanti punti quanti sono i sensori del nostro strumento ed in ordinata il valore dell'assorbanza relativo ad ogni sensore. L'unione di questi punti dà origine ad uno spettro che rappresenta, in funzione della lunghezza d'onda, l'intensità di energia elettromagnetica assorbita dal campione. Per fare un esempio, l'acquisizione da un campione, in questo caso di soia, di 81 spettri è rappresentata in figura 2.1 in cui è evidenziato lo spettro medio, così da dar modo al lettore di visualizzare in maniera immediata la singola osservazione.

Data la peculiarità di ogni composto, in termini di struttura atomica e livelli energetici, la forma degli spettri, data l'influenza dell'energia sui legami molecolari, sarà unica per ogni sostanza ed è proprio grazie a questa caratteristica che sarà possibile legare gli assorbimenti luminosi alla composizione chimica.

Dopo aver raccolto l'informazione spettrale, è necessario ottenere l'esatta quantificazione della composizione del campione attraverso analisi

chimiche in laboratorio. Ciò che la chemiometria persegue, infatti, è relazionare, grazie a un modello statistico, i valori di assorbimento alle analisi in laboratorio, cioè le quantità dei composti che caratterizzano la sostanza. Quest'ultime, quindi, si rendono necessarie ogniqualvolta non si ha l'esatta quantificazione chimica relativa alla matrice di dati da analizzare formata da n righe (le osservazioni) ed m colonne (le variabili), tante quante le lunghezze d'onda interessate dallo studio.

Ottenuta anche l'informazione analitica, si può passare alla ricerca e alla successiva stima di un modello di regressione che leghi l'analisi chimica agli assorbimenti luminosi; questa fase viene chiamata calibrazione. Lo scopo finale è quello della previsione, per la quale è necessario disporre dei coefficienti di regressione del modello, necessari per stimare la quantità incognita dei vari costituenti della materia dai relativi spettri incogniti.

Per ricercare un modello, quindi, dobbiamo effettuare una regressione multipla in cui più variabili influenzano la nostra dipendente, ossia la quantità in cui ciascun costituente oggetto d'interesse è presente nel campione. Caratteristica di queste variabili è la non conoscenza a priori né della loro importanza né dei loro possibili effetti sinergici, cioè quegli effetti che si possono cogliere solo considerando congiuntamente due o più variabili.

Non è inoltre possibile separare a priori l'informazione d'interesse, la presenza di rumore sperimentale e la possibile presenza di collinearità nelle variabili, problema che si verifica quando una o più variabili esplicative possono essere predette da altre. Poiché nel nostro caso i sensori sono posti ad una distanza di $2nm$, la presenza è quasi certa e, in generale, tale problema è ricorrente in tutte le applicazioni spettrometriche, dato che quasi sempre il numero delle incognite supera quello delle osservazioni.

Poiché, come detto, l'assorbanza è relazionabile alla composizione, per calcolare la quantità dei costituenti del composto oggetto di analisi, teoricamente sarebbe sufficiente creare un classico modello, da stimare tramite *OLS* (*Ordinary Least Squares*), con cui relazionare gli assorbimenti alle varie lunghezze d'onda e la quantificazione ottenuta tramite l'analisi chimica. Per applicare una regressione ai minimi quadrati, però, è necessario che la matrice delle esplicative sia a rango pieno poiché, per ottenere le

stime dei coefficienti di regressione, sarà necessaria l'inversione della stessa. La mancanza di tale presupposto non darà risultati attendibili e l'uso diretto del metodo di stima *OLS* non sarà consigliato ma sarà necessario intervenire a priori sulle esplicative per risolvere il problema.

Una soluzione consiste nel cercare di ridurre la numerosità delle variabili coinvolte, in modo da avere un numero di parametri da stimare ridotto. In tal modo la loro stima può avvenire in maniera adeguata, anche con poche osservazioni a disposizione. Per ridurre la loro numerosità possiamo eliminare alcune variabili esplicative (tramite ad esempio una procedura *stepwise*), creare nuove esplicative che colgano la maggior parte della variabilità presente nelle variabili iniziali oppure possiamo creare delle variabili latenti che, oltre a cogliere la variabilità presente nelle esplicative, vogliono anche essere massimamente correlate con la risposta.

I metodi chemiometrici hanno perciò la finalità di separare l'informazione utile dal rumore e di risolvere la quasi certa presenza di ridondanza, attraverso l'uso di variabili ausiliarie create allo scopo di sintetizzare l'informazione rilevante nei dati e di ridurre il numero delle esplicative.

Fra i metodi più adottati per risolvere i problemi legati a questa tipologia di dati, troviamo modelli che utilizzano variabili latenti, non direttamente osservabili ma collegate a quelle di partenza, come la *Principal Component Regression* (*PCR*) e la regressione *Partial Least Square* (*PLSR*). Entrambi si basano sostanzialmente su una doppia regressione ai minimi quadrati in cui la prima ha lo scopo di determinare, con diversi criteri, le variabili ausiliarie.

Dato che l'importanza di tali modelli è la stabilità delle previsioni, la convalida incrociata avrà un ruolo chiave sia per la scelta del numero di variabili latenti da includere nel modello, sia per saggiarne le prestazioni prima di poter definire compiuto il passaggio della calibrazione.

2.1 Analisi delle Componenti Principali

L'analisi delle componenti principali (*PCA*, *Principal Component Analysis*) è una tecnica statistica volta a ridurre la quantità di variabili di interesse attraverso l'uso di variabili ausiliarie collegate alle prime ed indipendenti fra loro. Tale tecnica, introdotta da Karl Pearson nel 1901 (Pearson

1901) venne poi sviluppata nella forma attuale da Harold Hotteling nel 1933 (Hotelling 1933).

La costruzione di queste variabili latenti, attraverso una combinazione lineare di quelle di partenza, ha lo scopo di massimizzare l'informazione contenuta nei dati e di ridurre il numero di variabili, gestendo il "trade-off" tra semplificazione ed accuratezza. Le variabili ausiliarie incaricate di evidenziare e sintetizzare l'informazione dalle variabili di partenza, sono appunto le componenti principali.

Date quindi $x_1, \ldots, x_m$ variabili iniziali si ricercano $z_1, \ldots, z_m$ variabili ausiliarie ruotate in un nuovo sistema di riferimento i cui assi rappresentano, in ordine decrescente, le direzioni di massima varianza dei dati iniziali. Per estrarre l'informazione possiamo utilizzare sia la matrice di varianza/covarianza sia la matrice di correlazione che, sebbene possano portare alla soluzione dello stesso problema, restituiscono risultati di differente interpretazione. Nel corso dell'analisi utilizzeremo la matrice di varianza/covarianza e la indicheremo con S.

La prima componente principale, che vogliamo sia il più "informativa" possibile, è quella che massimizza linearmente la variabilità nei dati e sarà quindi della forma

$$z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = a_1 x$$

in cui i coefficienti a_1 sono la soluzione del problema di ottimizzazione vincolata

$$\max_{a_1} a_1' S a_1$$

sotto la condizione

$$a_1' a_1 = 1$$

in modo che la varianza non venga arbitrariamente aumentata e vi sia l'ortonormalità della matrice dei pesi. Il problema così riformulato può quindi essere risolto con l'uso dell'autovettore γ_1 e dell'autovalore λ_1, propri della matrice S. La derivazione delle altre componenti principali si ricava dalla soluzione successiva del massimo vincolato, sotto l'aggiuntivo vincolo di indipendenza con le precedenti variabili latenti trovate. Questo avviene sequenzialmente ottimizzando la varianza residua ed ottenendo così i successivi autovalori $\lambda_i \, (i = 2, \ldots, m)$.

Sfruttando i risultati appena ottenuti ed il "Teorema di Scomposizione Spettrale" (Gregorio e Salce 2010), otteniamo

$$\underset{m\times m}{S} = \Gamma\Lambda\Gamma' = [\gamma_1,\ldots,\gamma_m]\begin{bmatrix}\lambda_1 & 0 & \ldots & 0\\ 0 & \lambda_2 & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \lambda_m\end{bmatrix}\begin{bmatrix}\gamma_1'\\ \vdots\\ \gamma_m'\end{bmatrix} \tag{2.1}$$

in cui Λ è la matrice diagonale degli autovalori e Γ è la matrice ortogonale le cui colonne sono gli autovettori associati agli autovalori della matrice S.

Gli elementi di ogni singolo autovettore γ_i, coefficienti compresi tra -1 e 1, rappresentano le correlazioni fra l'i-esima componente principale e le variabili di partenza. Per questo motivo non è possibile dare una descrizione a parole delle componenti principali prima dell'analisi ma è necessario guardare le correlazioni con le variabili iniziali per capire l'informazione che portano e, quindi, il loro significato.

La matrice diagonale Λ contiene invece, in corrispondenza di ogni autovettore, l'autovalore a quest'ultimo associato, che esprime la quota di varianza che tale direzione cattura. L'auspicata riduzione del numero di variabili avviene proprio in base ai singoli λ_i: maggiore sarà il valore dell'autovalore, maggiore sarà l'informazione portata dal suo associato autovettore. Valori di lambda molto bassi indicheranno componenti principali che hanno catturato prevalentemente rumore e non informazione.

