UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI PRINCIPI E IMPIANTI DI INGEGNERIA CHIMICA "I. Sorgato"

TESI DI LAUREA IN INGEGNERIA CHIMICA

METODOLOGIE PER LA SELEZIONE DELLE VARIABILI IN INGRESSO NELLO SVILUPPO DI SENSORI SOFTWARE BASATI SU TECNICHE STATISTICHE MULTIVARIATE

Relatore: Prof. Massimiliano Barolo
Correlatore: Ing. Pierantonio Facco

Laureando: MATTEO TONIZZO

ANNO ACCADEMICO 2006-2007

Riassunto

In questa Tesi viene presentata l'applicazione di algoritmi di selezione di variabili di processo per la realizzazione di sensori *software* per la stima di proprietà non misurabili in linea. Sono stati confrontati modelli di predizione ottenuti mediante proiezione su strutture latenti (PLS), calibrati su tutte le misure disponibili, con modelli costruiti su un sottoinsieme ottimo di misure.

Gli algoritmi di selezione di variabili utilizzati sono 4: tre di questi - Stepwise Regression (Draper e Smith, 1998), SROV (Shacham e Brauner, 2003), AS (Muradore *et al.*, 2006) - sono algoritmi che selezionano iterativamente una variabile alla volta in base a criteri diversi di tipo numerico o statistico, mentre il quarto algoritmo - VIP (Chong e Jun, 2005) - seleziona le variabili solo dopo aver costruito il modello PLS contenente tutte le variabili predittive iniziali.

L'applicazione degli algoritmi ha riguardato due processi differenti: il primo è un processo di distillazione binaria, il secondo un processo *batch* di produzione di una resina. Si è trovato che le misure di qualità stimate dai modelli PLS contenenti sottoinsiemi di variabili selezionate con l'algoritmo VIP sono più accurate rispetto alle stime prodotte dai modelli PLS contenenti sottoinsiemi di variabili selezionate con gli altri algoritmi, ma solo se il modello iniziale di partenza è costruito opportunamente. Rispetto agli altri algoritmi, inoltre, VIP seleziona sempre variabili importanti dal punto di vista del processo e permette sempre di ottenere un sottoinsieme significativo di variabili predittive.

La selezione preventiva delle grandezze sulle quali progettare lo stimatore ha permesso di diminuire il numero di misure in ingresso al modello; l'utilizzo di un sottoinsieme ridotto di variabili di processo consente di ridurre il rischio di avarie degli strumenti di misura.

Indice

Introduzione

Nei moderni siti industriali la strumentazione di controllo è presente in grande quantità e permette la registrazione con intervalli di campionamento molto brevi, nell'ordine dei secondi, di decine o centinaia di variabili la cui traiettoria è così seguita lungo tutta la durata del processo, sia esso continuo o discontinuo.

Spesso, però, l'output di un processo non è direttamente misurabile in linea; questo accade soprattutto per le misure di qualità, a causa dell'alto costo di impianto e manutenzione degli analizzatori in linea (gascromatografi, o altra strumentazione). Si preferisce, quindi, prelevare un campione in campo e analizzarlo in laboratorio; di conseguenza, non è possibile conoscere in tempo reale lo stato complessivo del sistema, il tipo di regolazione da eseguire e il risultato di tale regolazione: ciò introduce un sensibile ritardo (*delay*) nell'anello di regolazione del processo, che rende più difficile e meno affidabile il controllo del sistema.

Per superare questo inconveniente, nella pratica industriale si attua il controllo del processo monitorando alcune variabili sensibili ad ogni cambiamento delle condizioni del sistema, come temperature, pressioni, densità e portate: se si conosce la relazione esistente tra le condizioni del processo e le variabili monitorate, e tra queste ultime e le variabili di qualità, è possibile costruire un modello di controllo del processo, che per ogni condizione del sistema, così come monitorata dalle variabili osservate, indichi le eventuali correzioni da apportare attraverso le variabili manipolabili per ottenere la specifica desiderata per le variabili di qualità, senza dover attendere il risultato delle analisi di laboratorio. In questo modo si elimina il ritardo all'interno dell'anello di regolazione.

Il cuore di questo modello è lo *stimatore inferenziale* o *sensore software*: esso è costituito da una o più equazioni, le quali permettono di determinare in modo istantaneo il valore della variabile di output, dopo aver ottenuto in ingresso le misure delle variabili osservate.

La condizione più semplice di controllo è quella in cui si dispone di un modello matematico meccanicistico del processo che si sta regolando; questo modello matematico è un'espressione rigorosa che correla le variabili monitorate (*input*) con le variabili manipolate e con le variabili di qualità (*output*). Solitamente, il modello deterministico, detto anche modello a principi primi, è esplicitato mediante una o più equazioni di bilancio di materia, energia e quantità di moto (equazioni della fluidodinamica), i cui parametri sono coefficienti reali del processo, aventi significato fisico. Grazie al modello deterministico, è possibile conoscere a priori il risultato di ogni azione sul processo e

quindi regolare e controllare il processo rispetto ad ogni esigenza produttiva e in modo da risolvere ogni possibile inconveniente.
Molto spesso, però, il modello deterministico non è noto: ciò significa che non si conosce la reale relazione esistente tra variabili di *input* e di *output*, o che non sono noti o ricavabili alcuni parametri di questo modello, oppure che il processo è così complesso che non si ha interesse a studiare a fondo il modello deterministico o a regolarlo mediante tale modello. In assenza di un modello a principi primi, si può cercare di approssimare il reale comportamento del processo mediante una equazione empirica di regressione basata esclusivamente sui dati disponibili dalle apparecchiature di misura. Le tecniche di analisi statistica multivariata basate sul concetto di variabile latente forniscono gli strumenti necessari per estrarre da questa moltitudine di dati le informazione necessarie separandole dal rumore e dai disturbi presenti nel processo. L'applicazione dei metodi di regressione ed analisi statistica multivariata permette di identificare o stimare la relazione esistente tra i due set di variabili, osservate e di qulità; in questo modo, si ottiene un modello matematico empirico del processo che costituisce la base del modello di regolazione.
Un modello di questo tipo è detto modello *black-box* o empirico.
L'applicazione dei modelli empirici è complicata quando la strumentazione dell'impianto dispone di molte variabili predittive e dipendenti monitorate: in questo caso diventa infatti difficile inserire tutte le variabili nel modello empirico di regressione, o selezionare manualmente le variabili utili per il modello, anche perché le variabili sono altamente correlate tra loro. Inoltre, molto spesso le variabili monitorate hanno un rapporto segnale su rumore molto basso, cioè seguono una traiettoria difficilmente riconoscibile in quanto mascherata dal rumore di misura: in questi casi è difficile separare l'informazione utile dal rumore e capire la reale correlazione tra la variabile predittiva monitorata e le altre variabili del processo.
Tutti questi inconvenienti rendono meno robusti gli stimatori basati su tecniche di analisi statistica multivariata: diventa quindi essenziale effettuare una selezione preliminare delle variabili predittive da utilizzare nel modello di regressione.
La Tesi si propone l'obiettivo di testare diverse metodologie di selezione di variabili che riducano il numero di variabili monitorate e quindi il costo di monitoraggio e il rischio di avarie dei sensori. Le variabili selezionate dai diversi algoritmi saranno successivamente utilizzate e confrontate nello sviluppo di sensori *software* basati su tecniche statistiche multivariate nella stima delle qualità di un prodotto non misurabile direttamente in linea.

Capitolo 1

Regressione lineare multivariata

Vengono presentati due metodi di regressione lineare multipla: la regressione ai minimi quadrati classici (*Ordinary Least Squares regression*, OLS) e la Proiezione su Strutture Latenti (*Projection on Latent Structures*, PLS). Ciò implica la definizione di elementari operazioni di calcolo matriciale e di alcune funzioni, tecniche e operatori basilari della statistica. Insieme all'algoritmo di regressione, si espongono anche le tecniche di inferenza statistica utili ad ottenere stime puntuali ed intervalli di confidenza dei parametri di regressione, intervalli di confidenza per la stima dell'errore puntuale, test statistici per la verifica di ipotesi.
Successivamente, si espongono alcuni criteri per giudicare la correttezza statistica della regressione e la capacità predittiva del modello.

1.1 Modelli lineari multivariati

Si disponga di n misure sperimentali di k variabili indipendenti $x_1,\ldots, x_k$, e di altrettante misure della variabile dipendente y, con $n > k$.
Un modello lineare multivariato coinvolge una o più variabili dipendenti, rappresentate come funzione lineare di più variabili indipendenti o predittive. Il modello di equazione lineare multivariata è:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij} \qquad \forall i = 1,\ldots,n \quad , \tag{1.1}$$

dove y_i è l'i-esimo valore della variabile dipendente, $x_1,\ldots, x_k$ sono le k variabili indipendenti e $\beta_0, \beta_1,\ldots, \beta_k$ sono $k + 1$ parametri matematici o fisici.
L'analisi di un processo produttivo raramente conduce ad equazioni come (1.1), sia che si abbia a disposizione un modello deterministico lineare sia che si utilizzi un modello empirico: infatti le misure campionate delle variabili di processo, sia predittive che dipendenti, risultano affette da rumore di misura e portano un errore stocastico sistematico all'interno del modello. L'equazione lineare viene rappresentata come in (1.2):

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad \forall i = 1,\ldots,n \quad . \tag{1.2}$$

Questo modello può essere applicato sia ai sistemi deterministici che ai modelli empirici: nel primo caso la relazione (1.2) è esatta, rappresenta il reale comportamento del processo e i parametri non sono affetti da errore; nel secondo caso l'equazione (1.2) è imprecisa, in quanto non rappresenta con sicurezza la reale struttura del processo e identifica valori solo stimati per i parametri di processo e per la variabile dipendente.
La differenza tra i modelli deterministici e quelli empirici dipende anche dal significato dell'errore ε_i: nei modelli meccanicistici questo è un errore casuale a distribuzione normale facilmente quantificabile, mentre nei modelli empirici è il risultato di tutte le imprecisioni, come l'inesattezza del modello, il mancato inserimento di variabili di processo importanti (o l'inclusione di variabili trascurabili) e la stima errata dei parametri di regressione.
L'equazione (1.2), nel caso dei modelli empirici, viene ottenuta con i metodi della regressione statistica, stimando i coefficienti di regressione in modo da minimizzare l'errore. In questo caso i coefficienti di regressione parziale $\beta_0,\ldots, \beta_k$ non hanno significato fisico ma solo statistico: permettono di costruire un'equazione che riproduce correttamente il profilo sperimentale della variabile dipendente per ogni punto delle variabili predittive, minimizzando l'errore ε_i.
L'applicazione del modello di regressione lineare ai dati di processo è resa complicata da numerose difficoltà: sia la variabile dipendente che le variabili predittive sono affette da errore di misura, le variabili sono molto correlate tra loro, i dati sono imprecisi o in alcuni casi mancanti, a causa della rottura o malfunzionamento di un sensore. Inoltre, il rapporto segnale/rumore delle variabili può essere molto basso, introducendo rumore sistematico all'interno dell'equazione di regressione: questo fatto è molto importante, perché induce l'equazione di regressione, attraverso i coefficienti di regressione, a modellare l'errore stesso misurato dalle variabili predittive e dipendenti per minimizzare l'errore.
Ciò significa che anche un'equazione di regressione statisticamente significativa che ottenga un buon *fitting* dei dati sperimentali e minimizzi l'errore in (1.2) può essere non utilizzabile se prima non viene separata l'informazione utile dal rumore.
Prima di presentare il metodo di regressione ai minimi quadrati e le tecniche statistiche multivariate verranno introdotte alcune nozioni di chemometria; senza soffermarsi sugli aspetti teorici, si presenteranno le nozioni utili nel corso della tesi.

1.2 Fondamenti di chemometria di processo

Si disponga di n misure sperimentali di una variabile di processo $x_1,\ldots,x_n$; matematicamente, si definisce *media campionaria* di una variabile aleatoria la statistica $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i) \quad , \tag{1.3}$$

dove il campione di n misure è assunto essere rappresentativo della popolazione.
Una variabile è detta *centrata* rispetto alla media se le n misure che la compongono sono corrette sottraendo ad ognuna di esse la media della variabile stessa; la variabile così ottenuta, designata con il simbolo x^*, ha media nulla.
Si definisce *varianza campionaria* della variabile aleatoria x la statistica s^2 o $var(x)$:

$$s_x^2 = \mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad , \tag{1.4}$$

dove $\bar{x}$ è la media campionaria della variabile aleatoria x, ed n – 1 sono i gradi di libertà della funzione. La varianza è una misura della intensità con cui la variabile x oscilla intorno al valore atteso, cioè della sua variabilità.
Si definisce *errore standard* del campione, s, la radice quadrata della varianza campionaria.
Una variabile è detta *normalizzata* se è centrata e le n misure sono divise per la deviazione standard della variabile stessa; la variabile così ottenuta, definita con il simbolo $\tilde{x}$, è adimensionale, ha media nulla e varianza unitaria.
Si disponga di n misure sperimentali di k variabili di processo; i dati così ottenuti possono essere raccolti in una matrice $\mathbf{X}$ di dimensione $n \times k$, che contiene in ogni colonna le misure ordinate di ogni variabile aleatoria, e in ogni riga le misure, ottenute ad un determinato istante di campionamento t, di tutte le variabili. L'elemento della matrice è indicato come x_{ij}, i-esima misura della j-esima variabile predittiva, $1 \leq i \leq n$, $1 \leq j \leq k$.
Si definisce *covarianza campionaria* di due variabili aleatorie la statistica $cov(x_1, x_2)$:

$$\mathrm{cov}(x_1, x_2) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \quad . \tag{1.5}$$

La covarianza è un indice della relazione lineare esistente tra due variabili, cioè della loro variabilità relativa. La covarianza può assumere valori diversi a seconda dell'unità di misura con cui vengono riportate le variabili, e di conseguenza non è un buon indicatore. In generale, comunque, un valore assoluto elevato indica l'esistenza di una forte correlazione tra due variabili, viceversa nel caso il valore della covarianza tenda a zero.
Si definisce correlazione di due variabili aleatorie la statistica $corr(x_1, x_2)$:

$$corr(x_1, x_2) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{1i}-\bar{x}_1}{s_1}\right)\left(\frac{x_{2i}-\bar{x}_2}{s_2}\right) = \frac{1}{n}\frac{\mathrm{cov}(x_1, x_2)}{\sqrt{\mathrm{var}(x_1)\mathrm{var}(x_2)}} \quad . \tag{1.6}$$

La correlazione, è un indice adimensionale, compreso tra –1 e 1, della relazione lineare esistente tra due variabili normalizzate. Un valore di correlazione positivo indica che le due variabili tendono a crescere o a decrescere simultaneamente, mentre un valore negativo indica che tendenzialmente le due variabili hanno carattere opposto. Infine, due variabili linearmente indipendenti l'una dall'altra hanno correlazione nulla. Ciò non significa che le due variabili sono indipendenti tra loro, cioè che sono ortogonali, ma solo che non esiste nessuna relazione lineare in grado di esprimere una variabile in funzione dell'altra; può però esistere una funzione di qualsiasi altro tipo in grado di mettere in relazione le due variabili.

Si definisce *matrice di varianza-covarianza* il prodotto matriciale $n^{-1}\,(\mathbf{X}^*)^{\mathrm{T}}(\mathbf{X}^*)$, dove $\mathbf{X}^*$ è la matrice, organizzata come $\mathbf{X}$, contenente le variabili centrate. La matrice di varianza-covarianza ha dimensioni $k \times k$, dove k è il numero di variabili della matrice $\mathbf{X}^*$; essa è costituita dai valori di varianza campionaria di tutte le variabili disposti in ordine lungo la diagonale, e dai valori di covarianza campionaria all'esterno della diagonale: ad esempio, nella posizione (i,i) si troverà la varianza della variabile i, mentre nelle posizioni (i,j) e (j,i) si leggerà la covarianza delle variabili i e j. Da quanto detto si capisce che la matrice di varianza-covarianza è una matrice simmetrica.

La matrice $(\mathbf{X}^*)^{\mathrm{T}}(\mathbf{Y}^*)$ è detta *matrice di covarianza tra variabili indipendenti e dipendenti*: essa esprime l'entità della relazione lineare esistente tra ogni variabile dipendente e tutte le variabili indipendenti: una variabile x fortemente correlata a y ha un valore di covarianza molto elevato. La covarianza è quindi un indicatore della reale importanza di ogni variabile indipendente alla costruzione del modello di regressione tra variabili dipendenti e predittive, ed è un possibile criterio discriminante per la scelta delle variabili più idonee alla costruzione di un modello.

Si definisce *matrice di correlazione* il prodotto matriciale $(n-1)^{-1}\,\widetilde{\mathbf{X}}^{T}\,\widetilde{\mathbf{X}}$, dove $\widetilde{\mathbf{X}}$ è la matrice, organizzata come $\mathbf{X}$, contenente le variabili normalizzate. Anche la matrice di correlazione ha dimensione $k \times k$ ed è simmetrica, ma i suoi elementi hanno valori diversi: assumono valore unitario lungo la diagonale ed oscillano tra –1 ed 1 all'esterno di questa.

Analogamente al caso della covarianza, può essere costruita una matrice di correlazione tra variabili predittive e dipendenti: $corr(\mathbf{X}, \mathbf{Y}) = (n-1)^{-1}\,\widetilde{\mathbf{X}}^{T}\,\widetilde{\mathbf{Y}}$.

Le proprietà della media e della varianza sono utilizzate nell'inferenza statistica per fornire informazioni sui parametri stimati di un'equazione di regressione; ad esempio, se si ha una stima degli errori ε_i, dalla varianza degli errori si può ricavare un valore che indichi la probabilità di ogni punto di essere uguale al valore atteso e un intervallo, detto intervallo di confidenza, all'interno del quale avere il 95% di probabilità di trovare il valore atteso.

L'intervallo di confidenza che fornisce il 100(1 – α)% di probabilità che l'intervallo contenga il valore atteso μ della statistica è dato dall'espressione (1.7):

$$\frac{\bar{x} - t_{\alpha/2,n-1}s}{\sqrt{n}} \leq \mu \leq \frac{\bar{x} + t_{\alpha/2,n-1}s}{\sqrt{n}} \quad , \tag{1.7}$$

dove $t_{\alpha/2,n\text{-}1}$ è il valore, ottenuto da tabella (Draper e Smith, 1998), che fornisce la probabilità cumulata per una distribuzione di probabilità *t* di Student standardizzata compresa tra – α/2 e 1 – α/2, di ampiezza 1 – α. Solitamente α = 0.05, di modo che l'intervallo di confidenza fornisca una probabilità del 95%.

Si definisce funzione di distribuzione *t* di Student di una variabile aleatoria *x* la distribuzione di probabilità riportata nell'espressione (1.8):

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \cdot \frac{1}{\left[\left(\frac{x^2}{k}\right)+1\right]^{\left(\frac{k+1}{2}\right)}} \quad , \tag{1.8}$$

dove *k* è il numero di gradi di libertà e Γ la distribuzione di probabilità gamma; la distribuzione *t* sostituisce la distribuzione normale quando non è disponibile il valore della varianza della popolazione, o quando il campione di misure non è sufficiente per accertare la distribuzione normale.

La distribuzione *t* approssima il comportamento della distribuzione normale, risultando però più dispersa: ciò significa che, per avere la stessa probabilità di una distribuzione normale, l'intervallo di confidenza deve essere più ampio.

La somiglianza tra distribuzione *t* e distribuzione normale aumenta al crescere del numero di gradi di libertà: maggiore è il numero di punti disponibili per calcolare una statistica (ad esempio, il coefficiente di regressione), più piccolo è l'intervallo di confidenza che si ottiene sulla stima della statistica stessa.

Gli intervalli di confidenza conducono ad uno strumento statistico molto importante: la verifica di ipotesi.

A volte è importante sapere se un coefficiente di regressione è nullo: si deve quindi verificare l'ipotesi h_0: $b = 0$ contro il rigetto dell'ipotesi stessa. Ciò può essere verificato assumendo che il valore stimato da verificare abbia distribuzione *t* e stabilendo un intervallo di confidenza (solitamente del 95%) entro il quale il valore stimato deve essere considerato nullo. Dopo aver stabilito queste premesse, si calcola sulla base del campione disponibile il valore della statistica e si rigetta l'ipotesi h_0 con una probabilità del 95% se il valore stimato non rientra all'interno dell'intervallo detto. Se invece il valore stimato rientra all'interno dell'intervallo, si dice che l'ipotesi non è stata rigettata: non si può

parlare infatti di accettazione dell'ipotesi, perché non si ha la certezza che il valore trovato sia il valore atteso, o che il valore atteso si trovi all'interno dell'intervallo considerato.
L'estensione del concetto di funzione di distribuzione e funzione cumulativa al caso di più variabili è molto semplice ed avviene mediante la definizione della funzione di distribuzione congiunta, che esprime la probabilità che più funzioni di distribuzione assumano un definito *set* di valori; per ottenere la funzione cumulativa corrispondente si deve risolvere l'integrale multiplo calcolato per i vari intervalli stabiliti per ciascuna funzione. Se le varabili sono indipendenti, cioè non correlate tra loro in alcun modo, l'integrazione è semplice; se, invece, le variabili sono correlate, il calcolo dell'integrale è reso difficoltoso dal fatto che ogni funzione è implicitamente contenuta nella definizione delle altre funzioni. Sarà quindi necessaria un'integrazione numerica.
Anche nel caso di distribuzioni congiunte è possibile calcolare stime puntuali e intervalli di confidenza utilizzando tecniche di inferenza statistica: la trattazione è però molto onerosa, e non viene riportata nella Tesi.
E' utile comunque notare che la distribuzione *F* è l'equivalente della distribuzione normale per una distribuzione bivariata, e viene utilizzata per stimare la probabilità che una equazione di regressione sia esatta.

1.3 Regressione lineare ai minimi quadrati (*OLS*)

La regressione ai minimi quadrati (*Ordinary Least Squares*, OLS) risolve un sistema di n equazioni lineari come l'equazione (1.1), in cui sono incogniti i $k + 1$ coefficienti di regressione β_0, β_1,…, β_k corrispondenti ad ogni variabile predittiva più l'intercetta, minimizzando la varianza dell'errore ε.
La regressione ai minimi quadrati è possibile solo se il numero di campioni disponibili è superiore al numero di variabili predittive; infatti se $n < k$ il sistema non è risolvibile in modo univoco, ma si hanno infinite soluzioni. Nel caso invece si abbia $n = k$ la soluzione esatta del sistema lineare è possibile ed è unica: si ottiene per sostituzione o mediante algoritmi numerici come la decomposizione di Gauss.
L'equazione di regressione multivariata ai minimi quadrati è riportata in (1.9):

$$\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_k x_{ik} = b_0 + \sum_{j=1}^{k} b_j x_{ij} \qquad \forall i = 1,...,n \quad , \tag{1.9}$$

dove b_0, b_1,…,b_k sono le stime dei coefficienti di regressione parziale ottenute risolvendo il sistema di equazioni normali ricavate differenziando l'errore totale $\Sigma_i \varepsilon_i$ rispetto ad ogni incognita β_0, β_1,…, β_k (Montgomery, 2003).
Si definisce scarto o residuo dell'equazione di regressione la differenza (1.10) tra valore calcolato e sperimentale della variabile dipendente:

$$e_i = y_i - \hat{y}_i = y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij} \quad . \tag{1.10}$$

Oltre a tenere conto delle imprecisioni elencate per ε_i nel § 1.1, e_i tiene anche conto dell'imprecisione che si introduce utilizzando valori stimati non esatti per i coefficienti di regressione. I residui, nel caso il modello sia esatto, hanno queste caratteristiche:

1. si distribuiscono normalmente e non sono tra loro autocorrelati. La verifica di questa proprietà non è necessaria per la stima dei coefficienti di regressione, che è comunque possibile qualsiasi sia la distribuzione dei residui, ma per poter applicare i test statistici di verifica per la bontà della regressione;
2. hanno media nulla: $(\Sigma_i\, e_i) / n = 0$;
3. sono non correlati con la variabile calcolata: $\mathrm{cov}(e_i, \hat{y}_i) = 0$.

La varianza degli scarti è $s^2 = (n - k - 1)^{-1}\, \Sigma_i(e_i^2)$; la varianza degli scarti rappresenta una stima *unbiased* della varianza dell'errore ε.

Questa statistica viene definita spesso anche come scarto quadratico medio o *Mean Square Error* (MSE), e dipende sia dal modello di regressione che dai dati disponibili: al crescere del numero di punti sperimentali essa diminuisce, perché aumenta il valore al denominatore.

Il criterio del metodo di regressione ai minimi quadrati diventa quindi la minimizzazione della funzione $\mathrm{MSE} = (n - k - 1)^{-1}\, \Sigma_i(y_i - \hat{y}_i)^2$, in cui viene sostituito l'errore con lo scarto.

MSE è un buon criterio di giudizio della bontà di un'equazione, perché è un indice della dispersione degli errori rispetto alla retta o iperpiano di regressione.

Quanto più l'errore di stima sui coefficienti di regressione è elevato, tanto più il modello è errato (soffre di *lack of fit*) e non minimizza la funzione MSE.

Il metodo di regressione ai minimi quadrati può essere applicato anche a più variabili dipendenti contemporaneamente, affiancando i vettori colonna delle variabili dipendenti e creando la matrice $\mathbf{Y} = [y_1, \ldots, y_p]$.

In notazione matriciale l'equazione (1.2) si scrive $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, dove $\mathbf{y}$ è il vettore $n \times 1$ di osservazioni, $\mathbf{X}$ è la matrice $n \times (k+1)$ di dati delle variabili predittive, $\boldsymbol{\beta}$ è un vettore $(k+1) \times 1$ di coefficienti di regressione ed $\boldsymbol{\varepsilon}$ è il vettore $n \times 1$ degli errori.

Il sistema di equazioni normali del metodo di regressione ai minimi quadrati si scrive, in modo compatto:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \quad . \tag{1.11}$$

La soluzione dell'espressione matriciale (1.11) è:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\, \mathbf{X}^{\mathrm{T}}\mathbf{y} \quad . \tag{1.12}$$

Si ottiene quindi l'equazione di regressione lineare multivariata ai minimi quadrati in formato matriciale:

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} \quad , \tag{1.13}$$

dove $\hat{\mathbf{y}}$ è il vettore della variabile dipendente calcolata e **b** il vettore delle stime dei coefficienti di regressione parziali.

La soluzione (1.13) delle stime dei coefficienti di regressione parziale minimizza la funzione MSE.

Le stime b_j dei coefficienti di regressione parziali β_j calcolati con il metodo ai minimi quadrati hanno le seguenti proprietà (Montgomery, 2003):

1. minimizzano MSE, indipendentemente dalle proprietà della distribuzione degli errori e_i;
2. hanno deviazione standard $\mathbf{s}_b = diag\{((\mathbf{X}^T\mathbf{X})^{-1}s^2)^{1/2}\}$: più la deviazione standard è piccola più precisa è la stima del coefficiente; se il modello è errato o soffre di multicollinearità l'errore standard è molto alto e la precisione dei valori calcolati $\hat{y}$ molto inferiore, perché i termini diagonali dell'inversa della matrice di varianza-covarianza $\mathbf{X}^T\mathbf{X}$ assumono valori molto elevati;
3. sono stime attendibili di $\boldsymbol{\beta}$, se le variabili indipendenti sono tra loro non correlate o se i residui e_i sono statisticamente indipendenti.

La soluzione del problema ai minimi quadrati è possibile solo se la matrice $(\mathbf{X}^T\mathbf{X})^{-1}$ esiste, cioè se la matrice **X** ha rango massimo; questo accade solo se **X** è non singolare, con tutte le sue colonne linearmente indipendenti.

Nel caso le variabili siano tra loro linearmente correlate e non sia possibile effettuare l'inversione della matrice $\mathbf{X}^T\mathbf{X}$, si deve rinunciare ad ottenere una stima ai minimi quadrati dei coefficienti di regressione, utilizzando metodi alternativi per la risoluzione di sistemi lineari di equazioni come la decomposizione QR.

Infine, è importante notare che la regressione ai minimi quadrati classici si basa su un'assunzione che viene raramente rispettata dai dati di processo, ovvero l'esattezza delle misure sperimentali delle variabili predittive: l'assenza di questo requisito fa cadere la validità dell'equazione di regressione ai minimi quadrati, perché il modello non è in grado di verificare statisticamente l'esattezza delle stime dei coefficienti di regressione, l'entità del *bias*, il valore degli intervalli di confidenza, l'ampiezza degli intervalli di predizione e la attendibilità delle stime effettuate con dati di convalida.

1.4 Metodi di regressione multivariati

I metodi di regressione multivariata sono strumenti di regressione nati dall'esigenza di superare i limiti della regressione OLS nel trattare variabili predittive fortemente correlate tra loro e affette da rumore.