Dal punto di vista geometrico, quindi, tale tecnica si prefigge di individuare un sottospazio su cui sia possibile proiettare le osservazioni attraverso una rotazione delle variabili. Tale rotazione del sistema di riferimento è compiuta dalla matrice Γ (chiamata anche dei *loadings*) che per questo è anche detta "matrice di rotazione". L'applicazione di tale matrice ai dati iniziali, secondo la formula 2.2, consente di ricavare gli *scores* (indicati con T), ossia le coordinate delle variabili nel nuovo spazio:

$$\underset{n\times m}{T} = \underset{n\times m}{X} \times \underset{m\times m}{\Gamma}. \tag{2.2}$$

Per dare un'idea di come tale tecnica operi, possiamo dare uno sguardo alla figura 2.2 che rappresenta, nello spazio iniziale ed in quello delle com-

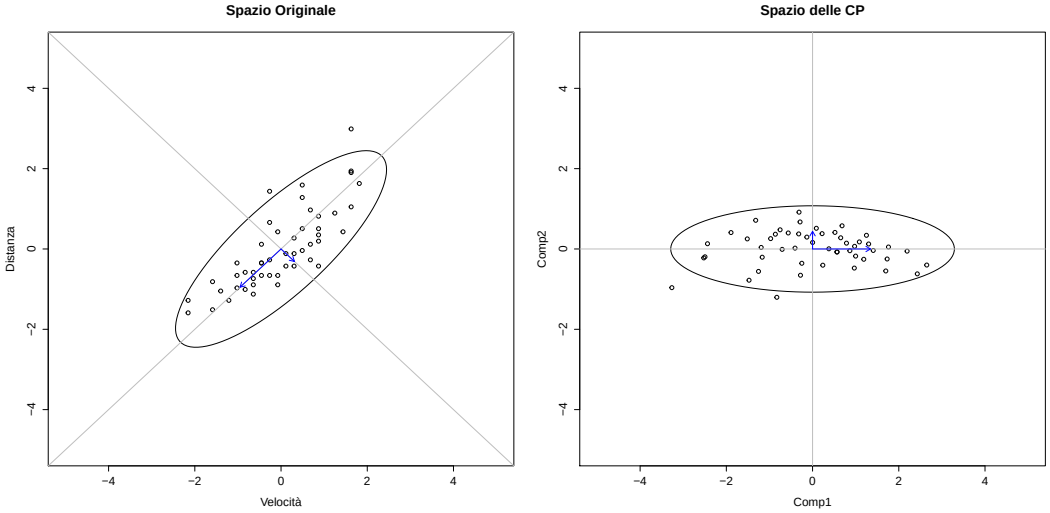


Figura 2.2: Esempio di rotazione mediante componenti principali di un sistema bivariato utilizzando il dataset `cars`.

ponenti principali, due variabili correlate fra loro. Il grafico mostra come l'asse maggiore e l'asse minore dell'ellisse che racchiude i punti sono, rispettivamente, la prima e la seconda componente principale; le lunghezze degli assi dell'ellisse, inoltre, sono proporzionali alla radice quadrata dei rispettivi autovalori.

2.2 I modelli di regressione: *PCR* e *PLS*

Finora abbiamo prevalentemente descritto la costruzione delle variabili latenti senza far riferimento a metodi di regressione e senza spiegare il criterio di riduzione delle variabili.

Se l'analisi delle componenti principali ha uno scopo prevalentemente esplorativo, la riduzione delle variabili può avvenire selezionando un numero $p < m$ di componenti principali tali che esse esprimano una quota sufficiente di variabilità cumulata, solitamente intorno all'80%.

Quando però la *PCA* è un passo intermedio per arrivare ad un modello di regressione in cui le componenti principali vengono prese come esplicative, il criterio appena descritto può essere molto dannoso.

2.2.1 Il pretrattamento dei dati

Prima di passare alla creazione e alla stima del modello vero e proprio, è necessario pretrattare gli spettri. Poiché abbiamo più di una variabile, l'intervallo delle stesse deve essere standardizzato in qualche modo affinché il loro utilizzo non dia informazioni errate. In chemiometria, generalmente, la standardizzazione ha il significato di dare a priori la stessa importanza alle variabili (Forina 2010). Successivamente si potranno dare ad esse pesi diversi a seconda del ruolo che svolgono nel contesto ma, per iniziare l'applicazione di un metodo, si è soliti dar loro pari importanza.

Nel seguito indicheremo con $w_{i,j}$, indipendentemente dal procedimento spiegato, il dato dell'i-esima ripetizione e della j-esima lunghezza d'onda dopo l'applicazione del relativo pretrattamento.

Fra le tecniche più adottate per la standardizzazione troviamo l'*autoscaling* di riga, detto anche *SNV* (*Standard Normal Variate*). Tale procedimento ha forma

$$w_{i,j} = \frac{x_{i,j} - \bar{x}_i}{sd(x_i)}$$

in cui $x_{i,j}$ è il j-esimo valore dell'assorbanza per l'i-esimo spettro, $\bar{x}_i$ e $sd(x_i)$ sono, rispettivamente, la media e la deviazione standard stimata dell'assorbanza dell'i-esimo spettro.

Un secondo metodo di standardizzazione, usato prevalentemente in applicazioni nel vicino infrarosso, è l'*MSC* (*Multiplicative Scatter Correction*). Esso inizia calcolando la quantità media di riferimento per ogni colonna, $\bar{x}_j$, e ponendo in una relazione lineare questi valori con la rispettiva media, secondo l'equazione

$$x_{i,j} = \alpha_i + \beta_i \bar{x}_j.$$

Successivamente, l'*MSC* sfrutta la stima di tali parametri per la standardizzazione:

$$w_{i,j} = \frac{x_{i,j} - \alpha_i}{\beta_i}.$$

Dopo aver applicato alternativamente uno di questi due metodi è possibile proseguire con i pretrattamenti.

Di norma si rimuove l'andamento di fondo dei dati mediante una procedura di *detrend*. Solitamente questa avviene linearmente ed è possibile

eseguirla quando le colonne j della nostra matrice hanno un significato fisico, nel nostro caso la lunghezza d'onda degli spettri. Stimato un classico *OLS* del tipo $x_{i,j} = \beta_{0i} + \beta_{1i} \cdot j$, il dato pretrattato risulterà

$$w_{i,j} = x_{i,j} - \beta_{0i} - \beta_{1i} \cdot j.$$

Successivamente le curve degli spettri, singolarmente, subiscono una derivazione, solitamente di grado non superiore al primo, con lo scopo di far risaltare eventuali picchi che possono spiegare differenze analitiche. Matematicamente, la derivazione viene portata a termine sfruttando la similitudine col rapporto incrementale come:

$$w_{i,j} = x_{i,j} - x_{i,j+k}$$

dove k rappresenta il gap, cioè la distanza in passi fra le lunghezze d'onda interessate dal calcolo.

Infine, le righe della nostra matrice di dati vengono sottoposte ad uno *smoothing* tramite media mobile, a passo k predeterminato, per eliminare le piccole vibrazioni nella rilevazione degli spettri. Il pretrattamento avrà quindi forma

$$w_{i,j} = \frac{1}{k+1} \sum_{c=j}^{j+k} x_{i,c}.$$

2.2.2 Adattamento dei modelli

Per valutare la bontà dei nostri modelli, avremmo bisogno di una misura in grado di quantificare quanto vicine sono le previsioni dei nostri modelli ai veri valori. Nei problemi di regressione, una delle misure usate più di frequente è l'errore quadratico medio o *MSEP* (dall'inglese *Mean Squared Error of Prediction*):

$$\text{MSEP} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

dove $\hat{f}(x_i)$ è la previsione effettuata dal modello della relativa variabile risposta y_i. L'errore che commettiamo è quindi soggetto alla semplifica-

zione che abbiamo dato di $f(x)$, l'ignota vera funzione che ha determinato la risposta, utilizzandone una versione approssimata $\hat{f}(x)$.

Per ricercare un modello soddisfacente possiamo quindi provarne di più o meno complessi al fine di ricercare il valore dell'*MSE* più basso. É importante, però, porre molta attenzione alle osservazioni utilizzate per il calcolo dell'errore. Il valore atteso dell'errore quadratico medio può infatti essere riscritto (Azzalini e Scarpa 2004), per una generica osservazione, come

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 \tag{2.3}$$

che decompone l'errore in due componenti fondamentali, la varianza e la distorsione del modello. In linea generale, infatti, un modello con molti parametri avrà una varianza elevata ma una distorsione ridotta. Contrariamente, una funzione determinata da una semplice regressione lineare, data da una retta parallela e passante per i punti, avrà una bassa varianza ma un'elevata distorsione. L'*MSE*, somma di queste misure, subirà quindi un'iniziale decrescita all'aumentare dei parametri inclusi nel modello, per poi aumentare quando l'aggiunta di nuove variabili non si giustifica con una sufficiente diminuzione della distorsione. Le due misure ricavate dalla 2.3 non possono quindi essere minimizzate contemporaneamente e questo loro rapporto viene spesso indicato con *Bias-Variance Tradeoff*.

Minimizzare l'*MSE* sul campione utilizzato per la stima, data l'influenza della distorsione, può portare informazioni errate sulla qualità del modello e far cadere nel problema del sovradattamento o *overfitting*. Tale problema si presenta quando, all'interno di un modello di regressione, vengono inserite un gran numero di esplicative o, meglio, un numero eccessivo. Se infatti da un lato l'aggiunta di variabili migliora l'adattamento del modello ai dati, dall'altro un numero eccessivo può portare lo stesso a focalizzarsi maggiormente su peculiarità legate alle fluttuazioni casuali dei dati utilizzati per la stima che su caratteristiche chiave del fenomeno. Il dar peso ad uno specifico gruppo di dati, esaltandone le singolarità che non si ripresenteranno in un nuovo campione, porterà quindi all'amplificazione del rumore di fondo e ad un peggioramento del potere predittivo del modello.

Poiché l'interesse dei modelli chemiometrici è proprio nella stabilità delle previsioni e non nell'adattamento degli stessi ai dati, l'*MSE*, per tro-

vare un punto sufficientemente buono in cui sia la varianza sia la distorsione sono contenute, dovrà essere calcolato su osservazioni estranee a quelle utilizzate per la sua stima. A tal fine, una delle tecniche più utilizzate, è la convalida incrociata, o *cross-validation* (*CV*) (Azzalini e Scarpa 2004).

Le tecniche che inizialmente furono utilizzate in chemiometria prevedevano la divisione del campione iniziale in due blocchi, uno per la stima ed uno per la verifica (*single evaluation set*). Sebbene sia il metodo più rapido, l'utilizzo di una sola partizione dei dati da utilizzare come verifica dava comunque risultati poco affidabili poiché il calcolo della distorsione era affidato ad un solo gruppo.

La procedura *k-fold cross validation*, invece, prevede la divisione del campione iniziale in k gruppi di egual dimensione. I modelli con diverso numero di regressori verranno quindi stimati utilizzando $k-1$ gruppi (*training set*), per poi andare a determinarne le previsioni per la restante parte dei dati. Il calcolo della bontà di adattamento verrà quindi svolta confrontando le previsioni ottenute dai vari modelli con il loro relativo vero valore presente nel k-esimo gruppo, che verrà perciò usato come insieme di verifica (*validation set*). La procedura appena descritta verrà ripetuta tante volte quanti sono i k gruppi in cui si è deciso di partizionare le osservazioni e si prenderà poi un valore medio come indicatore della qualità di adattamento. Così facendo, infatti, ogni gruppo svolgerà sia il ruolo di calibrazione sia quello di verifica del modello.

Ulteriore estensione della procedure appena descritta è la *double-cross validation* (*DCV*), introdotta da Svante Wold, che rappresenta «il metodo migliore per valutare il numero delle componenti significative» (Forina 2010). Essa inizia applicando la convalida incrociata a k gruppi su tutte le osservazioni. I $k-1$ gruppi incaricati di stimare il modello vengono ulteriormente divisi applicandone un secondo ciclo. Quest'ultima ripetizione è quella incaricata di ricercare un modello. Dopo l'ottimizzazione svolta dal ciclo interno, come descritto nel paragrafo precedente, il k-esimo gruppo, inizialmente escluso dal processo, sarà utilizzato per ottenere una stima realistica dell'errore commesso.

Nel corso dell'analisi, infine, ricorreremo all'uso di altri due indici collegati fra loro e con l'*MSE*: il *Predicted Residual Error Sum of Squares* (*PRESS*), definito come la sommatoria dei residui, e l'R^2. Anche questi

due, per i motivi sopra discussi, non saranno calcolati sui dati utilizzati durante la stima del modello bensì in convalida incrociata.

2.2.3 Outlier

Per migliorare ulteriormente la qualità della calibrazione, verranno ricercati fra gli spettri dei possibili outlier. L'analisi, che sarà eseguita sugli *scores*, verrà condotta principalmente tramite due metodologie: la distanza di Mahalanobis ed un test derivato dal test *t* di Student, chiamato outlier T.

La prima può essere utilizzata in due momenti dell'analisi. La sua applicazione può risultare utile prima della creazione di un modello, per evitare che quest'ultimo risenta di acquisizioni di dati anomali, o nella successiva fase di previsione, per andare ad eliminare spettri incogniti molto diversi da quelli utilizzati per la stima. Geometricamente, infatti, la formula di Mahalanobis quantifica la distanza che vi è fra due oggetti. Nel nostro caso, il punto dal quale misureremo la distanza dei vari punti, sarà la media degli *scores*. Il calcolo della distanza di Mahalanobis (al quadrato) è quindi

$$D^2 = (x_p - \mu_p)' S_p^{-1} (x_p - \mu_p)$$

in cui μ_p rappresenta la media per componente principale, x_p e S_p rappresentano, rispettivamente, gli *scores* e la matrice di covarianza delle prime p componenti principali. Dopo aver calcolato D^2 per tutte le osservazioni, essa verrà divisa per la deviazione standard ($sd(D^2)$), in modo da ottenere una misura fra distanze standardizzata. I campioni che avranno un valore $D^2/sd(D^2)$ maggiore di 10 saranno considerati outlier.

La seconda misura per la ricerca di osservazioni anomale sfrutta i residui di un modello di regressione. Iniziamo calcolando il modello con tutte le n osservazioni per poi ottenerne le previsioni, in modo da condurre la ricerca su una misura che lega sia la variabile dipendente che la risposta.

Ad ogni passo si calcola la varianza campionaria corretta detta *SEC*, *Standard Error of Calibration*[1], come

$$\text{SEC} = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n - 1 - k}}$$

[1] *SensoLogic Calibration Workshop* 2010.

in cui y_i ed $\hat{y}_i$ indicano l'i-esimo dato analitico e predetto, k rappresenta il numero di variabili nel modello ed n il numero delle osservazioni. La determinazione dell'outlier T può quindi avvenire applicando

$$T = \frac{\hat{y}_i - y_i}{\mathrm{SEC}\sqrt{1-H}}$$

che segue una distribuzione t di Student con $n-1$ gradi di libertà. Il valore H dipende sia dal numero dei parametri del modello sia dal numero di osservazioni all'interno del *training set* (n_t):

$$H = \frac{k+1}{n_t}.$$

Dopo una prima iterazione della ricerca di outlier T, vengono ritenute anomale quelle osservazioni che hanno un valore di T, in modulo, maggiore di 2.5. Supponiamo di aver trovato n_o campioni outlier. Il processo appena descritto ricomincerà quindi dalla ricerca del modello, utilizzando questa volta per la sua stima $n - n_o$ campioni. La ricerca di outlier T avviene infatti, come svolto da WinISI, in due interazioni.

2.3 *Principal Component Regression (PCR)*

Un modello di regressione basato su componenti principali (*Principal Component Regression*) ricerca una relazione lineare fra la variabile risposta e le variabili latenti descritte nell'analisi delle componenti principali.

La *PCA* dunque ci permette di passare da variabili osservabili fortemente collineari che non ci consentirebbero di utilizzare il classico *OLS*, a variabili latenti massimamente informative ed indipendenti fra loro.

Iniziamo quindi col ricavare le variabili Z da utilizzare nella regressione:

$$\underset{(n\times m)}{Z} = \underset{(n\times m)}{X} \times \underset{(m\times m)}{\Gamma'} \tag{2.4}$$

in cui X è la matrice iniziale dei dati e Γ la matrice di rotazione definita nell'analisi delle componenti principali. Abbiamo così determinato tante componenti principali quante sono le variabili di partenza. Per decidere il numero ottimo di componenti da includere nel modello ricorriamo alla convalida incrociata.

Per ogni possibile numero di regressori da includere nel modello viene verificata la capacità predittiva in base alle statistiche precedentemente presentate che quantificano l'errore che esso commette. Dalla ripetizione di tali misurazioni sui vari gruppi, si vedrà con quale numero p di regressori si sono ottenuti i migliori risultati, portando a considerare tale valore come ottimo.

Trovato quindi il numero ideale di componenti principali da utilizzare, possiamo andare a stimare un classico modello *OLS* utilizzando come esplicative le variabili latenti sopra calcolate. Abbiamo quindi:

$$Y = Z_p \beta_p^* + \varepsilon$$

dove con p a pedice si intendono vettori o matrici riferite alle prime p componenti principali. Il vettore dei coefficienti β_p^* viene calcolato utilizzanto le variabili ottenute da 2.4, secondo l'usuale formulazione dei minimi quadrati, come

$$\hat{\beta}_p^* = (Z_p' Z_p)^{-1} Z_p' Y.$$

I coefficienti appena ricavati, poiché calcolati su variabili ausiliarie, non sono direttamente applicabili alle variabili di partenza, le lunghezze d'onda. E' necessario cambiare nuovamente il sistema di riferimento utilizzando la matrice di rotazione che precedentemente ci ha fornito le esplicative. I nostri coefficienti, trasformati per adattarsi alle lunghezze d'onda, risulteranno quindi da:

$$\hat{\beta} = \Gamma \hat{\beta}^*.$$

Con i β così calcolati è possibile andare a quantificare i costituenti oggetto d'indagine da un coerente spettro incognito, cioè tale per cui la calibrazione sia rappresentativa. Sarà ad ogni modo necessario riapplicare agli spettri incogniti gli stessi pretrattamenti che sono stati utilizzati per giungere al calcolo dei coefficienti di regressione, prima di poter procedere alla quantificazione vera e propria dai vari assorbimenti spettrali.

2.4 *Partial Least Square Regression* (*PLSR*)

L'approcio basato su un modello di regressione *Partial Least Square*, una delle tecniche più utilizzate in chimica analitica (Forina 2010), presenta alcuni punti di contatto con la *PCR* ma anche notevoli vantaggi nella gestione di variabili con rumore, solitamente collineari e che spesso presentano dati mancanti. Il metodo fu inizialmente proposto da Herman Wold nel 1975 e successivamente venne sviluppato con l'aiuto del figlio Svante; quest'ultimo, date le caratteristiche del procedimento, proporrà successivamente il termine più calzante di *Projection to Latent Structure* (Wold, Sjöström a e Eriksson 2001).

La *PLSR*, similmente alla *PCR*, è una tecnica per la creazione di un nuovo set ridotto di variabili ausiliarie su cui effettuare una regressione lineare. Mentre la regressione in componenti principali, però, utilizza solo le esplicative X per la costruzione delle variabili latenti, la regressione ai minimi quadrati parziali sfrutta anche le informazioni contenute nella risposta Y. La *PCR* quindi, sfruttando solo una parte dell'informazione disponibile nel dataset, presenta un lato negativo: non è detto, infatti, che le direzioni che meglio rappresentano i predittori siano anche quelle che meglio spiegano la risposta. La *PLSR* non solo determina tramite *PCA* le direzioni di massima varianza fra le esplicative, ma pesa anche tali direzioni a seconda del loro rapporto con Y.