Questi metodi di regressione permettono di estrapolare, comprimere e sintetizzare le informazioni utili dal gran numero di misure quantitative e qualitative che un sistema di monitoraggio di un processo può registrare; non essendo possibile controllare contemporaneamente la traiettoria di tutte le variabili monitorate per captare eventuali fuori norma, diventa importante capire quale di queste variabili deve essere preferibilmente tenuta sotto controllo e quale può essere lasciata variare senza che ciò crei danni per la qualità del prodotto, quali variabili portano approssimativamente le stesse informazioni e quali si comportano in maniera analoga in risposta ad uno stesso fenomeno (*assignable cause*). Inoltre, molte variabili sono tra loro correlate e quindi la dimensione effettiva del processo sotto controllo è inferiore al numero di variabili monitorate.

Le tecniche statistiche multivariate più conosciute sono la Proiezione su Strutture Latenti (*Projection on Latent Structures*, PLS) e la Analisi delle Componenti Principali (*Principal Components Analysis*, PCA): entrambi i metodi sono stati adottati inizialmente nell'area della chimica analitica detta chemometria (Wold, 2001), per poi diffondersi nel campo dell'ingegneria chimica, soprattutto nell'ambito del controllo di processo.

I metodi di regressione multivariata vengono definiti anche come LVMR (*Latent Variable Multivariate Regression*) in quanto presuppongono che esista una struttura sottostante i dati descritta dalle variabili latenti; i metodi PLS e PCA sono strumenti per catturare questa struttura (Burnham *et al.*, 1999).

Compito dei metodi PCA e PLS è:

1. determinare la dimensione effettiva del sistema e cogliere la struttura delle correlazioni tra le variabili; questo avviene mediante la definizione delle variabili latenti (LV); le variabili latenti, in numero inferiore alle variabili originali e tra loro ortogonali, costituiscono le direzioni del nuovo iperpiano, di dimensioni ridotte rispetto alle dimensioni del processo, ma comunque in grado di descrivere i dati.

   Per capire il concetto di LV, si può fare l'esempio di una reazione all'interno di un processo continuo, dove le fonti reali di variabilità indipendenti tra loro, cioè le LV, sono la composizione dell'alimentazione, le proprietà delle materie prime e l'attività del catalizzatore, mentre tutti i vari strumenti di misura, forniscono solo diverse misure del modo in cui cambiano queste tre variabili latenti durante il processo;
2. l'estrazione delle informazioni utili dai dati disponibili e la compressione (o proiezione) di tali informazioni lungo le direzioni prestabilite dalle LV.

Rispetto ad un metodo OLS classico, i metodi PCA e PLS presentano molti vantaggi (Kresta *et al*., 1991; Kourti e MacGregor, 1995):

1. sono capaci di gestire l'inclusione nel modello di variabili altamente correlate tra loro. In questo modo si evita di dover escludere una variabile importante dal modello solo perché fortemente correlata con un'altra variabile già compresa nel modello, perdendo parte dell'informazione essenziale di un processo. Questo è un vantaggio rispetto al metodo OLS, che invece è incapace di gestire variabili tra loro correlate a causa della non invertibilità dell'inversa di una matrice singolare;
2. riducono le dimensioni del problema e permettono una semplice interpretazione grafica dei risultati;
3. manipolano matrici in cui mancano alcuni dati;
4. permettono la regressione anche con dati di variabili predittive affette da errore, cosa non possibile con la regressione ai minimi quadrati classici;
5. minimizzano il rumore presente nelle variabili latenti mediando il *rapporto segnale/rumore* proveniente dalle variabili predittive (MacGregor *et al*., 1991);
6. forniscono stime più robuste dei coefficienti di regressione parziale (Geladi e Kowalski, 1986).

Nella tesi si farà un uso esteso solo della tecnica PLS; tuttavia, si antepone all'introduzione del PLS la presentazione del PCA perché si ritiene che ciò agevoli la comprensione della tecnica PLS.

1.4.1 Analisi delle componenti principali

L'Analisi della Componenti Principali (PCA) individua, all'interno di un set di variabili, la struttura correlativa esistente tra le variabili stesse ed esplicita la dimensione effettiva dello spazio rappresentato da queste variabili mediante la costruzione delle variabili latenti.

Data la matrice **X** dei dati delle variabili di processo, di dimensioni $n \times k$, dove n sono i campioni o misure disponibili e k le variabili predittive, il metodo PCA calcola dei vettori, detti componenti principali (PC) o variabili latenti (LV), che esprimono le direzioni di massima variabilità complessiva dei dati disponibili, cioè le direzioni lungo le quali si dispongono la maggior parte delle traiettorie delle variabili. Le variabili latenti sono, per costruzione, ortogonali tra loro; la prima variabile latente indica la direzione di massima variabilità delle variabili, la seconda variabile latente indica la seconda direzione di massima variabilità ortogonale alla prima, etc.

Per ottenere le componenti principali, il metodo PCA crea due nuovi vettori per ogni PC:

1. un vettore *loading*, indicato con il simbolo **p**;
2. un vettore *score*, indicato con il simbolo **t**.

Per ottenere i vettori *loadings* e gli *scores* si scompone la matrice **X** dei dati in una serie di *k* matrici di rango 1, ognuna delle quali rappresenta una componente principale e la cui somma è la matrice **X** stessa. Ognuna di queste matrici è prodotto vettoriale dei due vettori di *loading* e *score* di ogni PC:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \ldots + \mathbf{t}_k\mathbf{p}_k^T = \mathbf{T}\mathbf{P}^T \quad , \tag{1.14}$$

Con **T** e **P** matrici degli *scores* e dei *loadings* ordinati come vettori colonna.
Come detto, l'informazione ottenuta da variabili di processo è corrotta da rumore: di conseguenza, non tutte le *k* direzioni principali saranno necessarie per rappresentare l'informazione contenuta nei dati. Se vengono mantenute le prime *A* direzioni principali o variabili latenti, la matrice **X** può essere rappresentata come nell'equazione 1.15:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \ldots + \mathbf{t}_r\mathbf{p}_r^T + \mathbf{E}_A = \mathbf{T}\mathbf{P}^T + \mathbf{E}_A \quad , \tag{1.15}$$

con $A \leq k$, ed $\mathbf{E}_A$ matrice dell'informazione residua.
Si presume che l'informazione utile sia contenuta nelle prime *A* variabili latenti, e il rumore nella matrice residua $\mathbf{E}_A$. In realtà questo raramente accade: la determinazione del numero opportuno di variabili latenti da includere in un modello PCA o PLS rimane una questione ancora irrisolta, anche se sono state presentate molti ipotesi (Valle, 1999; Hoskuldsson, 1996). In assenza di certezze, il metodo più comunemente impiegato rimane quello della *cross – validation* (Wold, 1978).
L'equazione (1.15) può essere così interpretata: il processo monitorato attraverso le variabili contenute nella matrice **X** può essere descritto attraverso *A* variabili fittizie artificiali, dette variabili latenti, le quali assicurano di rappresentare la maggior parte della variabilità utile del processo e di non essere correlate tra loro grazie all'ortogonalità dei *loadings*. Le variabili latenti costituiscono le nuove variabili del sistema e sono combinazione lineare delle variabili predittive, colonne di **X**. Le variabili latenti hanno un'importanza diversa nel modello, proporzionale alla quantità di variabilità della matrice **X** che riescono a spiegare.
Il metodo più conosciuto per ottenere le direzioni principali e le variabili latenti è la decomposizione ai valori singolari della matrice $\mathbf{X}^T\mathbf{X}$ (*Singular Value Decomposition*, SVD), strumento numerico per la determinazione degli autovalori ed autovettori di una matrice; si ricorda che $svd(\mathbf{X}^T\mathbf{X}) = \mathbf{P}\Lambda\mathbf{P}^T$, dove Λ è la matrice diagonale degli autovalori della matrice di covarianza, ordinati in modo decrescente, e **P** è la matrice contenente gli autovettori associati, coincidente con la matrice dei *loadings*.
E' possibile ottenere un'interpretazione geometrica ed analitica dei vettori di *scores* e *loadings*:

1. la matrice dei *loadings* $\mathbf{P} = [\mathbf{p}_1,\ldots, \mathbf{p}_a,\ldots, \mathbf{p}_A]$ ha dimensioni $A \times p$, ed ogni elemento p_{aj} rappresenta il contributo della j-esima variabile predittiva alla a-esima variabile latente; geometricamente, p_{aj} è il coseno direttore, compreso tra 0 e 1, che quantifica l'angolo di rotazione della a-esima variabile latente rispetto alla j-esima variabile predittiva nello spazio originario. Ogni vettore $\mathbf{p}_a$, di dimensioni $1 \times p$, definisce una nuova direzione nello spazio cartesiano in funzione delle variabili originali.
   I *loadings* sono vettori ortonormali.
   I vettori nella matrice **P** sono ordinati in modo decrescente secondo il valore del corrispondente autovalore λ_a di Λ, il quale è un indice della frazione di varianza della matrice **X** spiegata da ogni direzione principale: la prima variabile latente è associata al più grande autovalore, la seconda variabile latente al secondo autovalore, etc.
   In questo modo le prime variabili latenti dell'equazione 1.15 sono in grado di spiegare la maggior parte dell'informazione lineare delle variabili;
2. la matrice degli *scores* $\mathbf{T} = [t_1,\ldots, t_a,\ldots, t_A]$ ha dimensioni $n \times A$; gli *scores* sono il risultato della proiezione della matrice **X** lungo le direzioni principali individuate dai vettori *loadings*, secondo $\mathbf{T} = \mathbf{XP}$.

Per meglio capire il ruolo degli *scores* e dei *loadings* in un modello PCA può essere utile anche descrivere l'algoritmo NIPALS per ottenere i due vettori (Geladi and Kowalski, 1986), calcolando singolarmente e iterativamente una variabile latente alla volta:

1. prendere un vettore colonna (cioè una variabile) $\mathbf{x}_j$ di **X** e denominarlo $\mathbf{t}_a$: $\mathbf{x}_j = \mathbf{t}_a$;
2. calcolare il vettore *loading* corrispondente: $\mathbf{p}_a = \mathbf{X}\mathbf{t}_a \,/\, \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$; questo equivale a risolvere una regressione ai minimi quadrati delle colonne di **X** su $\mathbf{t}_a$ senza problemi di correlazione, perché la matrice $\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$ è in realtà uno scalare; in questo modo si cattura anche la direzione di massima variabilità della matrice **X**. La matrice **X**, per i passi successivi al primo, viene sostituita dalla matrice residua $\mathbf{E}_a$, definita al punto 6;
3. normalizzare $\mathbf{p}_a$ a norma unitaria: $\mathbf{p}_a = \mathbf{p}_a \,/\, \| \mathbf{p}_a \|$;
4. ricalcolare il vettore *score*: $\mathbf{t}_a = \mathbf{X}\mathbf{p}_a \,/\, \mathbf{p}_a^{\mathrm{T}}\mathbf{p}_a$; anche in questo caso si ha una regressione ai minimi quadrati senza problemi di inversione, perché $\mathbf{p}_a^{\mathrm{T}}\mathbf{p}_a$ è uno scalare (unitario);
5. confrontare $\mathbf{t}_a$ nel punto 2 con $\mathbf{t}_a$ al punto 4: se la differenza relativa è nulla (con uno scarto di $10^{-8} \div 10^{-6}$), passare al punto 6, altrimenti tornare al punto 2;
6. calcolare i residui della matrice **X**: $\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$; si nota che $\mathbf{E}_0 = \mathbf{X}$.

Il modello PCA fornisce anche utili indici statistici, come Q_{res} e T^2, per giudicare l'aderenza di ogni osservazione all'iperpiano costituito dalle k variabili latenti e la similarità tra uno *score* e gli altri. Informazioni su queste variabili possono essere trovate in Wise e Gallagher (1996) e in Kourti e MacGregor (1995).

1.4.2 Proiezione su strutture latenti

La proiezione su strutture latenti (PLS) è un metodo di regressione che costruisce un'equazione di regressione lineare tra gli *scores* delle variabili predittive e gli *scores* delle variabili dipendenti; analogamente al metodo PCA, il metodo PLS decompone sia la matrice delle variabili predittive **X** che la matrice delle variabili dipendenti **Y**, creando direzioni principali che descrivano la massima variabilità di **X** cogliendo allo stesso tempo la massima covarianza tra gli *scores* di **X** e **Y**.

Le matrici **X** e **Y** possono essere decomposte con due PCA secondo le relazioni esterne:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad , \tag{1.16}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \quad . \tag{1.17}$$

Se si utilizzano gli *scores* di **X** come nuove variabili predittive e gli *scores* di **Y** come nuove variabili dipendenti, è possibile ottenere una equazione di regressione interna tra gli scores. La relazione interna è:

$$\mathbf{U} = \mathbf{T}\mathbf{b}_{PLS} + \mathbf{F}^* \quad \text{e} \quad \hat{\mathbf{U}} = \mathbf{T}\mathbf{b}_{PLS} \quad , \tag{1.18a,b}$$

dove il vettore $\mathbf{b}_{PLS} = [b_1,\ldots, b_a,\ldots, b_A]^{\mathrm{T}}$ è il vettore dei coefficienti di regressione tra le variabili latenti, **F*** è la matrice degli scarti (analoga ad **e** nella regressione OLS) ed $\hat{\mathbf{U}}$ è la matrice degli *scores* di Y calcolati. $\mathbf{b}_{PLS}$ è stato ottenuto con la regressione ai minimi quadrati tra i singoli *scores* di **X** e **Y**: $b_{a,PLS} = \mathbf{u}_a^{\mathrm{T}}\mathbf{t}_a / \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$ e $\hat{\mathbf{u}}_a = b_{a,PLS}\mathbf{t}_a$.

Il modello così calcolato non è però il migliore possibile, perché non sempre gli *scores* di **X** e **Y** sono correlati tra loro; si può quindi ottenere un'equazione di regressione che rappresenta solo una minima parte della variabilità di **Y** o che non minimizza **F***.

Si è quindi cercato di decomporre simultaneamente le matrici **X** e **Y**, ruotando parzialmente gli *scores* $\mathbf{t}_a$ in modo da renderli più correlati agli *scores* $\mathbf{u}_a$. Per fare questo si sono introdotti anche i vettori *weights*, simili nel significato ai vettori *loadings*.