Il processo inizia quindi pesando singolarmente l'influenza di ogni variabile esplicativa X_j sulla risposta Y calcolando il coefficiente

$$\phi_{1j} = X_j' Y$$

che è proporzionale alla correlazione fra X_j e Y (Hastie, Tibshirani e Friedman 2008) e avrà lo scopo di ruotare parzialmente gli *scores* della matrice X per aumentarne la correlazione con la dipendente. La determinazione della prima variabile latente definita secondo la regressione ai minimi quadrati parziali sarà quindi ricavata come

$$Z_1 = \sum_{j=1}^{m} \phi_{1j} X_j$$

che rappresenta la prima direttrice nel nuovo sistema di riferimento ed è

maggiormente influenzata da quelle variabili X_j che risultano altamente collegate alla risposta.

La derivazione delle successive variabili latenti, similmente a quanto visto in *PCA*, si ricavano ottimizzando l'informazione residua. Per determinare la seconda, ad esempio, si regredisce ogni variabile iniziale su Z_1 e si ottengono i residui. Questi possono essere visti come l'insieme di informazioni non spiegate dalla prima variabile latente e verranno quindi utilizzati per il calcolo di Z_2, nello stesso modo in cui è stata calcolata la prima variabile ausiliaria.

La *PLSR* quindi non massimizza l'informazione delle variabili esplicative tanto bene quanto fa la *PCA*, ma svolge un ruolo migliore nello spiegare la risposta. Essa può quindi essere vista come una modellazione di variabili latenti comuni al principio generatore che lega X ed Y.

Dopo aver ricavato tutte le Z_i ($i = 1, \ldots, p$), similmente alla *PCR*, la *PLSR* utilizza un classico *OLS* per la creazione di un modello ed il numero p di variabili latenti che entrano nella regressione, viene nuovamente scelto tramite convalida incrociata.

Capitolo 3

Applicazione

In questo capitolo presenteremo i risultati delle analisi applicate ad un dataset di esempio. Per ogni costituente d'interesse stimeremo i modelli descritti nel capitolo precedente ed andremo a fare alcune considerazioni sui risultati, per determinare quale metodologia, per ciascuna variabile dipendente, sia meglio seguire. Il primo passo, ad ogni modo, è il pretrattamento dei dati, di cui in seguito se ne darà anche un'interpretazione visiva.

Per ogni funzione che verrà impiegata nell'analisi, sarà riportato il riferimento all'appendice A, in cui verrà fornito il codice R necessario alla loro implementazione.

3.1 I dati

Il dataset che andiamo ad analizzare contiene 81 spettri rilevati su campioni di soia con lo spettrometro poliSPEC^{NIR}, di fabbricazione IOR3. Le rilevazioni sono eseguite con un range spettrale da 900 a $1680nm$, con risoluzione ottica a $2nm$. Il dataset risultante è perciò una matrice composta da $n = 81$ righe ed $m = 390$ variabili, una per ogni sensore. I costituenti oggetto di analisi, cioè le sostanze di cui si vuole quantificare la presenza all'interno della soia, sono tre: le proteine, l'umidità ed i grassi.

Poiché i campioni sottoposti ad analisi non hanno una composizione omogenea, è necessario effettuare più rilevazioni per ottenere un dato significativo. La semplice acquisizione, infatti, che richiede pochi centesimi di secondo, sarà specifica per l'area ed il campione di soia che è stata posta

sotto al rilevatore dello spettrofotometro. Per ottenere una calibrazione rappresentativa, quindi, è necessario rilevare più spettri per ogni campione. Nel nostro caso la singola osservazione è la media di 10 rilevazioni. Dopo aver ottenuto la rilevazione spettrale media per tutti i campioni, è possibile utilizzare i dati così acquisiti per eseguire una calibrazione che si adatti alla "popolazione".

Il valore di riferimento di tali quantità viene ricavato da un'analisi chimica in laboratorio ed è inizialmente espresso come moli su litro ($mol\,L^{-1}$). La standardizzazione di tali concentrazioni, che è il valore che sarà usato nell'analisi, ci consente di trattare i costituenti come quantità percentuali. Ad ogni modo i tre costituenti oggetto di analisi non sono esaustivi di tutta la composizione della soia tant'è che, mediamente, le tre quantità insieme rappresentano il 73% di tutte le sostanze che compongono i nostri campioni.

3.2 Analisi esplorative

Poiché le variabili che dobbiamo prevedere sono di tipo quantitativo, iniziamo con un'analisi esplorativa, per avere un'idea delle quantità oggetto di studio. Di seguito, in figura 3.1 ed in tabella 3.1, vengono mostrate le caratteristiche delle tre variabili: `Proteine`, `Umidita` e `Grassi`.

Gli istogrammi ed i boxplot mostrano le differenze che vi sono fra queste quantità. Oltre alla diversa concentrazione delle variabili, notiamo che le loro distribuzioni presentano forme diverse. Le proteine ed i grassi hanno una distribuzione abbastanza simmetrica e, come messo in evidenza dai boxplot, alcune osservazioni sono identificate come outlier. Per quanto riguarda invece l'umidità, le misurazioni si presentano con una distribuzione più omogenea. Il boxplot, infatti, ha una distanza interquartile molto più ampia degli altri due costituenti e non evidenzia campioni anomali.

Iniziamo ora a trattare la matrice delle esplicative, cioè la discretizzazione per singola lunghezza d'onda delle curve spettrali. Come preannunciato, i dati dovranno essere pretrattati per migliorare la qualità dei modelli che verranno stimati. Come primo passo quindi applichiamo le funzioni per il calcolo dell'*SNV* (funzione A.1) ed eseguiamo la derivata ed

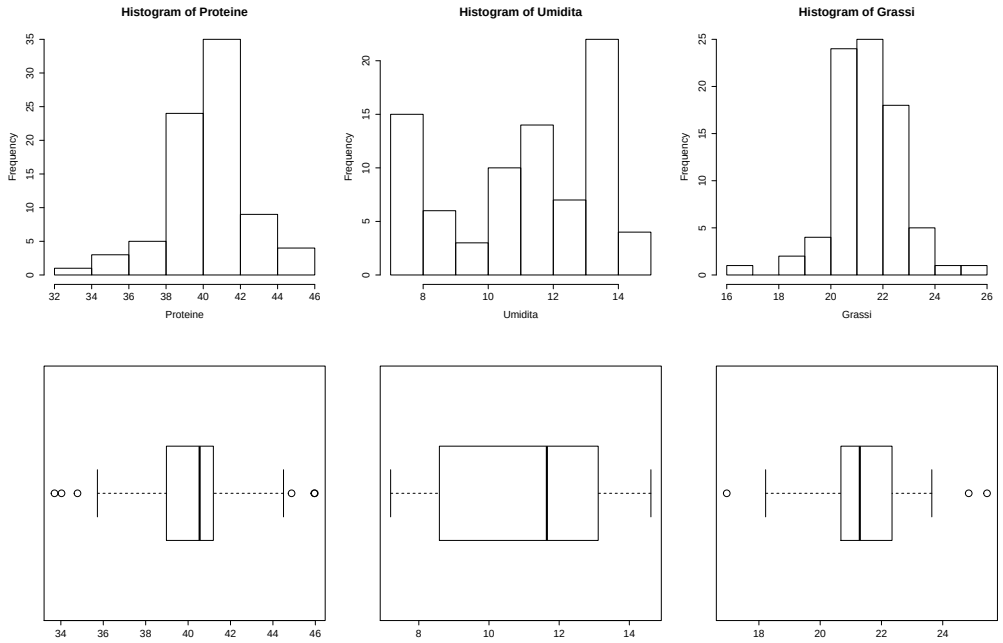


Figura 3.1: Istogrammi e boxplot per le tre variabili dipendenti.

	Proteine	Umidità	Grassi
Min.	33.7	7.196	16.95
1st Qu.	38.98	8.586	20.68
Median	40.54	11.65	21.29
Mean	40.25	11.17	21.42
3rd Qu.	41.19	13.11	22.35
Max.	45.97	14.61	25.45

Tabella 3.1: Sommario delle variabili dipendenti: `Proteine`, `Umidita` e `Grassi`.

il lisciamento degli spettri (funzioni A.2 e A.3). Applicando l'*autoscaling* come mostrato in appendice, viene eseguito l'*SNV* come fatto da WinISI. Il trattamento infatti, che normalmente prevederebbe la standardizzazione sia in media che in varianza, viene eseguita dal *software* di riferimento con la sola divisione per la deviazione standard. Questo comunque non comporterà errori futuri poiché, prima di eseguire uno dei metodi di regressione presentati nel capitolo precedente, i dati verranno scalati anche in media. Le funzioni del detrend (lineare) e del lisciamento, per quanto riguarda i passi su cui effettuare il relativo calcolo, sono così impostate per l'emulazione di WinISI e, a meno di trascurabili errori numerici,

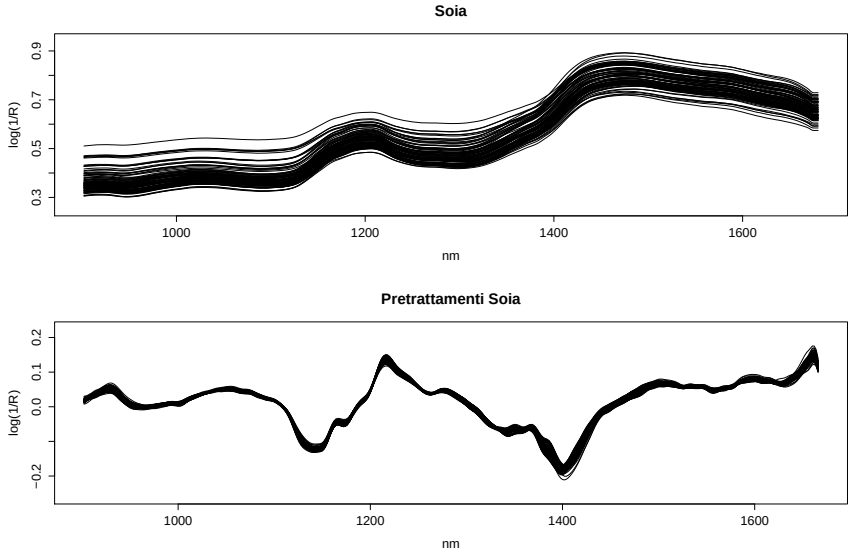


Figura 3.2: Curve spettrali prima e dopo l'applicazione dei pretrattamenti.

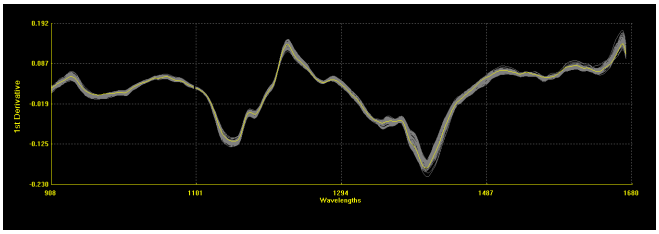


Figura 3.3: Grafico degli spettri pretrattati con WinISI.

producono gli stessi risultati.

In figura 3.2 vengono mostrati gli spettri prima e dopo l'esecuzione delle funzioni sopra descritte. Si può vedere come l'effetto sia stato quello di avvicinare fra loro gli spettri e mettere in evidenza picchi di assorbimento che prima risultavano essere solo deboli cambiamenti di pendenza delle curve spettrali. In figura 3.3, viene riportato infine l'insieme delle curve pretrattate con WinISI per mostrarne la similitudine.

L'applicazione, sia del lisciamento tramite media mobile sia della derivata, fa perdere le ultime colonne della matrice, per impossibilità di calcolo. Da qui in poi quindi la matrice di dati pretrattata avrà $m = 383$ colonne.

Prima di iniziare la ricerca di un modello, andiamo ad indagare la possibile presenza di osservazioni anomale tramite la distanza di Mahalanobis, come presentato nel paragrafo 2.2.3. Il calcolo di tale quantità prevede l'inversione della matrice di covarianza degli *scores*, dato che la ricerca av-

viene sulle variabili ausiliarie. Poiché l'inversione dell'intera matrice non è possibile, data la presenza di molti autovalori nulli, è necessario troncarla. Per fare ciò sfruttiamo la funzione A.7 che, al variare del numero di componenti principali, determina la quantità di varianza spiegata. Decidendo a priori una soglia di informazione di cui si vuole tener conto, è possibile andare a ricercare, con la funzione A.8, il numero di componenti necessarie. Utilizziamo quindi il numero trovato per il calcolo della matrice di covarianza e la successiva determinazione della distanza di Mahalanobis, come in funzione A.9. Poiché nessun campione supera la soglia di rifiuto, possiamo iniziare la ricerca dei modelli con tutte le osservazioni a nostra disposizione.

3.3 Calibrazione

Per ricercare un modello facciamo uso del comando `mvr_dcv` (Filzmoser, Liebmann e Varmuza 2009) all'interno del pacchetto `chemometrics`. Tale funzione conduce una ricerca del numero ottimo di regressori da inserire in un modello, sia esso *PCR* o *PLSR*, in doppia convalida incrociata, come presentato al termine del paragrafo 2.2.2. La selezione del modello di regressione avviene tramite il parametro `method`.

Per ottenere tale analisi per una regressione in componenti principali, impostiamo l'opzione su `"svdpc"` (*Single Value Decomposition Principal Component*) che, per una matrice di dati centrata, fornisce gli stessi risultati dell'analisi delle componenti principali.

Volendo invece ottenere tale informazione per un modello di regressione che utilizza variabili latenti, definite dalla procedura PLS, l'impostazione di `method` può avvenire scegliendo fra tre differenti procedure: `kernelpls`, `widekernelpls`, `simpls` e `oscorespls`.

L'algoritmo `kernelpls`, proposto da Höskuldsson (Höskuldsson 1988), è basato sul calcolo delle cosiddette matrici kernel $X'YY'X$ e $Y'XX'Y$. Esso calcola iterativamente i *weights*, gli *scores* ed i *loading*, rimuovendo ad ogni step l'informazione estratta dalla matrice iniziale. L'algoritmo risulta molto veloce se il numero di variabili X ed Y non è troppo elevato e, per come sono determinate le matrici proprie dell'algoritmo, il numero di oggetti non ha impatto sulla loro dimensione.

La variante `widekernelpls` risulta maggiormente efficiente quando il numero delle variabili è molto superiore a quello delle osservazioni, con un rapporto di circa 1000 : 1. Data la dimensione del nostro dataset, questo non rientra nell'eventualità.

La procedura `simpls` calcola le variabili latenti dell'algoritmo *PLS* come combinazione lineare dei dati iniziali. I fattori così determinati saranno quelli che massimizzano la covarianza fra X ed Y, facendo sempre riferimento alle restrizioni di ortogonalità e normalità.

Infine, la strategia `oscorespls` prevede la determinazione delle variabili latenti estraendo le variazioni dalla matrice delle esplicative ortogonalmente a quella della dipendente. Utilizza poi i dati così modificati per massimizzare, non solo la correlazione fra le variabili, ma anche la covarianza.

Nel seguito, date le migliori prestazioni computazionali per studiare la variazione delle nostre statistiche in un modello *PLSR*, utilizzeremo la strategia `kernelpls`. Si noti comunque che, l'applicazione delle diverse metodologie al nostro caso, ha restituito lo stesso numero ottimo di regressori.

I restanti parametri `repl`, `validation`, `segments0` e `segment` controllano, rispettivamente, il numero di replicazioni, il metodo da utilizzare per la validazione (a scelta fra *leave one out*[1] e *k-fold cross-validation*, quella che utilizzeremo con $k = 6$) ed il numero di segmenti per il primo ed il secondo ciclo propri della doppia convalida incrociata.

L'algoritmo divide inizialmente le osservazioni in `segments0` blocchi, supponiamo k_0. A questo punto, i $k_0 - 1$ gruppi incaricati di stimare la regressione vengono divisi in altri `segments` partizioni, poniamo k. Il modello viene stimato, quindi, con $k - 1$ blocchi, utilizzandoli come insieme di "allenamento" e ne viene testata la performance sul restante k-esimo gruppo. Su questo viene registrato il numero ottimo di componenti con il quale si sono ottenuti i migliori risultati. La verifica finale della capacità predittiva del modello avviene infine, sul gruppo k_0, escluso dall'iniziale processo di calibrazione. L'intero processo viene quindi ripetuto `repl` = 100 volte.

[1] La strategia *leave one out* è una tipologia di convalida incrociata che prevede l'esclusione, a rotazione, di ogni singola osservazione.

Infine, il criterio da seguire per la selezione del modello, può avvenire in base a diverse strategie che vanno impostate con l'opzione `selstrat`.

Il metodo `hastie` prevede l'impiego della cosiddetta *standard-error-rule* (Hastie, Tibshirani e Friedman 2008). Questa regola prevede l'uso del modello più parsimonioso fra quelli che registrano un valore dell'errore prossimo al *MSEP* minimo. Essa tiene contro che la stima del valore è effettuata con errore e propende, quindi, per un approccio conservativo. Si sceglie perciò il modello più parsimonioso il cui *MSEP* è compreso in un intervallo centrato sul miglior modello. Operativamente, essa aggiunge al valore dell'errore quadratico medio, pesato dal coefficiente `sdfact` preimpostato a 2, l'errore standard del *MSEP*, dove in entrambi i casi la statistica di riferimento è presa nel suo minimo.

La strategia `diffnext` confronta la differenza fra il *MSEP* medio, aggiustato come nel criterio `hastie`, ed il valore degli *MSEP* con un numero minore di componenti. Se la differenza fra queste due misure risulta negativa, il modello è fra i possibili ottimi. Fra questi candidati, sarà ritenuto migliore il modello più complesso che presenta un numero di componenti minore rispetto a quello di riferimento.

Infine, la procedura `relchange` prevede l'uso combinato del metodo `hastie`. Prendendo in considerazione solo i modelli con complessità minore o uguale a quello con valore più basso, si usa il valore massimo e minimo del *MSEP* per relazionare i vari modelli e, se la differenza è più piccola di 0.005, vengono scartati quelli con un numero di regressori maggiori. Fra i restanti si seleziona poi il modello secondo il criterio `hastie`.

Nel seguito si utilizzerà il metodo `hastie` poiché permette di selezionare con semplicità il modello più stabile. Si noti comunque che nell'applicazione al dataset, nonostante una frequenza diversa del numero ottimo ed indipendentemente dal modello di regressione (*PCR* o *PLSR*), i tre metodi hanno portato alla determinazione dello stesso numero ottimo di regressori.