I *loadings* e gli *scores* sono calcolati ancora con l'algoritmo NIPALS, adattato alla regressione PLS (Geladi e Kowalski, 1986):

1. prendere un vettore colonna (cioè una variabile) $\mathbf{y}_j$ di **Y** e denominarlo $\mathbf{u}_a$: $\mathbf{y}_j = \mathbf{u}_a$;
2. calcolare il vettore *weight* di **X** corrispondente: $\mathbf{w}_a = \mathbf{X}\mathbf{u}_a / \mathbf{u}_a^{\mathrm{T}}\mathbf{u}_a$; regredendo **X** su $\mathbf{u}_a$ trovo la relazione lineare che permette di spiegare una variabile dipendente, $\mathbf{u}_a$, utilizzando l'informazione contenuta nella matrice delle variabili predittive. L'utilizzo del metodo ai minimi quadrati, inoltre, garantisce che ciò è stato fatto prendendo la direzione di massima variabilità dei dati di **X** e minimizzando la matrice residua **F***;
3. normalizzare $\mathbf{w}_a$ a norma unitaria: $\mathbf{w}_a = \mathbf{w}_a / \| \mathbf{w}_a \|$;
4. calcolare il vettore *score* di **X**: $\mathbf{t}_a = \mathbf{X}\,\mathbf{w}_a / \mathbf{w}_a^{\mathrm{T}}\,\mathbf{w}_a$;
5. calcolare il vettore *loading* di **Y**: $\mathbf{q}_a = \mathbf{Y}\mathbf{t}_a / \mathbf{t}_a^{\mathrm{T}}\,\mathbf{t}_a$;

6. normalizzare $\mathbf{q}_a$ a norma unitaria: $\mathbf{q}_a = \mathbf{q}_a / \| \mathbf{q}_a \|$;
7. ricalcolare il vettore *score* di **Y**: $\mathbf{u}_a = \mathbf{Y}\mathbf{q}_a / \mathbf{q}_a^T \mathbf{q}_a$;
8. confrontare il valore di $\mathbf{t}_a$ (4) tra un'iterazione e l'altra: se la differenza relativa è nulla (con uno scarto di $10^{-8} \div 10^{-6}$), passare al punto 9, altrimenti tornare al punto 2;
9. calcolare il vettore *loading* di **X**: $\mathbf{p}_a = \mathbf{X}\,\mathbf{t}_a / \mathbf{t}_a^T \mathbf{t}_a$;
10. normalizzare $\mathbf{p}_a$ a norma unitaria: $\mathbf{p}_a = \mathbf{p}_a / \| \mathbf{p}_a \|$;
11. calcolare il vettore *scores* di **X** ortogonale ai precedenti; $\mathbf{t}_a = \mathbf{X}\,\mathbf{p}_a / \mathbf{p}_a^T \mathbf{p}_a$;
12. ortogonalizzare il vettore *weight*: $\mathbf{w}_a = \mathbf{w}_a\,\mathbf{p}_a / \mathbf{p}_a^T \mathbf{p}_a$;
13. calcolare il coefficiente di regressione: $b_{a,PLS} = \mathbf{u}_a^T \mathbf{t}_a / \mathbf{t}_a^T \mathbf{t}_a$. La relazione interna rimane $\hat{\mathbf{u}}_a = b_{a,PLS}\mathbf{t}_a$, ma la forma di *scores* e *loadings* è cambiata;
14. aggiornare le variabili: $\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a^T$; $\mathbf{F}_a = \mathbf{F}_{a-1} - \mathbf{u}_a\mathbf{q}_a^T$; $\mathbf{F}_a^* = \mathbf{F}_{a-1}^* - \hat{\mathbf{u}}_a\mathbf{q}_a^T$.

In pratica, si combinano gli algoritmi di decomposizione PCA per le due matrici **X** e **Y** tra loro, scambiando informazioni tra gli *scores* e i *loadings* delle variabili predittive e dipendenti, di modo che sia **X** che **Y** siano modellate dalle stesse LV.
La matrice **F***, minimizzata grazie alla regressione ai minimi quadrati, esprime la differenza tra valori calcolati e sperimentali:

$$\mathbf{F}^* = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{T}\mathbf{Q}^T \quad . \tag{1.19}$$

A differenza dell'algoritmo PCA gli *scores* **t** sono combinazioni lineari delle variabili predittive di **X** pesate secondo i *weights* **w***, e non secondo i *loadings* **p**:

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* \quad . \tag{1.20}$$

In aggiunta al modello PCA, vale la relazione che lega la matrice **X** alla matrice **Y**:

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F}^* = \mathbf{X}\mathbf{W}^*\mathbf{Q}^T + \mathbf{F}^* = \mathbf{X}\mathbf{b} + \mathbf{F}^* \quad , \tag{1.21}$$

I coefficienti di regressione b_j che costituiscono il vettore **b** sono i coefficienti di regressione effettivi delle singole variabili predittive.
La differenza tra *weights* e *loadings* della matrice **X** è semplice: i primi forniscono i pesi di ogni variabile predittive a formare la variabile latente, i secondi trovano la direzione della LV e spiegano la varianza delle singole matrici; a differenza del PCA, le due caratteristiche (contributo alla LV e formazione della direzione principale) non coincidono.
Si nota inoltre che esiste una differenza tra i *weights* **w** e **w***: i primi si riferiscono alla matrice aggiornata dopo ogni iterazione (infatti sono all'interno dell'algoritmo NIPALS), mentre i secondi fanno riferimento alla matrice iniziale **X**. La relazione tra **W** e **W*** è data da: $\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}$.
Le proprietà degli *scores* e dei *loadings* sono analoghe a quanto descritto per il PCA; in aggiunta, anche i *weights* sono ortonormali.

Un modello PLS opportunamente costruito contiene variabili latenti che descrivono la varianza di **X** e **Y** con valori decrescenti e cumulativamente simili; se ad esempio, un modello spiega la maggior parte della varianza di **Y** ma poca varianza di **X**, significa che la matrice **X** contiene variabili poco correlate ad **Y** che eventualmente danneggiano la modellazione. Spesso, inoltre, accade che le variabili latenti in successione spieghino una buona proporzione di varianza di **X** e poca di **Y** e viceversa, alternativamente: la presenza di variabili predittive inutile quindi rende il modello più complesso, perché rende necessarie più variabili latenti per modellare la regressione.
La regressione condotta con il metodo PLS è rigorosamente lineare: per questo motivo, spesso vengono suggeriti metodi per eliminare il carattere dinamico dalle traiettorie, come il centramento, la normalizzazione, la trasformazione di variabili (Mejdell e Skogestad, 1991), o l'utilizzo di un *PLS* non lineare che adotta una relazione interna polinomiale o *spline* (Zamprogna *et al*., 2004). La scelta di utilizzare un modello PLS lineare semplice è motivata da precedenti analisi compiute (Zamprogna *et al*., 2004), che sottolineano la superiorità del metodo lineare rispetto a modelli non lineari.
In linea con i suggerimenti di letteratura, inoltre, (Geladi e Kowalski, 1986), i dati vengono centrati rispetto alla media di ogni variabile, per eliminare il carattere dinamico e non lineare, e normalizzati: in questo modo si conferisce uguale peso a tutte le variabili predittive nella costruzione del modello, evitando che le variabili aventi maggiori valori numerici o maggiore varianza (a causa della unità di misura o dell'apparecchiatura utilizzata) abbiano impropriamente maggior peso nel modello di regressione multivariato.

1.5 Analisi della regressione

Si presentano in questo paragrafo alcune tecniche per valutare qualsiasi modello di regressione. Si assume che le variabili abbiano distribuzione normale, anche se questo non sempre avviene.

1.5.1 Criteri di fitting

Vengono presentati, in questa sezione, alcuni metodi per valutare la validità di un'equazione di regressione.

1.5.1.1 Scarto quadratico medio (MSE)

Lo scarto o residuo è stato definito come $e_i = y_i - \hat{y}_i$. L'uguaglianza vale anche per i termini al quadrato e per tutti i punti sperimentali, di modo che:

$$SSE = \sum_{j=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{j=1}^{n} (y_i - \bar{y}_i)^2 - \sum_{j=1}^{n} (\hat{y}_i - \bar{y}_i)^2 = SST - SSR\,, \tag{1.22}$$

dove SSE è la somma dei quadrati dei residui (*Residuals Sum of Squares*), SSR la somma dei quadrati delle differenze tra valore calcolato e valore medio della variabile dipendente (*Regression Sum of Squares*), SST la somma dei quadrati delle differenze tra valore sperimentale e valore medio della variabile dipendente (*Total Sum of Squares*).
Lo scarto quadratico medio (MSE) è il rapporto tra SSE e il numero dei gradi di libertà del sistema, pari al numero di dati sperimentali meno il numero di coefficienti di regressione, compresa l'intercetta: MSE = SSE / (n - k - 1); analogamente per lo scarto quadratico medio di regressione MSR = SSR / k.
Spesso il grafico della funzione MSE rispetto al numero di variabili inserite nel modello individua un minimo, l'equazione ottima di regressione: infatti il valore di SSE diminuisce sempre aggiungendo variabili al modello (a meno che la variabile non sia palesemente inutile), perché migliora la rappresentazione dei punti, ma contemporaneamente diminuisce anche il valore al denominatore, perché il parametro k aumenta di un'unità per ogni nuova variabile aggiunta.
In presenza di rumore sulle variabili predittive, però, conduce ad *overfitting* perché tende ad includere molte più variabili del necessario per modellare anche la parte di varianza della variabile dipendente dovuta al rumore.
Inoltre, la minimizzazione di questo criterio non prescinde dalla verifica statistica dell'equazione ottenuta: spesso, infatti, un'equazione che minimizza lo scarto quadratico medio contiene coefficienti di regressione inutili perché prossimi a zero o con intervalli di confidenza troppo elevati.
In alternativa, vengono utilizzati i criteri MSEP (*Mean Square Error of Prediction*) e MSECV (*Mean Square Error of Cross-Validation*) che sono computazionalmente identici, ma calcolati su dati di convalida o attraverso la *cross – validation* dei dati di calibrazione (Wise e Gallagher, 1996).

1.5.1.2 Coefficiente di correlazione multipla (R^2)

Si definisce la statistica coefficiente di correlazione multipla R^2 il rapporto tra *SSR* e *SST*:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad , \qquad [0 < R^2 < 1] \quad . \tag{1.23}$$

Questa statistica indica la percentuale di varianza totale della variabile dipendente spiegata dalla equazione di regressione. Un'equazione è preferibile ad un'altra se ha un maggior valore del coefficiente di correlazione multipla.
Nota che se sono presenti misure ripetute di variabili dipendenti, cioè misure diverse della variabile y nelle stesse condizioni delle variabili indipendenti, è impossibile per il

coefficiente di correlazione multipla raggiungere il valore 100%, a causa della presenza dell'errore puro. In questo caso è possibile calcolare il rapporto tra il valore di varianza spiegata e il valore massimo ottenibile, in modo da ottenere un valore relativo di varianza spiegata.

Un altro inconveniente di R^2 consiste nella sua costante crescita all'aumentare del numero di parametri di regressione, indipendentemente dal fatto che l'aumento di questi provochi una complicazione, a volte inutile, dell'equazione di regressione. Se R^2 aumenta al crescere dei parametri del modello, può infatti accadere che, contemporaneamente, la somma quadratica media dei residui MSE e la varianza della variabile dipendente calcolata var($\hat{y}$) = $\sigma^2(\hat{y}) = \mathbf{X_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X_0}\sigma^2$ aumentino, rendendo il modello meno preciso ed adeguato.

Il coefficiente di determinazione multipla aggiustato è derivato da R^2 e tiene conto del numero totale di coefficienti di regressione, cioè dei gradi di libertà del sistema:

$$R_{adj}{}^2 = \frac{(SSR/k)}{(SST/(n-1))} = 1 - \frac{(SSE/(n-k-1))}{(SST/(n-1))} = 1 - (1-R^2)\frac{(n-1)}{(n-k-1)} \tag{1.24}$$

Essendo il denominatore una costante, $R_{adj}{}^2$ aumenterà solo se la nuova variabile predittiva aggiunta al modello ridurrà il valore del numeratore, il quale è composto da due termini: SSE, che generalmente diminuisce all'aumentare del numero di variabili significative nel modello, ed $n - k - 1$, che diminuisce sempre in quanto aumenta il numero k di variabili nel modello. Il rapporto tra questi due valori è una funzione il cui minimo, se esiste, indica la condizione di massimo $R_{adj}{}^2$ e ottimo del modello. L'utilizzo di $R_{adj}{}^2$ come parametro di *fitting* permette di evitare il cosiddetto *overfitting*, cioè l'aggiunta di variabili di regressione non realmente utili.

Il maggior problema con il parametro R^2 è l'incapacità di distinguere tra informazione utile e rumore; si può avere un'equazione di regressione molto precisa che minimizza lo scarto quadratico medio, ma che in realtà riproduce le stesse oscillazioni introdotte dal rumore di misura nei dati.

Se le variabili predittive vengono aggiunte una alla volta al modello di regressione calcolando di volta in volta i coefficienti di regressione, la varianza e il coefficiente di determinazione multipla, si possono stabilire due criteri di stop per l'aggiunta di parametri al modello: il primo, di tipo assoluto, blocca l'aggiunta di parametri al modello quando viene raggiunto un valore soglia di R^2, scelto dall'operatore, mentre il secondo criterio, di tipo incrementale, interrompe l'aggiunta di parametri se l'incremento relativo di R^2, cioè $(R^2_h - R^2_{h-1}) / R^2_{h-1}$, è inferiore ad un valore percentuale di soglia.

Zarzo e Ferrer (2005) suggeriscono di utilizzare come valore di soglia del secondo tipo una frazione (1%) del valore di varianza spiegata con il modello di regressione ad un solo parametro, l'intercetta.

1.5.1.3 PRESS

Si definisce il criterio PRESS (*PRediction Error Sum of Squares*) come somma dei quadrati delle differenze tra ogni osservazione y_i e il valore calcolato $\hat{y}_{(i)}$ ottenuto dalla equazione di regressione costruita su tutti i punti tranne y_i. Matematicamente:

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 \quad . \tag{1.25}$$

Il criterio di scelta indica come modello migliore quello che presenta il valore più piccolo di PRESS. Il criterio PRESS è molto simile a MSE ed a MSEP ma viene utilizzato solitamente nella *cross - validation*; infatti si ha MSECV = PRESS / (n - k - 1); versioni diverse dello stesso criterio si hanno eliminando blocchi di dati invece che un singolo valore alla volta per la fase di calibrazione (Kourti e MacGregor, 1995).
Altri criteri, qui non presentati, sono C_P di Mallows e AIC di Akaike, Q^2_{cum}, DMODX, DMODY (Gauchi e Chagnon, 2001) e RI (Lazraq *et al*., 2003).
Una discussione più ampia di questi criteri e del loro impiego nella selezione del numero ottimo di variabili predittive e del numero di LV si trova in Forina *et al*. (2004).