3.3.1 Modelli di regressione *PCR*

Eseguiamo la funzione `mvr_dcv` (A.11) che andremo a sfruttare per l'estrazione del numero ottimo di componenti. Per ogni gruppo in cui si è deciso di partizionare i dati, riportato in riga nella tabella 3.2, la funzio-

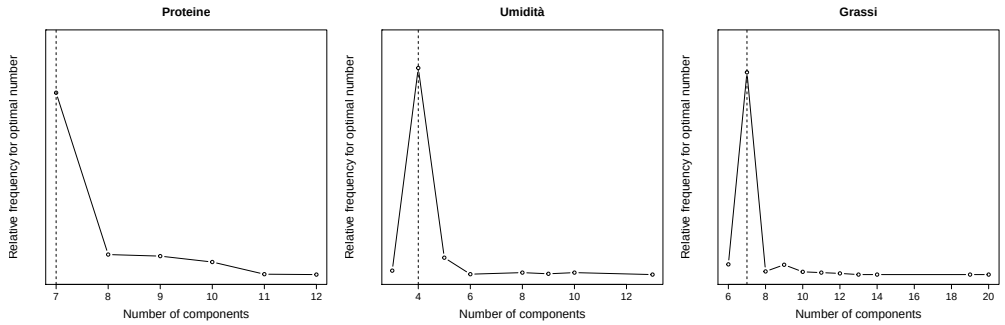


Figura 3.4: Distribuzione di probabilità stimata per il miglior numero di componenti principali per i tre costituenti.

ne registra il numero di regressori con il quale il modello ha ottenuto i risultati migliori, in base alla statistica di riferimento scelta. Dopo aver eseguito il controllo dei possibili modelli su ogni partizione, il processo viene ripetuto 100 volte ricampionando i dati. La tabella riportata mostra quindi solo le prime 10 ripetizioni.

	1 repl	2 repl	3 repl	4 repl	5 repl	6 repl	7 repl	8 repl	9 repl	10 repl
1 segm	7	7	7	10	7	7	7	7	10	8
2 segm	8	7	10	7	8	8	10	7	7	7
3 segm	7	7	7	7	7	9	7	7	7	7
4 segm	7	7	7	10	8	7	7	7	7	7
5 segm	7	7	10	7	6	7	7	7	8	7
6 segm	7	7	7	7	10	7	7	8	7	7

Tabella 3.2: Esempio della matrice di output della doppia convalida incrociata.

Attraverso la funzione `plotcompmvr` è possibile visualizzare la probabilità stimata per il numero ottimo di componenti, cioè la rappresentazione visiva dell'informazione contenuta nella matrice della tabella 3.2. Per i tre costituenti oggetto d'indagine, un esempio di tale grafico è mostrato in figura 3.4. In ascissa vi sono il numero di regressori che si sono rivelati ottimi almeno una volta in una partizione dei dati ed in ordinata la frequenza con cui si è ottenuto tale numero.

	3	4	5	6	7	8	9	10	11
Proteine	0.00	0.00	0.00	0.00	0.80	0.07	0.07	0.04	0.01
Umidità	0.02	0.86	0.09	0.00	0.00	0.01	0.01	0.00	0.00
Grassi	0.00	0.00	0.00	0.06	0.86	0.01	0.02	0.01	0.01

Tabella 3.3: Probabilità stimata per numero di regressori e costituente per i modelli PCR.

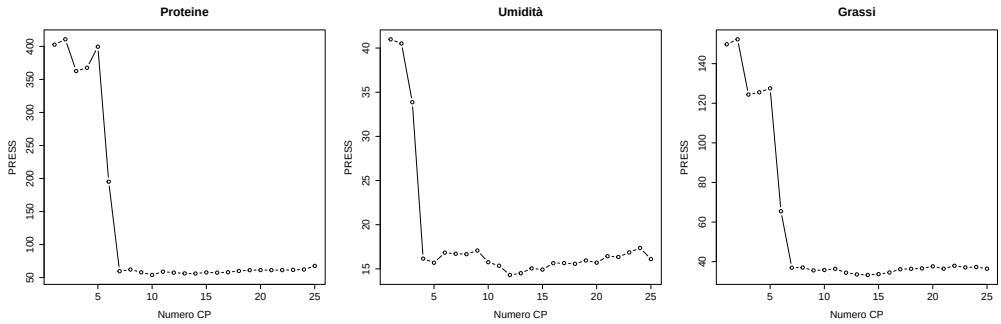


Figura 3.5: Andamento della statistica PRESS al variare del numero di componenti principali.

Su volontà di IOR3, a differenza di quanto fatto da WinISI, la ricerca del numero ottimo di componenti principali avviene per singolo costituente. Il programma di riferimento, infatti, calcola (sia per modelli *PCR* che *PLSR*) un numero di regressori comune all'intero dataset; nel caso delle componenti principali il numero ottimo da questo trovato è 7. Come mostrato dai grafici in figura 3.4 e dalla tabella 3.3, il numero ottimo di regressori scelti con R differisce fra i costituenti ed è 7 per `Proteine` e `Grassi` e 4 per `Umidita`.

Andiamo quindi ad utilizzare la funzione `pcr` (A.12), per ogni costituente, per stimare la regressione del paragrafo 2.3. Da tale oggetto andremo poi ad estrarre alcune misure di interesse per capire quanto bene funziona il modello trovato. Il parametro `ncomp` di tale funzione può essere utilizzato per diminuire il carico computazionale, impostando il massimo numero di componenti principali uguale all'ottimo trovato precedentemente. Per mostrare l'andamento delle statistiche di riferimento, però, impostiamo un valore massimo di ricerca più alto.

I grafici in figura 3.5 mostrano il variare della misura *PRESS* contro il numero di componenti principali. Come si può notare, il numero di variabili latenti scelte nel passaggio precedente, non è il minimo assoluto ma il primo minimo locale, per il quale si ha un sensibile abbassamento della statistica di riferimento. Se andassimo a scegliere il minimo assoluto, o il massimo se stessimo osservando l'andamento dell'R^2, otterremo un numero di parametri spesso simile ed un lieve miglioramento. Per quanto riguarda l'umidità, ad esempio, la scelta del minimo assoluto porterebbe all'utilizzo di 12 componenti principali. A fronte di un miglioramento dell'R^2 di meno del 5‰, decidiamo di non usare un numero triplo di re-

gressori. La ricerca chemiometrica, infatti, deve avere come obbiettivo la creazione di modelli semplici, cioè con un numero contenuto di parametri, per non rischiare di esaltare il rumore nei dati includendo informazioni scarsamente necessarie nei modelli di regressione. Il principio di massima parsimonia è quindi sempre da tenere a mente.

Andiamo ora ad estrarre i residui per il modello con il numero ottimo di componenti principali determinato in doppia convalida incrociata, al fine di condurre una ricerca di outlier mediante test *t*, come discusso nel paragrafo 2.2.3.

Con l'uso della sequenza A.14 replichiamo il metodo di calcolo di WinISI per l'individuazione di outlier T. Eseguiamo le linee di codice separatamente per i tre costituenti del dataset ed otteniamo i campioni ritenuti essere anomali. Il risultato di tale procedimento è mostrato in tabella 3.4.

	R	WinISI
Proteine	4, 41	4, 41, 65
Umidità	43, 70, 74	43, 70, 74
Grassi	63, 74	41, 63, 74

Tabella 3.4: Outlier rilevati da R e WinISI per i modelli PCR.

Dal confronto fra i due *software* vediamo come la selezione dei campioni da parte di R sia buona, replicando perfettamente WinISI nella determinazione degli outlier per il secondo costituente e rilevando per le proteine ed i grassi due campioni su tre che sono usciti dalla soglia. Poiché i calcoli per arrivare alla determinazione del valore dell'outlier T si basano sui residui del modello, non c'è da stupirsi che i risultati non siano perfettamente identici. Bisogna ricordare, infatti, che la stima del modello è soggetta ad errori numerici.

Reimpostiamo quindi la nostra matrice di dati escludendo i campioni individuati per rispettivo componente e stimiamo nuovamente il modello. Una seconda iterazione del processo non porta a determinare nuovi outlier e possiamo quindi ritenere di essere giunti al modello finale.

In figura 3.6 possiamo vedere i grafici di dispersione per relativo costituente delle previsioni effettuate dai tre modelli. Per giudicare buone le previsioni, i punti dovrebbero distribuirsi lungo la bisettrice del primo quadrante. A giudicare dai grafici, le previsioni sono abbastanza buo-

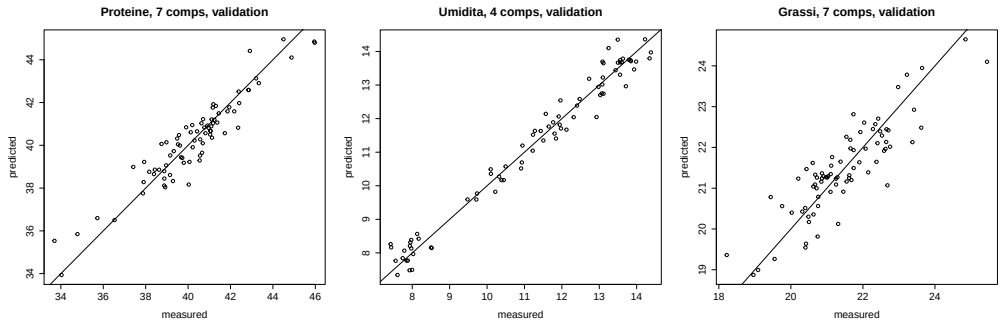


Figura 3.6: Confronto valori chimici e valori stimati con i modelli PCR.

ne, particolarmente accurate sembrano le stime di `Umidità`. In tabella 3.5 riportiamo un sommario delle variabili dipendenti, confrontandone la quantificazione tramite analisi chimica e tramite stima del modello.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Proteine	33.70	39.06	40.54	40.28	41.22	45.97
Stima Proteine	33.94	39.12	40.35	40.25	41.20	44.96
Umidità	7.42	9.54	11.71	11.21	13.11	14.38
Stima Umidità	7.34	9.59	11.65	11.22	13.21	14.36
Grassi	18.22	20.70	21.31	21.48	22.36	25.45
Stima Grassi	18.87	20.96	21.33	21.47	22.13	24.65

Tabella 3.5: Sommario delle variabili dipendenti dall'analisi chimica e dalla stima tramite modelli PCR.