1.5.2 Intervalli di confidenza

Condizione necessaria per poter sviluppare intervalli di confidenza è la distribuzione normale degli errori con media nulla e varianza costante; questo equivale ad assumere che ogni punto y_i è $N(b_0 + \sum_j b_j x_{ij}, s^2)$, cioè ha una distribuzione normale con valore atteso uguale al valore calcolato e varianza uguale alla varianza dell'errore rispetto alla retta di regressione.
Ogni stima dei coefficiente di regressione ha distribuzione normale, da cui si può ottenere il corrispondente intervallo di confidenza per i coefficienti di regressione ottenuti con il metodo ai minimi quadrati classici:

$$b_j - t_{n-k-1,\alpha/2} \sqrt{s^2 \left(X^T X \right)^{-1}_{jj}} \le b_j \le b_j + t_{n-k-1,\alpha/2} \sqrt{s^2 \left(X^T X \right)^{-1}_{jj}} \quad , \tag{1.26}$$

dove il termine sotto radice rappresenta l'errore standard del coefficiente di regressione. Se l'intervallo di confidenza comprende il valore nullo, il coefficiente di regressione calcolato non ha validità statistica e l'equazione di regressione si trova in condizione di *overfitting*.
Effettuare test e analisi di tipo statistico su un'equazione di regressione ottenuta con un modello PLS è molto difficile, perché il vettore di regressione stimato con i modelli PLS è intrinsecamente non lineare ed è ottenuto operando sulla particolare matrice di covarianza tra variabili predittive e dipendenti, $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$. Un metodo certo per la determinazione di intervalli di confidenza, intervalli di predizione e test di ipotesi sulla nullità di un coefficiente di regressione non è ancora stato sviluppato per i modelli PLS.

1.5.3 Significatività statistica della regressione

La condizione necessaria per effettuare correttamente test statistici su un'equazione di regressione ai minimi quadrati è poter disporre di residui normalmente distribuiti a varianza costante.

Il test di significatività della regressione serve a determinare se esiste effettivamente una relazione lineare tra la variabile dipendente y e un sottoinsieme delle variabili predittive, ovvero verifica l'ipotesi h_0: $b_1 = \ldots = b_k = 0$ contro il rigetto dell'ipotesi stessa.

Se è vera l'ipotesi h_0, allora la funzione MSR / MSE ha distribuzione F con k ed $n - k - 1$ gradi di libertà; di conseguenza, il test è rigettato e l'equazione di regressione è plausibile se vale la disuguaglianza (1.27):

$$f_0 = \frac{(\mathrm{SSR}/k)}{[SSE/(n-k-1)]} = \frac{MSR}{MSE} > F_{k,n-k-1,1-\alpha} \quad . \tag{1.27}$$

Accettare l'ipotesi nulla non significa negare una relazione tra variabili predittive e osservate, ma indica solo che la regressione lineare non è utile nel rappresentare la variabilità dei dati analizzati; analogamente, rifiutare l'ipotesi nulla non significa che la relazione lineare è l'unica possibile, ma certifica che la relazione lineare è utile a spiegare i dati, ed eventualmente può essere accompagnata da altri termini di ordine superiore al primo.

Molto spesso l'informazione fornita da un test di un'ipotesi non è sufficiente allo sperimentatore, o a chi analizza i dati, per poter trarre delle conclusioni certe.

Infine, Draper e Smith (1998) hanno fatto notare come un'equazione statisticamente significativa possa risultare inutile, se spiega solo una minima parte della variabilità della variabile dipendente e ha un piccolo valore del coefficiente di correlazione multipla R^2. Sarebbe infatti opportuno che il valore calcolato f_0 fosse almeno 4 volte più grande del valore di controllo, anche se tale rapporto dovrebbe dipendere dalla grandezza del rapporto tra intervallo di variabilità della variabile dipendete ed errore standard.

Esiste una stretta connessione tra coefficiente di correlazione multipla, MSE e il test F di bontà della regressione: tutti questi criteri si sono però dimostrati inaffidabili in caso di correlazione tra le variabili predittive (Montgomery, 2003).

1.5.4 Significatività statistica dei singoli coefficienti di regressione.

E' importante verificare che tutte le variabili predittive presenti nel modello siano realmente utili: infatti, l'aggiunta di ogni nuova variabile al modello aumenta la varianza dei punti calcolati $\hat{y}$, e può anche aumentare il criterio MSE, se la variabile aggiunta è superflua.

In generale si può dire che un coefficiente di regressione parziale è significativo se è sufficientemente maggiore del suo errore standard, cioè se $b_j \pm \sigma(b_j)$ copre un intervallo non contenente lo zero. Ciò equivale ad effettuare un test t nei confronti dell'ipotesi h_0: $b_j = 0$, dove b_j è il coefficiente di regressione della variabile j-esima di cui si vuole misurare l'utilità nel modello di regressione.

La statistica test t verifica l'ipotesi h_0 confrontando il valore $t_0 = b_j / s(b_j)$ contro il valore soglia $t_{n-k-1,1-\alpha/2}$. L'ipotesi h_0 è rigettata se $t_0 > t_{n-k-1,1-\alpha/2}$.

Esiste un altro metodo per verificare il contributo di ogni singola variabile alla regressione; questo metodo, detto test di significatività parziale dei coefficienti di regressione (*extra sum of squares*), permette di verificare il contributo di una variabile qualsiasi nel modello come se fosse l'ultima variabile predittiva aggiunta al modello.

Si consideri il modello contenente tutte le variabili (compresa la j-esima) e se ne calcoli $SSR(\beta) = \mathbf{b}^T\mathbf{X}^T\mathbf{y}$, con $k + 1$ gradi di libertà, $MSE = (\mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}) / (n - k - 1)$; si consideri poi il modello contenente tutte le variabili esclusa la j-esima, detto modello ridotto, e se ne calcoli $SSR(\beta_r) = \mathbf{b}_r^T\mathbf{X}_r^T\mathbf{y}$, con k gradi di libertà. L'aumento di varianza spiegata, a causa del coefficiente di regressione parziale aggiunto, tenuto conto dei k già presenti nel modello, è $SSR(\beta_{k+1}|\ \beta_r) = SSR(\beta) - SSR(\beta_r)$, avente un grado di libertà; questo parametro è detto *extra sum of squares* ed è possibile confrontarlo con la varianza totale tramite un test parziale F con 1 ed $n - k - 1$ gradi di libertà; si deve cioè verificare se:

$$f_0 = SSR(\beta_{k+1}|\ \beta_r) / MSE > F_{1,n-k-1,1-\alpha} \quad . \tag{1.28}$$

Se f_0 è maggiore la variabile in ingresso al modello è significativa e deve essere mantenuta.

Il test di significatività parziale è molto utile nella costruzione di modelli (*model building*), perché può essere usato iterativamente per verificare il contributo di ogni variabile rispetto a tutte le altre o, se esiste un ordine di ingresso delle variabili indipendenti nel modello, per verificare la significatività di ogni variabile ultima entrata rispetto a quelle già entrate. Si ottiene così un sistema di test F, la cui notazione è $SS_1 = SS(b_1 \mid b_0)$, $SS_2 = SS(b_2 \mid b_1, b_0)$, $SS_3 = SS(b_3 \mid b_2, b_1, b_0)$, fino all'ultima variabile candidata all'ingresso. Il risultato del test F parziale ripetuto con $SS_2 - SS_1$, $SS_3 - SS_2$ al numeratore, costituirà un discriminante per decidere l'ammissione di una variabile al modello.

In generale, i coefficienti di regressione b e b_r sono tra loro diversi; ciò significa che le variabili indipendenti non sono tra loro ortogonali, e che quindi la significatività statistica di ogni variabile nel modello è condizionata dalle altre variabili già presenti nel modello.
La non ortogonalità delle variabili ha effetto anche sul coefficiente di multipla correlazione R^2. Il contributo di ogni variabile indipendente a spiegare la varianza della variabile dipendente è diverso a seconda della posizione di ingresso della variabile indipendente nel sistema: una variabile che da sola è capace di fornire un elevato coefficiente di correlazione multipla (cioè è una variabile che supera il test F di validità della regressione), può risultare inutile in un modello in cui sono già presenti variabili correlate con essa, in quanto il suo test F parziale risulta essere non statisticamente significativo.
In questo caso, dopo l'ingresso di ogni variabile, può essere utile eseguire il test F parziale per tutte le variabili presenti nel modello, singolarmente, come se fossero le ultime in ingresso, per verificare le interazioni relative tra ogni variabile e se la loro presenza nel modello è ancora statisticamente significativa. Questo procedimento è indicato spesso come rotazione delle variabili nel modello di regressione.
Le varianze dei coefficienti di regressione nel modello completo e ridotto sono rispettivamente $\sigma^2(b) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ e $\sigma^2(b_r) = \sigma^2(\mathbf{X}_r{}^T\mathbf{X}_r)^{-1}$: la differenza tra queste due matrici è una matrice semidefinita positiva, e questo indica che la varianza dei coefficienti di regressione del modello completo è sempre maggiore della varianza dei coefficienti di regressione del modello ridotto. Da tale risultato si può dedurre che un modello con poche variabili di regressione è preferibile, da un punto di vista statistico, ad un modello con molte variabili.
Dopo aver descritto i metodi di regressione utilizzati nella tesi ed aver esplicitato alcuni criteri che serviranno a giudicare la bontà di un'equazione di regressione, si passa ora alla descrizione dei metodi di selezione di variabili.

Capitolo 2

Metodologie per la selezione delle variabili predittive

Si introducono quattro algoritmi di selezione di variabili: *Stepwise regression* (Draper e Smith, 1998), l'algoritmo SROV di Shacham e Brauner (2003), l'algoritmo AS di Muradore *et al*. (2006) e il metodo VIP (Chong e Jun, 2005). La descrizione degli algoritmi è preceduta da una discussione circa le caratteristiche ottime di un modello di regressione e dalla descrizione dei più comuni problemi nel campo della regressione multivariata e della scelta delle variabili. Tali algoritmi saranno poi applicati in funzione della costruzione di modelli PLS.

2.1 Sviluppo del modello

La regressione PLS è negativamente influenzata dalla presenza di variabili predittive non correlate con la variabile dipendente, o da dati strutturati che esprimono regolarità non connesse a **Y**. La presenza di variabili inutili rende necessario un maggior numero di LV per modellare i dati e rende inaffidabile il criterio di *cross validation*. Anche se questo metodo di regressione non è influenzato dalla presenza di variabili tra loro correlate, diventa necessario operare una scelta sulle variabili indipendenti, in modo da selezionare quelle maggiormente esplicative della variabile dipendente e aventi un elevato rapporto segnale/rumore, in modo da rendere le stime di qualità del prodotto più robuste e il modello meno influenzato dal rischio di avarie dei sensori. Il rischio connesso a tale selezione è quello di perdere parte dell'informazione utile contenuta nelle variabili escluse.

2.1.1 Scelta delle variabili

La scelta delle variabili è contrastata tra due esigenze:

1. il desiderio di rappresentare nel modo migliore possibile i dati sperimentali, minimizzando lo scarto quadratico medio, MSE, porterebbe a introdurre nel modello il maggior numero possibile di variabili predittive; questo criterio, se vengono utilizzate variabili tra loro correlate nel modello, conduce a *overfitting*, perchè introduce eccessive variabili nel modello che spiegano solo la parte della varianza

dovuta al rumore e impedisce l'estrapolazione dei risultati dell'equazione di regressione al di fuori dell'intervallo in cui sono stati calibrati i dati;

2. la necessità di ottenere un modello il più possibile semplice e che minimizzi la varianza media del dato calcolato ($\text{var}(\hat{y}) = k\sigma^2/n$) richiederebbe di inserire il minor numero possibile di variabili predittive.

Oltre a queste esigenze, di tipo statistico, si desidera che le variabili predittive prescelte siano facilmente misurabili, ottenibili in quantità con apparecchiature poco costose e raramente soggette a fuori servizio, soggette a piccoli errori di misura e rumore.

In particolare, la letteratura indica diverse soluzioni per rispondere a quattro tipi di problemi (Hocking, 1976):

1. quale criterio si deve adottare per determinare che una equazione contenente k coefficienti di regressione, è soddisfacente dal punto di vista statistico, indipendentemente dal *fitting*; sono già stati presentati sia il test F, sia il criterio R^2, anche se si è detto che questi parametri non sempre funzionano in modo soddisfacente. Un altro riferimento è costituito dai grafici dei residui di regressione, dai grafici normalizzati e dagli intervalli di confidenza;
2. quale criterio si deve adottare per determinare la superiorità di un'equazione di regressione rispetto ad un'altra e il sottoinsieme "migliore" di variabili predittive che determina il modello. Il criterio più utilizzato rimane sempre la minimizzazione dello scarto quadratico medio di predizione (MSEP). Questo criterio, però, pone il problema della distinzione tra modellazione dell'informazione e del rumore: un'equazione può riprodurre ottimamente i dati reali, ma essere inutile perché di fatto riproduce le oscillazioni casuali introdotte dalla strumentazione di misura nei dati raccolti;
3. quale criterio utilizzare per giustificare la selezione di una variabile rispetto ai suoi competitori; questo criterio può essere lo stesso che discrimina la scelta di un'equazione di regressione rispetto ad un'altra (test F per la singola variabile), oppure un criterio diverso;
4. quale algoritmo è capace di selezionare il *set* più appropriato di variabili in modo da risultare superiore agli altri rispetto al criterio di giudizio scelto, sapendo che la selezione dell'insieme ottimo dipende anche dallo stimatore che si intende costruire: un set di variabili può essere ottimale per l'utilizzo in un modello PLS, ma inutile per un'equazione di regressione ai minimi quadrati (Muradore *et al*., 2006).

Nella Tesi si cercherà di dare una risposta al punto 4: ciò comunque comporterà la necessità di verificare la capacità di *fitting* di un'equazione e la sua validità statistica, al fine di determinare l'algoritmo di selezione più appropriato.

L'utilizzo di metodi di selezione di variabili porta spesso a risultati contrastanti con la conoscenza del processo: questo può accadere quando una variabile importante del

processo, costretta dalle condizioni di esercizio ad oscillare all'interno di un intervallo ridotto, viene considerata ininfluente dal metodo di analisi statistica o presenta un coefficiente di regressione diverso da quello atteso. Il contrasto tra variabili importanti per il processo e variabili statisticamente utili si verifica anche quando i dati raccolti delle variabili indipendenti sono influenzati da diverse qualità di rumore, più o meno intenso, più o meno casuale.
Solo un'opportuna pianificazione degli esperimenti riduce questi fenomeni in funzione dell'ottenimento di un'equazione di regressione significativa. Visto che la regressione PLS è un metodo basato sulla correlazione tra variabili, però, si deve far attenzione a non calibrare l'equazione al di fuori delle condizioni di processo in cui deve poi essere utilizzata (Kourti e MacGregor, 1995).
In generale, quindi, la scelta delle variabili non può avvenire meccanicamente e non può prescindere da una conoscenza precisa e fondata de dati che si stanno elaborando.

2.2 Problemi nella scelta del modello di regressione

Gli algoritmi di selezione sono negativamente influenzati dalla presenza di due fenomeni che riguardano spesso i *sets* di dati di processo: la multicollinearità e l'autocorrelazione.

2.2.1 Multicollinearità

Per multicollinearità si intende il fenomeno di correlazione esistente tra 2 o più variabili indipendenti, che rende inaffidabili gli algoritmi di selezione di variabili di tipo *stepwise* (Gunst e Mason, 1977*b*).
La multicollinearità può essere indotta dal sistema di controllo del processo, ad esempio con la modifica di un *set-point* di una variabile in funzione del valore di un'altra variabile misurata.
Tra i metodi per riconoscere la multicollinearità, ci sono: l'analisi della matrice di correlazione, il calcolo del VIF (*Variance Inflation Factor*), il test *F*, l'analisi dei coefficienti di regressione parziale e degli intervalli di confidenza, l'analisi del determinante della matrice di varianza-covarianza $\mathbf{X}^T\mathbf{X}$ o degli autovalori della matrice $\mathbf{X}$ o $\mathbf{X}^T\mathbf{X}$, il calcolo del numero di condizionamento, il metodo di Belsley, basato sulla decomposizione ai valori singolari (Montgomery, 1991). Tutti questi criteri hanno basi numeriche più o meno forti, ma nessuno è in grado di dare assoluta certezza di distinguere la correlazione dal semplice rumore di misura che rende due o più variabili predittive simili tra loro. Inoltre, molti di questi indicatori non possiedono un valore soglia che determini la presenza di multicollinearità, e di conseguenza non possono essere utilizzati in un algoritmo per selezionare variabili.

Il modo migliore per evitare il malcondizionamento, senza rinunciare a nessuna variabile correlata con le altre, consiste nel trasformare in modo opportuno le variabili predittive.
La normalizzazione delle variabili contribuisce a diminuire la multicollinearità tra variabili, perché riduce le differenze numeriche tra le variabili, rendendole omogenee. Infatti, l'operazione di centramento rimuove le componenti non lineare e non essenziali delle variabili ed ha un effetto benefico sul calcolo delle matrici inverse, qualsiasi metodo si usi (Gauss, QR). Inoltre, la stima dei coefficienti di regressione ottenuta con variabili sottoposte a centramento è uguale alla stima delle variabili non pre-trattate. L'unico inconveniente di questo metodo risiede nel fatto che pesa in modo equivalente l'importanza di ogni variabile nella regressione, e ciò può influenzare la selezione delle variabili (Mejdell e Skogestad, 1991).

2.2.2 Autocorrelazione

L'assunzione fondamentale della regressione lineare ai minimi quadrati riguarda la distribuzione degli scarti: questi devono avere media nulla, varianza pari alla varianza dell'errore, ed essere casualmente ed identicamente distribuiti lungo una distribuzione normale.
Ciò non ha influenza sugli algoritmi di selezione delle variabili, ma invalida tutti i test e le analisi statistiche per la convalida del modello.
Il caso più diffuso di distribuzione non casuale degli scarti riguarda le serie temporali (*time series data*), ovvero le variabili il cui andamento di sequenza naturale influenza o crea una correlazione tra altre variabili. Ad esempio, è un caso di serie temporale l'analisi di regressione tra la quantità di prodotto ottenuta da una reazione e la temperatura del reattore, senza tener conto della portata di alimentazione dei reagenti decrescente nel tempo: quest'ultima variabile, non incorporata nel modello influenza in maniera opposta le due variabili, diminuendo la portata di prodotto (perché non c'è più reagente) ed aumentando la temperatura (perché la camicia o il sistema di riscaldamento forniscono la stessa quantità di calore ad una minor quantità di reagente).
In questo caso gli scarti delle altre variabili non sono indipendenti, ma sono correlati in serie tra loro. Si dice cioè che si è verificato un caso di autocorrelazione.
Le conseguenze più gravi dell'autocorrelazione sono:

1. l'inefficienza della regressione ai minimi quadrati, che fornisce stime influenzate dei coefficienti di regressione parziale;
2. l'ottenimento di uno scarto quadratico medio che sottostima la varianza dell'errore, fornendo intervalli di confidenza troppo piccoli rispetto alla verità per i coefficienti di regressione;
3. la non validità del test *F* per misurare la validità statistica della regressione.

L'autocorrelazione può essere risolta in due soli modi: inserendo all'interno del modello la variabile che genera l'autocorrelazione, oppure utilizzando tecniche speciali di regressione (Montgomery e Peck, 1992).
Uno dei metodi più efficaci nell'individuare l'autocorrelazione è l'analisi dei residui: il fenomeno, soprattutto nel grafico che riporta i residui in funzione del tempo, si manifesta raggruppando i residui in *clusters* dello stesso segno attorno alla media.
Il metodo statistico più diffuso per individuare l'autocorrelazione è il Durbin-Watson test (Draper e Smith, 1998), limitato dal fatto che può individuare solo strutture autocorrelative del primo ordine.

2.3 Algoritmi di selezione

Vengono presentati prima i modelli dedicati alla selezione di variabili per la regressione ai minimi quadrati e basati su algoritmi iterativi (*stepwise*), poi quelli per il metodo PLS.

2.3.1 Stepwise regression

È una procedura iterativa, basata sul metodo dell'*extra sum of squares*, che costruisce in sequenza modelli di regressione aggiungendo o togliendo una variabile alla volta. Questo algoritmo può essere utilizzato per selezionare variabili rispetto ad una variabile dipendente alla volta.
Il criterio che permette di aggiungere o togliere variabili è il test *F* parziale. La variabile candidata all'ingresso nel modello ad ogni iterazione è quella avente la maggiore correlazione con la variabile dipendente e che produce, di conseguenza, il valore più alto della statistica *F*.
Partendo da un set nullo di variabili, ad ogni iterazione *h*-esima avvengono diverse operazioni:

1. si aggiunge la variabile candidata *h*–esima al modello se $f_0 > F_{in}$, dove:

$$f_0 = \max_h \frac{SSR(x_h | x_1,...,x_{h-1})}{MSE(x_1,...,x_{h-1})} = \frac{b_h}{s^2(b_h)} \quad . \tag{2.1}$$

   Ciò equivale, come detto nel § 1.5.4, alla verifica di un test *F* per la significatività del *h*-esimo coefficiente di regressione (ipotesi h_0). F_{in} si ricava dal valore soglia di ingresso stabilito dalla distribuzione di probabilità $F_{1,n-k-1,\alpha}$, al variare del parametro α; solitamente $\alpha = 0.05$;
2. si elimina la variabile inclusa dal modello dalla matrice delle variabili **X**;
3. si aggiorna la variabile dipendente sottraendo ad ogni variabile una quantità proporzionale al grado di correlazione con la variabile inserita nel modello:

$$y^{(h)} = y^{(h-1)} - b_h x_h = y - b_0 - \sum_{j=1}^{h} b_j x_j = y - \hat{y}^{(h)} \quad , \tag{2.2}$$

dove $\hat{y}^{(h)}$ è la variabile dipendente calcolata nel modello contenente h variabili di regressione, mentre $y^{(1)},\ldots y^{(h-1)}$, $y^{(h)}$, rappresentano la matrice della variabile dipendente dei dati sperimentali aggiornata dopo ogni iterazione;

4. si aggiornano le variabili indipendenti (se correlate tra loro):

$$x_j^{(h)} = x_j^{(h-1)} - a_j^{(h)} x_h^{(h)} \quad , \tag{2.3}$$

dove $x_j^{(h)}$ e $x_j^{(h-1)}$ sono la j-esima variabile indipendente aggiornata dopo l'h-esima e la (h-1)-esima iterazione, e $a_j^{(h)}$ è il coefficiente di regressione della h-esima variabile entrata nel modello rispetto alla j-esima variabile della matrice **X** originale non ancora entrata nel modello;

5. tutte le variabili già presenti nel modello vengono controllate con la statistica F per verificare se la loro presenza è ancora utile al modello mediante un test analogo a (2.1). La variabile con il più piccolo valore del test F viene confrontata con F_{out} e scartata se $F_0 < F_{out}$, dove F_{out} si ricava dal valore soglia di uscita della distribuzione statistica $F_{1,n-k-1,\alpha}$, α parametro, di solito uguale a 0.10.

La procedura continua fino a quando non ci sono più variabili che hanno i requisiti per essere inserite o escluse dal modello. Aggiungendo una variabile alla volta, è possibile ottenere, in sequenza, una stima dell'utilità della variabile h-esima come se fosse l'ultima aggiunta al modello. I parametri manipolabili dall'operatore per influenzare la scelta delle variabili sono i due valori soglia α_{in} e α_{out} che influenzano il valore di $F_{1,n-k-1,\alpha}$ da confrontare con i valori f_0 dell'*extra sum of squares* di ogni variabile in ingresso o in uscita; di solito $F_{in} > F_{out}$ per impedire che tutte le variabili entrate nel modello escano subito dopo.

2.3.1.1 Forward selection

Il metodo è simile alla *stepwise regression*, con la differenza che le variabili possono solamente entrare nel modello e non uscire. Si inizia con un modello a una variabile e si continua fintanto che nessuna variabile supera più il vincolo di ingresso $F_0 > F_{in}$. Il maggior problema, utilizzando questo metodo, consiste nel fatto che non si può verificare, ad ogni passo, se alcune variabili presenti nel modello sono diventate inutili a seguito dell'introduzione successiva di altre variabili predittive.

2.3.1.2 Backward elimination

L'algoritmo inizia con tutte le variabili predittive presenti nel modello, ed ad ogni passo elimina una variabile (quella con il più piccolo valore di correlazione con la variabile

dipendente), con il vincolo che superi il test di uscita $F_0 < F_{out}$; l'iterazione termina quando non esistono più variabili che possono essere eliminate.

2.3.1.3 Commenti sugli algoritmi stepwise

Tutti i metodi descritti dovrebbero essere sottoposti a controlli successivi, come l'analisi dei residui, l'analisi dei punti più influenti, o i test per la mancanza di adeguatezza. Inoltre, per migliorare la rappresentazione dei dati, si potrebbe prendere in considerazione la possibilità di aggiungere o aggregare più variabili indipendenti tra loro eventualmente trasformandole tramite i logaritmi o le potenze.

Non sempre i tre metodi di selezione *stepwise*, cioè *stepwise regression*, *backward elimination* e *forward selection*, forniscono lo stesso risultato allo stesso problema di regressione. Questo non significa che uno o più dei tre modelli sono errati, ma che i modelli trovano la migliore equazione rispetto alle caratteristiche con cui vengono eseguiti, e che quindi tutti i diversi set di variabili trovati rappresentano possibili equazioni di regressione ottimale, in funzione dell'utilizzo che se ne deve fare. La diversa selezione di variabili è dovuta alla correlazione presente tra variabili, cioè alla multicollinearità.

In generale, l'algoritmo di *forward selection* include più variabili nel modello, in quanto non comprende una procedura di uscita di variabili dal modello, mentre gli algoritmi *stepwise regression* e *backward elimination* consentono di ottenere risultati simili.

A volte si privilegia il metodo di *backward elimination*, perché parte dal modello completo e permette di tener conto dell'effetto sinergico di più variabili nello spiegare la varianza dei dati originari, soprattutto se le variabili sono correlate: questo effetto non può invece essere messo in luce dagli altri due algoritmi, che utilizzano come modello di partenza l'equazione contenente la sola intercetta. Tuttavia, partire dal modello completo espone la regressione al rischio di avere una matrice dei dati $\mathbf{X}$ singolare, bloccando di fatto l'algoritmo (Draper e Smith, 1998).

L'ordine con cui le variabili di regressione entrano nel modello non è un'indicazione dell'importanza della variabile nella regressione: spesso variabili entrate tra le prime nel modello con l'algoritmo *stepwise regression* escono dopo poche iterazioni rimanendo inutilizzate.

In ogni caso, i risultati di ogni algoritmo possono essere modificati semplicemente cambiando il valore dei parametri α_{in} e α_{out} del test F parziale. Non esiste concordanza sul valore da attribuire a questi parametri, e in letteratura esistono molte indicazioni discordanti tra loro.

Il motivo per cui la *stepwise regression* è molto criticata come metodo di selezione di variabili è dovuto al fatto che tale algoritmo è costruito solamente per selezionare variabili tra loro ortogonali, o al limite leggermente correlate, e non affette da errore di misura: come dimostrato da Kabe (1963), l'utilizzo di questo algoritmo con variabili correlate

porta alla sottostima dei coefficienti di regressione delle variabili scelte successivamente dall'algoritmo (in proporzione alla correlazione esistente tra le variabili) e alla distorsione dei valori della varianza dei coefficienti di regressione (Hocking, 1976): di conseguenza, il test *F* basato proprio sul rapporto tra coefficiente di regressione e varianza del coefficiente risulta completamente falsato. La quantificazione del *bias* dei coefficienti di regressione, della variazione degli intervalli di confidenza e della validità statistica dell'equazione di regressione costruita su un sottoinsieme di variabili correlate con il metodo ai minimi quadrati è riportata in Draper e Smith (1998): si può dimostrare che le equazioni di regressione ai minimi quadrati costruite su un sottoinsieme delle variabili di regressione ricavano stime *biased* dei coefficienti di regressione parziale anche se a varianza inferiore alla varianza dei coefficienti ottenuti dall'equazione di regressione del modello completo (solo se la selezione compiuta ha significatività statistica). Questo ha influenza sia sulla stima della variabile dipendente che sull'errore stesso della stima.
In generale, quindi, eliminare variabili dal modello migliora la precisione della stima dei parametri, introducendo però un *bias* nel sistema; se questo *bias* è minore della riduzione della varianza provocata dalla eliminazione di variabili, il modello ridotto può essere giudicato positivamente. Un discorso più ampio è riportato in Gunst e Mason (1977*a*).
Il *bias* è nullo solo se le variabili eliminate sono ortogonali a quelle selezionate dall'algoritmo.