A questo punto, andiamo a calcolare l'errore quadratico medio e l'indice di adattamento R^2 per i nostri definitivi modelli di regressione in componenti principali. In tabella 3.6 viene fornito, insieme alle due quantità appena annunciate, anche il relativo valore della statistica R^2 di WinISI.

	MSEP	R^2	WinISI R^2
PCR Proteine	0.54	0.89	0.90
PCR Umidità	0.13	0.97	0.97
PCR Grassi	0.37	0.76	0.77

Tabella 3.6: Statistiche di adattamento dei modelli PCR.

Si noti che essa coincide con quella calcolata in R solo per il secondo costituente, per il quale sono stati rilevati gli stessi outlier. Per gli altri due

costituenti, invece, l'adattamento dei modelli di WinISI sembra leggermente migliore ma ciò è probabilmente dovuto al fatto di aver eliminato un numero superiore di campioni "scomodi". L'adattamento dei modelli, comunque, oltre che molto simile al programma di riferimento sembra nel complesso buono. Come preannunciato dai grafici dei valori previsti, l'umidità è il costituente che è stato meglio predetto, ottenendo un R^2 pari a 0.97 con il solo utilizzo delle prime 4 componenti principali. Si è quindi ottenuto lo stesso adattamento di WinISI con tre componenti principali in meno.

3.3.2 Modelli di regressione *PLS*

Torniamo ad utilizzare la funzione `mvr_dcv` (A.11) impostando questa volta l'opzione `method` su `kernelpls`, per ottenere la doppia convalida incrociata per variabili latenti. Il grafico 3.7 mostra il grafico di probabilità stimata per il miglior numero di variabili latenti da includere nel modello; in tabella 3.7 viene riportata la frequenza con cui sono stati ottenuti i valori.

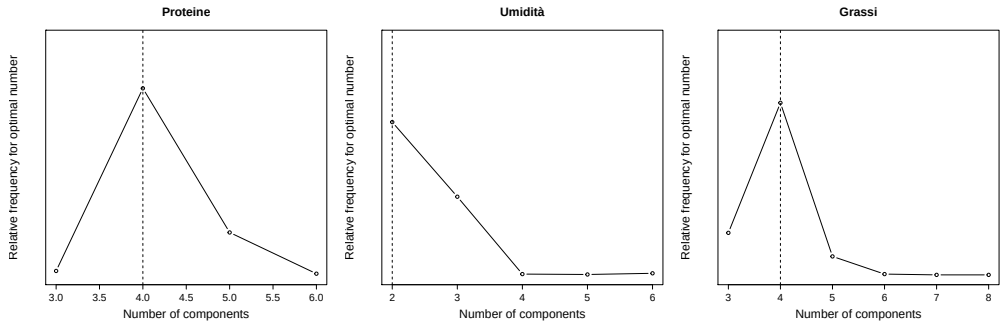


Figura 3.7: Distribuzione di probabilità stimata per il miglior numero di variabili latenti per i tre costituenti.

	2	3	4	5	6
Proteine	0.00	0.02	0.79	0.18	0.01
Umidità	0.65	0.33	0.00	0.00	0.01
Grassi	0.00	0.18	0.73	0.08	0.00

Tabella 3.7: Probabilità stimata per numero di regressori e costituente per i modelli PLSR.

Similmente al grafico ottenuto per le componenti principali, vediamo che `Proteine` e `Grassi` presentano lo stesso numero di variabili latenti e,

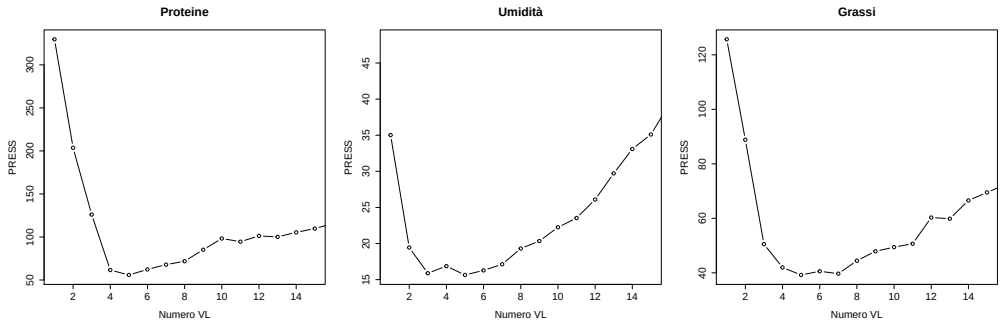


Figura 3.8: Andamento della statistica PRESS al variare del numero di variabili latenti.

nuovamente, `Umidita` richiede un numero minore di regressori. Da questo semplice esempio possiamo notare una caratteristica dei modelli di regressione *PLS*: l'utilizzo di un numero generalmente inferiore di parametri. Essi, infatti, restituiscono il minor numero necessario di variabili, consentendo a questa classe di modelli di essere molto stabili in previsione (Höskuldsson 1988). D'altronde il fatto di ruotare le variabili iniziali sia in direzione di massima varianza sia di maggior correlazione con la dipendente, non poteva che portare ad un utilizzo di parametri minore, dato che le variabili latenti, così costruite, sono maggiormente informative per la stima. Anche WinISI, che aveva determinato l'utilizzo di 7 componenti principali, con la *PLSR* si ferma a 5 variabili latenti. Per noi, il numero di variabili ausiliarie da utilizzare per costituente, sarà invece, come mostrato dal grafico e dalla tabella precedenti, 4, 2 e 4.

Eseguiamo quindi la funzione `plsr` (A.13) per stimare la classe di modelli presentata nel paragrafo 2.4. Anch'essa potrebbe essere limitata all'ottimo precedentemente trovato e, nella fase applicativa vera e propria, tale impostazione è notevolmente vantaggiosa. Poiché siamo interessati in questo momento a vedere come le nostre statistiche di adattamento cambiano al variare del numero di regressori, manteniamo un numero superiore.

I rispettivi grafici del *PRESS* per la *PCR* non davano così immediato riscontro di quanto detto nel capitolo precedente. Per tutti i costituenti è facile notare che un iniziale aumento del numero di variabili latenti provoca un veloce abbassamento della statistica *PRESS*, per poi farla progressivamente aumentare quando i parametri diventano meno significativi in ottica di miglioramento del potere predittivo del modello. Nuovamente

notiamo che il numero scelto non è il minimo assoluto bensì il numero per cui si è ottenuta una sensibile decrescita dal valore precedente anche se, in questo caso, i minimi relativi ed assoluti non sembrano differire di molto.

Andiamo quindi ad estrarre dalla `plsr`, per il numero di variabili latenti trovato in doppia convalida incrociata, le previsioni dei tre modelli per il successivo calcolo dell'outlier T. Questa volta, dopo una prima esecuzione dell'algoritmo A.14 per la determinazione degli outlier e la relativa esclusione dei campioni individuati, non si è ottenuto il modello definitivo. Una seconda iterazione per verificare nuovamente la presenza di campioni anomali ha infatti dato esito positivo. Il risultato finale dei due cicli è riportato in tabella 3.8.

	R	WinISI
Proteine	4, 41, 63	41, 63
Umidità	18, 43, 55, 70	74
Grassi	60, 63	63, 74

Tabella 3.8: Outlier rilevati da R e WinISI per i modelli PLSR.

Si nota che la determinazione degli outlier differisce in maniera più marcata fra i due *software* soprattutto per il secondo costituente. Mentre per le proteine ed i grassi la determinazione è abbastanza simile, per l'umidità le differenze sono chiare. Si fa nuovamente presente che tutti i calcoli sono soggetti a errori numerici e che il metodo ritenuto migliore per il nostro dataset, `kernelpls`, potrebbe non essere lo stesso adottato da WiniISI.

Escludiamo i campioni trovati nella doppia iterazione per la determinazione dell'outlier T ed andiamo a stimare i modelli *PLSR* definitivi. Calcoliamo quindi le previsioni dei costituenti e confrontiamole con i dati chimici iniziali. Il grafico di tali quantità per i diversi costituenti è mostrato in figura 3.9.

Per tutte le variabili i grafici sembrano dare segnali positivi evidenziando una buona corrispondenza fra valore stimato e misurato. La variabile che sembra essere stata peggio prevista è `Grassi` in cui qualche concentrazione particolarmente alta o bassa della sostanza, sembra discostarsi in maniera più decisa dalla retta di riferimento. Andiamo quindi

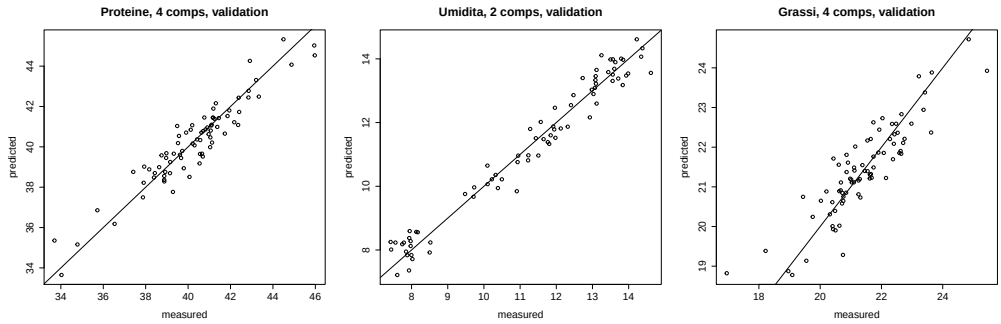


Figura 3.9: Confronto valori chimici e valori stimati con i modelli PLSR.

a vedere un sommario dei dati chimici e dei valori predetti per i singoli costituenti.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Proteine	33.70	39.02	40.55	40.28	41.23	45.97
Stima Proteine	33.65	39.00	40.35	40.23	41.35	45.33
Umidità	7.42	9.71	11.77	11.27	13.11	14.61
Stima Umidità	7.21	9.67	11.52	11.24	13.33	14.61
Grassi	16.95	20.67	21.29	21.40	22.31	25.45
Stima Grassi	18.78	20.78	21.39	21.42	22.14	24.72

Tabella 3.9: Sommario delle variabili dipendenti dall'analisi chimica e dalla stima tramite modelli PLSR.