2.3.2 SROV (Shacham e Brauner, 2003)

L'algoritmo di selezione di variabili SROV (*Stepwise Regression on Ortogonal Variable*) è stato ricavato dall'esigenza di ottenere un'equazione di regressione ai minimi quadrati statisticamente valida anche in presenza di errore di misura nella variabile indipendente (Shacham e Brauner, 2007).
L'interazione tra errore nella variabile predittiva, correlazione tra le variabili, trasformazione delle variabili e intervallo dei dati sperimentali influenza il tipo di equazione di regressione che si può ottenere (Shacham e Brauner, 1997), sia per il numero massimo di variabili predittive che si possono selezionare in un modello di regressione, sia per il valore degli intervalli di confidenza dei coefficienti di regressione.
Se si vuole introdurre la variabile *h*-esima nonostante il rischio di apportare ulteriore rumore al modello di regressione, si deve essere certi che l'informazione che contiene non sia già rappresentata dalle altre variabili predittive e che il rumore non modifichi la correlazione tra variabili. Ciò significa che si deve verificare che il residuo e_h ottenuto regredendo la *h*-esima variabile predittiva sulle altre variabili inserite nel modello deve essere superiore al rumore di misura ε_h contenuto nella variabile *h*-esima. Questo confronto

è effettuato attraverso la varianza dello scarto, definito come errore di troncamento (Brauner e Shacham, 1998*a*), e dell'errore di misura.
Si definisce rapporto tra errore di troncamento e rumore il coefficiente *TNR*:

$$TNR_h = \frac{\|e_h\|}{\|\varepsilon_h\|} \quad . \tag{2.4}$$

L'errore di troncamento contiene informazione utile solo se $TNR_h > 1$; in caso contrario, l'errore di troncamento è inferiore all'errore sperimentale e l'aggiunta della variabile contribuirebbe a descrivere solo la variabilità del rumore nell'equazione di regressione.
Il denominatore del criterio *TNR* è spesso solo stimato, se non si conosce l'errore di misura; inoltre, anche l'errore di troncamento non è preciso, proprio perché dipendente anche dall'errore sulle variabili predittive già inserite nel modello. Per avere la certezza che la variabile *h*-esima possa essere inserita nel modello, conviene che il criterio sia molto maggiore dell'unità.
L'unico inconveniente con questo criterio è la necessità di disporre di una stima dell'errore di ogni variabile predittiva.
Il vantaggio che deriva dall'utilizzo di questo criterio è duplice: si definisce il ruolo della precisione e del *range* dei dati disponibili nel determinare l'equazione ottima di regressione e non si utilizzano dati della variabile dipendente, che essendo anch'essa affetta da errore di misura potrebbe comportare ulteriori complicazioni (Brauner e Shacham, 1999).
Il criterio *TNR* può essere utilizzato all'interno di un algoritmo di selezione di variabili predittive, con opportuni accorgimenti.
L'algoritmo SROV inserisce iterativamente le variabili predittive al modello, tenendo conto sia dell'errore nelle variabili predittive e dipendente, sia della correlazione tra variabili, sia dei problemi numerici derivanti dalla risoluzione di un *set* di equazioni lineari (Shacham e Brauner, 1999 e Shacham e Brauner, 2003).
Affinché il modello non soffra della presenza di variabili correlate, le variabili sono selezionate dopo la loro ortogonalizzazione reciproca; per evitare di inserire variabili affette da rumore, si utilizza il criterio *TNR*; infine, per evitare problemi di inversione numerica, si utilizza il metodo QR al posto del metodo di Gauss: la matrice **Q** è ortogonalizzata con il metodo di Gram-Schmidt.
Per rendere il calcolo più agevole, inoltre, il criterio *TNR* viene semplificato, come si vede nell'equazione seguente 82.5), scritta per ogni variabile *j*-esima candidata all'ingresso nel modello al passo *h*-esimo:

$$TNR_j^{(h)} = \frac{\left\| e_j^{(h)} \right\|}{\left\| \delta x_j^{(h)} \right\|} = \frac{\left\| e_j^{(h)} \right\|}{\left\| \left(x - x_\varepsilon \right)_j^{(h)} \right\|} \quad , \tag{2.5}$$

dove la parte superiore rappresenta la varianza della variabile *j*-esima dopo il suo aggiornamento ad ogni passo; il denominatore, invece, è più difficile da calcolare perché rappresenta il rumore contenuto nel modello, che cambia ad ogni iterazione: $\delta x_j^{(h)}$ è infatti una stima del rumore presente nella variabile *j*-esima candidata all'ingresso dopo *h* iterazioni.

Il calcolo del rumore al denominatore dell'equazione (2.5) avviene con l'utilizzo di metodi di perturbazione del sistema: si aggiunge un errore casuale a distribuzione normale ai dati disponibili, avente media pari all'errore assoluto sperimentale e varianza nota, e si conduce la regressione su entrambi i dati, separatamente. La differenza tra i dati della matrice residua nei dati non perturbati (x) e perturbati (x_ε) costituisce una stima del rumore in media e varianza.

All'interno dell'algoritmo viene calcolato anche un altro criterio, con l'obbiettivo di controllare se l'informazione portata dalla *j*-esima variabile predittiva è significativa per la variabile dipendente:

$$CNR_j^{(h)} = \left(\frac{\left| \left(y^{(h)} \right)^T x_j^{(h)} \right|}{\sum_{i=1}^{n} \left(\left| x_{ij}^h \varepsilon_{iy}^{(h)} \right| + \left| y_i^h \delta x_{ij}^{(h)} \right| \right)} \right) = \left(\frac{\left| \left(y^{(h)} \right)^T x_j^{(h)} \right|}{\left| x^T \left(y - y_\varepsilon \right) \right| + \left| y^T \left(x - x_\varepsilon \right) \right|} \right) \quad . \tag{2.6}$$

Anche in questo caso viene utilizzata l'informazione contenuta nel *set* di dati perturbato per calcolare il rumore contenuto nelle variabili.

L'algoritmo è presentato nei seguenti punti, per il passo *h*-esimo:

1. prendere la variabile predittiva *k* maggiormente correlata a *y* e calcolare $TNR_k^{(h)}$ e $CNR_k^{(h)}$;
2. se i due criteri sono maggiori di 1, includere la variabile al modello; se almeno uno dei due è inferiore, eliminare la variabile predittiva e ritornare al punto 1;
3. aggiornare la matrice **X**, decomposta con il metodo QR:

$$r_j^{(h)} = \frac{\left(x_j^{(h)} \right)^T x_k^h}{\left(x_k^{(h)} \right)^T x_k^h} \quad ; \tag{2.7}$$

$$q_j^{(h+1)} = x_j^{(h+1)} = x_j^{(h)} - x_k^{(h)} r_j^{(h)} \quad ; \tag{2.8}$$

L'aggiornamento delle variabili indipendenti riguarda solo le variabili non ancora incluse nel modello, perché le altre sono già state eliminate riducendo la matrice dei dati di una colonna;

4. calcolare il coefficiente di regressione della variabile ortogonale:

$$\breve{b}_k = \breve{b}_h = \frac{\left(y_h\right)^T x_k^h}{\left(x_k^{(h)}\right)^T x_k^h} \quad ; \tag{2.9}$$

5. aggiornare la variabile dipendente e tornare al punto 1:

$$y^{(h+1)} = y^{(h)} - \breve{b}_k \, x_k^{(h)} \quad . \tag{2.10}$$

I coefficienti di regressione, alla fine della selezione, devono essere trasformati per poter essere validi anche per le variabili non ortogonalizzate:

$$b = R^{-1}\breve{b} \quad . \tag{2.11}$$

Dopo aver terminato l'algoritmo si può effettuare una ulteriore ricerca del *set* di variabili che minimizza la varianza, ruotando il *set* di variabili incluse nel modello, selezionando ogni variabile come se fosse l'ultima inclusa e confrontando questa variabile contro le altre escluse dal modello per vedere se è possibile inserire variabili alternative: infatti le variabili non sono ortogonali e quindi la scelta di ogni variabile del modello dipende anche dall'ordine con cui sono state scelte le precedenti variabili. Tale rotazione prosegue finchè tutte le variabili non vengono riselezionate come ultime in ingresso al modello.
Si nota che non esiste un criterio di stop assoluto per la selezione di variabili, ma l'inserimento termina quando nessuna variabile mostra un criterio *TNR* o *CNR* superiore all'unità. Shacham e Brauner (2007) garantiscono che l'algoritmo conduce ad una equazione statisticamente significativa senza bisogno di ulteriori controlli e verifiche.
Viene inoltre fatto notare come l'algoritmo non necessiti di criteri statistici per la regressione: in realtà, il criterio *TNR* è un rapporto tra varianze molto simile a quello utilizzato nell'*extra sum of squares* in cui il test *F* per la verifica della distribuzione è sostituito da un rapporto che è tanto maggiore quanto il numeratore ha distribuzione normale a media nulla e varianza pari alla varianza dell'errore iniziale.
Il codice MATLAB dell'algoritmo prevede diverse opzioni; le più interessanti riguardano la possibilità di pre-trattare i dati mediante diverse forme di trasformazioni di variabili, introdurre l'errore in valore assoluto o in percentuale, di inserire un'intercetta.

2.3.3 Algoritmo AS (Muradore et al., 2006)

L'algoritmo AS (Algoritmo di Selezione) di Muradore *et al*. (2006) è anch'esso di tipo *stepwise*, in quanto seleziona una variabile alla volta e calcola in modo iterativo i coefficienti di regressione.
L'algoritmo di regressione presentato da Muradore è un evoluzione dei metodi di regressione basati sugli *orthogonal descriptors* (Xu e Zhang, 2006), migliorato in quanto

l'ordine di ortogonalizzazione delle variabili è selezionato mediante il coefficiente di correlazione con la variabile predittiva.

L'obiettivo originale della selezione è costruire uno stimatore ai minimi quadrati classico che minimizzi la varianza dell'errore, cioè la matrice residua della variabile dipendente.

Rispetto all'algoritmo SROV, l'algoritmo AS è in grado di regredire più variabili dipendenti contemporaneamente. Inoltre, il criterio di stop riguarda il modello complessivo di regressione e non l'ingresso di ogni singola variabile predittiva: si utilizza un valore soglia della varianza spiegata per ammettere variabili all'interno del modello.

L'algoritmo è costituito, per l'iterazione *h*-esima in cui si seleziona la *k*-esima variabile predittiva, dai seguenti passi:

1. calcolo della correlazione tra le variabili predittive e ognuna della variabili dipendenti; si seleziona la variabile predittiva che massimizza la somma delle correlazioni con le variabili dipendenti;
2. calcolo del coefficiente di regressione della variabile predittiva selezionata con il metodo ai minimi quadrati:

$$b_k = b_h = \left(\left(x_k^{(h)}\right)^T x_k^{(h)}\right)^{-1} Y^{(h)}\left(x_k^{(h)}\right)^T \quad . \tag{2.12}$$

   Compiendo la regressione con una sola variabile predittiva alla volta si eliminano le problematiche connesse all'inversione della matrice $\mathbf{X}^T\mathbf{X}$;
3. si aggiorna la matrice delle variabili dipendenti:

$$Y^{(h)} = Y^{(h-1)} - b_k x_k^{(h)} \quad ; \tag{2.13}$$

4. si aggiorna la matrice delle variabili indipendenti, eliminando la colonna *k*-esima e calcolando la matrice residua per le altre variabili predittive:

$$x_j^{(h)} = x_j^{(h-1)} - b_k x_k^{(h)} \quad ; \tag{2.14}$$

5. criterio di stop. Si interrompe un ulteriore ciclo se:

$$R_y^2 > R_{y,soglia}^2 \quad \text{oppure} \quad \frac{\left(R_y^2\right)^k - \left(R_y^2\right)^{k-1}}{\left(R_y^2\right)^k} < \varepsilon \quad , \quad \varepsilon \propto 10^{-3} - 10^{-1} \quad . \tag{2.15}$$

Questo algoritmo è molto semplice da implementare e sicuramente conduce ad una selezione ottima di sensori, ma dovrebbe essere irrobustito con un criterio per l'ingresso delle variabili che abbia validità statistica (test *F*) o numerica (criterio *TNR*), in modo da evitare comunque l'*overfitting*. Infatti non può essere determinato il valore soglia di varianza spiegata, se prima non si conosce la quantità di varianza dipendente dall'errore stesso; il rischio è quello di inserire variabili predittive inutili che, in seguito agli aggiornamenti, diventano correlate alla variabile dipendente.

Un ulteriore problema con questo algoritmo riguarda il valore dei coefficienti di regressione ottenuti: il calcolo di ogni singolo coefficiente di regressione dipende dal numero di volte che la variabile predittive è stata aggiornata, cioè del numero di variabili già inserito nel modello, quindi non può essere considerato indipendente dagli altri coefficienti di regressione e statisticamente significativo.
I coefficienti di regressione ottenuti con questo algoritmo permettono di ottenere un'equazione di regressione che minimizza lo scarto quadratico medio, ma soffrono dello stesso problema prima descritto a proposito degli algoritmi *stepwise regression* e SROV.
Un criterio di stop alternativo alla varianza spiegata potrebbe essere:

- presa la variabile meno correlata alla variabile dipendente (o creata una variabile artificiale completamente casuale, a distribuzione normale, media nulla e varianza uguale alla varianza più grande delle variabili predittive reali), interrompere la selezione quando questa variabile viene selezionata dall'algoritmo.

2.3.4 Metodi di selezione di variabili per il modello PLS: VIP

Sono stati sviluppati diversi metodi di eliminazione di variabili implementati direttamente sull'algoritmo di NIPALS: OSC (*Orthogonal Scatter Correction*) sviluppato da Wold *et al*. (1998), O-PLS (*Orthogonal-PLS*) creato da Trygg e Wold (2002), IVS (*Interactive Variable Selection*), di Lindgren *et al*. (1994), e UVE-PLS, di *Centner et al*. (1996). Nessuno di questi, tuttavia, si è imposto rispetto agli altri o ha mostrato *performances* ottime.
Un ulteriore metodo per analizzare l'importanza di una variabile predittiva in un modello di regressione e per verificarne la correlazione con altre variabili è fornito da Lorho *et al*. (2006). Il metodo proposto parte dalla considerazione che il metodo IVS non è corretto perché il *loadings* di ogni variabile dipende dal numero di variabili presenti nel modello, in quanto deve essere $\|\mathbf{p}\| = 1$; se questo vincolo viene abbandonato, le variabili predittive possono essere valutate senza vincoli nella importanza con cui contribuiscono ad ogni LV e complessivamente, al modello. Il *loading* di correlazione tra una LV *a* e una variabile predittiva x_j si definisce:

$$r_{aj} = corr(x_j, t_a) = \frac{x_j^T t_a}{\sqrt{x_j^T x_j t_a^T t_a}} = p_{aj} \frac{\sqrt{t_a^T t_a}}{\sqrt{x_j^T x_j}} \quad . \tag{2.16}$$

I *loadings* di correlazione sono invarianti rispetto alla scala di rappresentazione e possono essere rappresentati nello stesso grafico dei *loadings* sia per le variabili di **X** che di **Y**. Il grafico contiene una circonferenza centrata all'origine, che indica l'utilizzo complessivo (100%) dell'informazione contenuta nella variabile per spiegare la singola LV del modello PLS: a seconda della vicinanza del *loadings* alla circonferenza si ha una quantificazione

del reale utilizzo della variabile e della sua vicinanza ad altre variabili predittive. Infatti la somma dei quadrati delle coordinate dei *correlation loadings* di ogni variabile predittiva indica la varianza spiegata da quella variabile predittiva, mentre la posizione nel piano indica l'effetto di quella variabile rispetto alle altre.
Un altro metodo di selezione è VIP, acronimo di *Variable Importance in Projection*, (Chong e Jun, 2005) il quale calcola un parametro di importanza per ogni variabile predittiva in base al valore pesato di ogni *loadings* nelle LV:

$$VIP_j = \sqrt{\frac{p\sum_{a=1}^{A}\left(b_a^2 t_a^T t_a \left(w_{aj} \Big/ \left\|w_{aj}\right\|\right)^2\right)}{\sum_{a=1}^{A}\left(b_a^2 t_a^T t_a\right)}} \quad . \tag{2.17}$$

Tutte le variabili aventi il criterio $VIP_j > 1$ sono ammesse al *set* ridotto.
VIP effettua sempre le stesse scelte indipendentemente dal rapporto segnale/rumore delle variabili, mentre varia se cambiano il numero di LV utilizzate dal modello.
Nel seguito, si utilizzerà, tra i metodi proposti in questo paragrafo, solo il criterio VIP per selezionare le variabili, a causa della sua semplicità di utilizzo e delle ottime *performances* ottenute.
Altri algoritmi per la selezione di variabili si trovano in Forina *et al*. (2004).