Non ci resta che andare a verificare la bontà dei nostri modelli tramite il calcolo dell'errore quadratico medio e dell'R^2 per verificare le nostre congetture.

	MSEP	R^2	WinISI R^2
PLS Proteine	0.55	0.89	0.89
PLS Umidita	0.18	0.96	0.97
PLS Grassi	0.37	0.79	0.76

Tabella 3.10: Statistiche di adattamento dei modelli PLS.

Come anticipato dai grafici di dispersione della variabile dipendente, notiamo dalla tabella 3.10 che per la variabile `Umidita` è stato fatto un ottimo lavoro con solo 2 variabili latenti, ottenendo praticamente lo stesso R^2 di WinISI ma con meno della metà delle variabili. Nonostante le

considerazioni fatte sulla variabile `Grassi`, sebbene sia fra i tre costituenti quella descritta dal peggior modello, abbiamo ottenuto un buon valore della statistica R^2, di poco superiore a quella di WinISI.

Alla luce dei risultati ottenuti, riportati di seguito in tabella 3.11, non ci resta che decidere che modello utilizzare a seconda del costituente. Poiché l'attenzione dei modelli è posta sulla precisione della previsione, siamo portati a scegliere l'uso di un modello di regressione in componenti principali per `Proteine` ed `Umidità` ed uno in variabili latenti per `Grassi`.

	R^2		MSE	
	PCR	PLSR	PCR	PLSR
Proteine	0.89	0.89	0.54	0.55
Umidità	0.97	0.96	0.13	0.18
Grassi	0.76	0.79	0.37	0.37

Tabella 3.11: Sommario dei valori R^2 e MSE ottenuti dai modelli stimati per i vari costituenti.

A questo punto il processo della calibrazione può definirsi compiuto. Il passo finale è andare ad estrarre, dagli oggetti creati con le funzioni `pcr` e `plsr`, i coefficienti di regressione necessari per prevedere la concentrazione a partire da spettri incogniti. Ciò che è stato mostrato, infatti, è la parte di calcolo affidata ad un *personal computer*. Dopo aver estratto i coefficienti ed averli inseriti all'interno dello spettrometro poliSPEC^{NIR}, sarà possibile effettuare nuove rilevazioni sulla soia, coerentemente con la calibrazione fornita allo strumento, ed ottenere direttamente sul luogo di rilevazione la quantificazione delle sostanze oggetto d'interesse.

Capitolo 4

Conclusioni

All'interno della relazione abbiamo visto come la chemiometria sia una valida alternativa alle analisi chimiche. I vantaggi possono essere riassunti in un minor costo ed una maggior velocità di analisi, senza contare che questa procedura rappresenta un metodo non distruttivo che richiede scarsa, se non nulla, preparazione della sostanza.

La stima dei modelli statistici, eseguiti per la previsione dei costituenti, ha messo in luce come le stime siano sostanzialmente dotate di buona precisione e possano sostituire i classici metodi di laboratorio. Tali modelli, creati seguendo le due principali metodologie, non sono comunque gli unici disponibili nel campo delle analisi chemiometriche. Al variare del dataset, invero, potrebbe essere più vantaggioso l'uso di diverse strategie (come una procedura stepwise, la regressione Lasso o Ridge) al fine di determinare una più vasta gamma di modelli su cui effettuare poi la scelta.

Lo stesso pretrattamento dei dati potrebbe essere opinabile (Forina 2010) ma, dati gli scopi di emulazione del software creato, tale eventualità non è stata presa in considerazione. Tuttavia, laddove ci fosse un beneficio, questo risulterebbe comunque di scarsa entità e non stravolgerebbe i risultati a cui si è giunti.

Anche la regressione in componenti principali, d'altra parte, risulta poco affidabile in caso di dataset con un maggior numero di variabili rispetto alle osservazioni, a causa di un inconsistente rotazione degli assi. Quest'ultimo aspetto, spesso non approfondito dai tecnici del settore, è però un problema frequente. La dimostrazione teorica, superiore alla praticità

di questo contesto, non viene ad ogni modo presentata.

Nonostante ciò, l'utilizzo di una regressione su strutture latenti non soffre di tale disparità fra osservazioni e variabili, permettendoci comunque di raggiungere risultati affidabili, dalle qualità simili alla regressione in componenti principali.

Bibliografia

[1] Adelchi Azzalini e Bruno Scarpa. *Analisi dei dati e data mining*. Springer, 2004.

[2] Peter Filzmoser, Bettina Liebmann e Kurt Varmuza. «Repeated double cross validation». In: *Journal of Chemometrics* 23 (2009).

[3] Livio Finos. *Dispensa Classificazione e Analisi di Dati Multidimensionali*. 2014.

[4] Michele Forina. *Fondamenta per la Chimica Analitica*. 2010. URL: http://gruppochemiometria.it/.

[5] Heide Garcia e Peter Filzmoser. *Multivariate Statistical Analysis using the R package chemometrics*. 2011. URL: http://cran.r-project.org/web/packages/chemometrics.

[6] Enrico Gregorio e Luigi Salce. *Algebra Lineare*. Libreria Progetto, 2010.

[7] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2008.

[8] Agnar Höskuldsson. «PLS regression methods». In: *Journal of Chemometrics* 2 (1988).

[9] Harold Hotelling. «Analysis of a complex of statistical variables into principal components». In: *ournal of Educational Psychology* (1933).

[10] Karl Pearson. «On lines and planes of closest fit to systems of points in space». In: *Philosophical Magazine Series* (1901).

[11] *SensoLogic Calibration Workshop*. 2010.

[12] Roberto Todeschini. *Introduzione alla Chemiometria*. EdiSES, 2003.

[13] Svante Wold, Michael Sjöström a e Lennart Eriksson. «PLS-regression: a basic tool of chemometrics». In: *Chemometrics and Intelligent Laboratory Systems* 58 (2001).

Appendice A

Codice R

Prima di eseguire le funzioni di seguito presentate, carichiamo la libreria `pracma`, che verrà sfruttata per il detrend, e `chemometrics` ad opera di K. Varmuza e P. Filzmoser (Garcia e Filzmoser 2011). All'interno di tale pacchetto troveremo buona parte delle funzioni necessarie. Esso, oltre a contenere alcune funzioni scritte dai suoi autori, importa automaticamente anche tutte le librerie utili ad affrontare un'analisi chemiometrica, dai modelli di regressione alla validazione incrociata.

Codice A.1: *Standard Normal Variate*

```
SNV.f=function(x){
  SNV=x/apply(x,1,sd)
  SNV
}
```

Codice A.2: Derivata

```
derivate.f=function(x,der=1,gap=4){
  der=-t(diff(t(x),lag=gap,differences=der))
  der
}
```

Codice A.3: Lisciamento

```
smooth.f=function(x,by=4){
  n=dim(x)[1]
  m=dim(x)[2]-by+1
  smooth=matrix(0,n,m)
```

```
  for(i in 1:n){
    for(j in 1:m){
      smooth[i,j]<-mean(x[i,j:(j+by-1)])
    }
  }
  smooth
}
```

Codice A.4: *Multiplicative Scatter Correction*

```
MSC.f=function(x){
  library(pls)
  t=t(x)
  MSCt=msc(t)
  MSC=t(MSCt)
  MSC
}
```

Codice A.5: Applicazione pretrattamenti

```
A=SNV.f(spettri)
B=t(detrend(as.matrix(t(A))))
C=derivate.f(B)
D=smooth.f(C)
```

Codice A.6: Scomposizione spettrale

```
E=eigen(cov(D))
VEC=t(E$vec)
VAL=E$values
SCO=D%*%t(VEC)
```

Codice A.7: Esplorazione varianza

```
varexpl=colMeans(pcaCV(D,amax=50,repl=10,segments=6,plot.
   opt=FALSE)$ExplVar)
```

Codice A.8: Selettore CP

```
soglia=0.95
```

```
for(i in 1:length(varexpl)){
  if(varexpl[i]>=soglia) {
    break
  }
  else(i=i+1)
  numeroCP=i
}

numeroCP
```

Codice A.9: Mahalanobis

```
sigma=cov(SCO[,1:numeroCP])
mu=colMeans(SCO[,1:numeroCP])
x=SCO[,1:numeroCP]
D2=mahalanobis(x,mu,sigma)
out=which(D2/sd(D2)>10)
```

Codice A.10: Impostazione del dataframe

```
df=data.frame(D,comp)
```

Codice A.11: Doppia *cross-validation*

```
pcr=mvr_dcv(comp~.,data=df,ncomp=30,validation="CV",method=
   "",segments0=6,segments=6,repl=100)
```

Codice A.12: Regressione *PCR*

```
pcrv=pcr(comp~.,data=df,ncomp=25,validation="CV",method="
   svdpc",segments=6,repl=1000)
```

Codice A.13: *PLS*

```
plsv=plsr(comp~.,data=df,ncomp=25,validation="CV",method="
   kernelpls",segments=6,repl=1000)
```

Codice A.14: Outlier T

```
SEC=sqrt(sum((pcrv$residuals[,,NCpcr])^2)/(81*5/6-1-NCpcr))
test.t=(pcrv$residuals[,,NCpcr])/(SEC*sqrt(1-(8/((5/6)*81))
   ))
```

```
names(test.t)=NULL
which(abs(test.t)>2.5)
```