2.4 Conclusioni

L'incapacità di ottenere regressioni statisticamente accettabili con il metodo ai minimi quadrati dopo aver selezionato le variabili con un algoritmo *stepwise* è un problema al momento non risolto.
Tuttavia questi metodi di selezione possono rivelarsi utili per la selezione preliminare di variabili, nel caso non vengano poi presi in considerazione i coefficienti di regressione ottenuti con questi metodi, ma solo le variabili selezionate per la costruzione di modelli di regressione PLS. Un pregio di questi algoritmi, infatti, è quello di spazzare l'intero spazio occupato dalle variabili predittive, mediante l'aggiornamento iterativo delle variabili, che costringe l'algoritmo a selezionare ad ogni passo la variabile più correlata con la direzione ortogonale alla direzione del passo precedente. Con questo significato gli algoritmi *stepwise*, SROV e AS verranno applicati nei successivi capitoli e confrontati con il criterio VIP.

Capitolo 3

Selezione di variabili in un processo di distillazione binaria

In questo Capitolo verrà presentato un esempio di selezione di variabili con le tecniche analizzate precedentemente. I dati provengono da un impianto pilota e da simulazioni dettagliate di un processo di distillazione binaria. Lo scopo finale dell'analisi è quello di ottenere uno stimatore, basato su un modello PLS costituito da un insieme ridotto di variabili, per mezzo del quale poter predire la composizione dei prodotti della separazione.
Il principale problema è la natura intrinsecamente dinamica e non lineare dei dati a disposizione.

3.1 Introduzione

Si applicano, in questo Capitolo, i metodi di selezione delle variabili introdotti nel Capitolo precedente, insieme con i criteri per giudicare il modello costruito spiegati nel Capitolo 1.
La parte sperimentale è preceduta da una introduzione che presenta il processo studiato, l'impianto con cui i dati sono stati raccolti, e il trattamento cui sono stati sottoposti i dati prima di essere resi disponibili per la loro analisi.

3.1.1 Dati simulati e caratteristiche dello studio

Si considera di seguito il processo di distillazione continua di un sistema binario. Si fa riferimento all'impianto pilota di distillazione del Dipartimento di Principi e Impianti di Ingegneria Chimica dell'Università di Padova, di cui si riporta uno schema semplificato in Figura 3.1. La colonna si compone di 30 piatti, può essere esercita in continuo oppure in modalità *batch* e viene attualmente impiegata per la distillazione di miscele idroalcoliche. L'impianto è provvisto di un apparato di misurazione in linea che consente di acquisire, ad intervalli di tempo regolari, il valore delle temperature del fondo colonna, del distillato, dell'alimentazione e dei piatti numero 4, 8, 12, 18, 22, 26 e 30. Sono inoltre installati in

linea dei misuratori di densità per la corrente di alimentazione e per i prodotti di testa e di fondo.

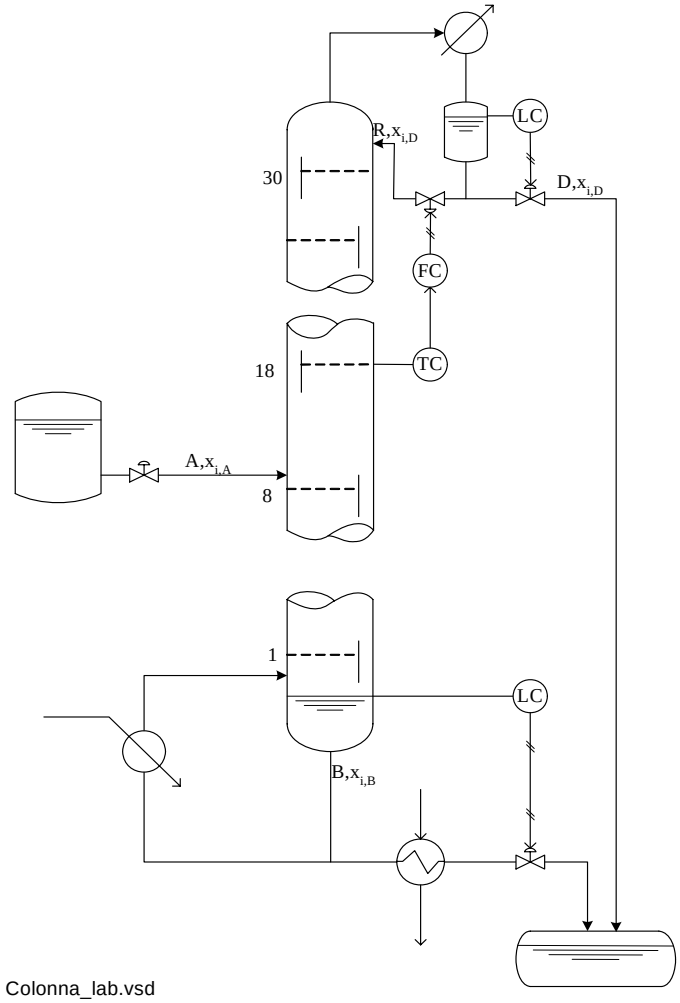


Figura 3.1. *Schema dell'impianto pilota distillazione.*

Per la colonna dell'impianto descritto è stato messo a punto un programma di simulazione, *distillazione.exe*, implementato in linguaggio *Fortran 90* (Marchi, 1990). Tale programma è utilizzato come sorgente dei dati simulati. Il motivo per cui si ricorre ad un modello della colonna, pur disponendo di un impianto reale, risiede nel fatto che le tecniche di modellazione dei dati che si intendono utilizzare necessitano di un numero di osservazioni per gli ingressi e per le risposte sufficientemente elevato da rappresentare la dinamica del processo. Nel primo caso, l'acquisizione di misure di temperatura e densità è effettuata in linea ad opera del sistema automatico di misura. Per quanto riguarda le composizioni, è difficile riuscire ad ottenere un elevato numero di misurazioni poiché i dati di concentrazione non sono forniti in linea, ma vengono determinati fuori linea mediante gascromatografia.

Sono state effettuate sei simulazioni o prove differenti di esercizio della colonna di distillazione, corrispondenti a diverse condizioni dinamiche, con l'obiettivo di osservare la risposta del processo a tipologie differenti di disturbi che possono influenzare il comportamento della colonna e la qualità delle frazioni molari di distillato e residuo. I disturbi interessano le portate di alimentazione, riflusso e vapore di fondo colonna.

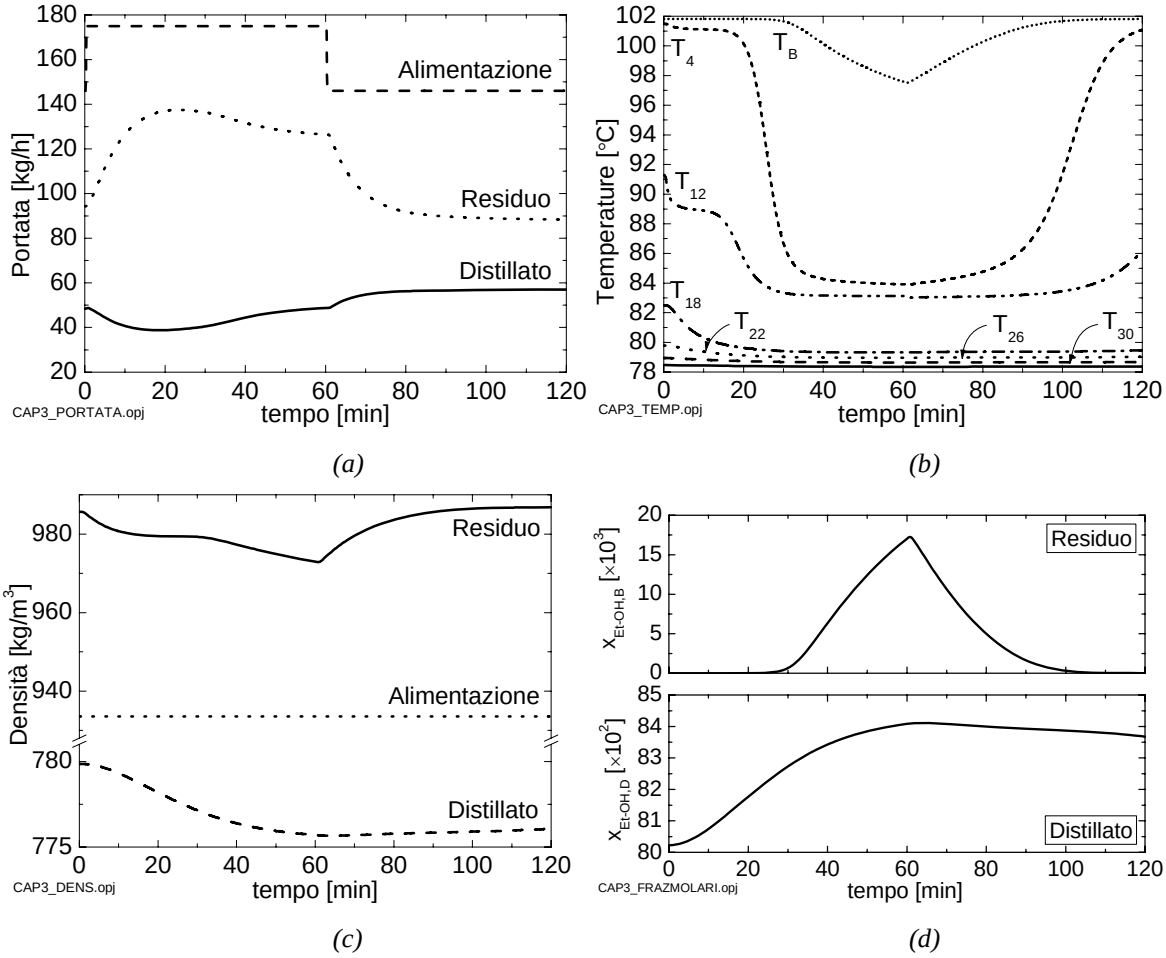


Figura 3.2. *profili simulati di (a) portate, (b) temperature, (c) densità e (d) frazioni molari di residuo e distillato a seguito del disturbo introdotto in colonna (variazione a gradino della portata di alimentazione).*

Per ogni simulazione sono state registrate 2-3 misure al minuto di ogni variabile, dall'inizio del disturbo fino al ritorno in stato stazionario della colonna; di queste misure, un campione di 200 misure selezionato casualmente costituisce il set di calibrazione: in totale, quindi, si dispone di 1200 misure di temperatura, densità e composizione per la costruzione del modello di regressione, organizzate come riportato in Zamprogna *et al.* (2005). Tutte le rimanenti misure costituiscono il set di convalida.

I dati di temperatura e densità vengono raccolti nella matrice $\mathbf{X}_{\mathrm{TD}}^{1200\times 12}$, che consiste di 1200 campioni per ciascuna delle 9 temperature misurabili (nell'ordine: T_{30}, T_{26}, T_{22}, T_{18}, T_{12}, T_4 T_{A}, T_{B}, T_{D}) e per ciascuna delle misure di densità dell'alimentazione, del prodotto di fondo

colonna e di testa. Infine, vengono selezionati 1200 campioni per la composizione di residuo e del distillato, organizzati in una matrice $\mathbf{Y}^{1200\times2}$.

Tabella 3.1. *Elenco delle variabili monitorate nel processo, associate al rispettivo simbolo.*

Variabile	Simbolo
T piatto 30	T_{30}
T piatto 26	T_{26}
T piatto 22	T_{22}
T piatto 18	T_{18}
T piatto 12	T_{12}
T piatto 4	T_4
T alimentazione	T_A
T fondo colonna	T_B
T riflusso del distillato	T_D
Densità alimentazione	ρ_A
Densità residuo	ρ_B
Densità distillato	ρ_D

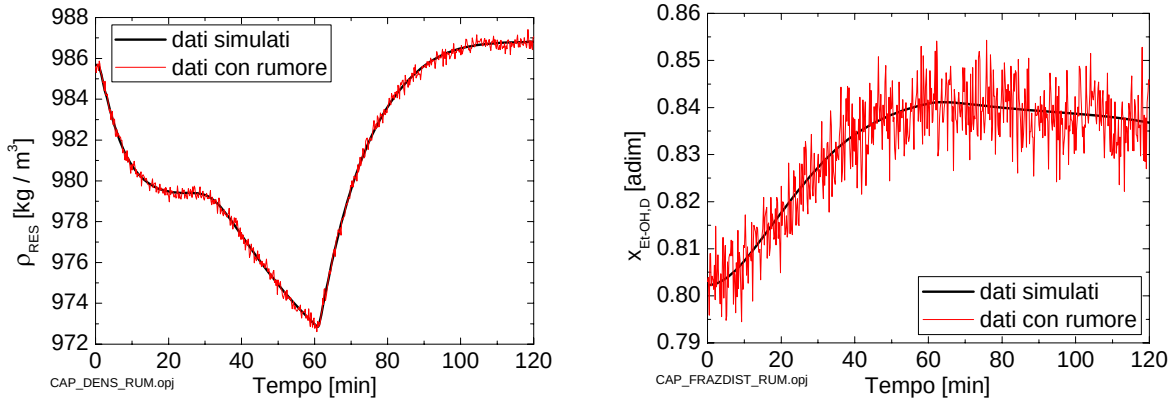
Le misure di calibrazione e convalida sono contenute nel file `DATI_CAL_CONV.mat`.
Si nota che le simulazioni hanno durate e intervalli di campionamento diversi; tuttavia, per rendere omogeneo il contributo di ogni simulazione alla costruzione del modello, si è scelto un set di calibrazione identico, per numero di campionamenti, per ogni simulazione.
Per rendere più verosimile la simulazione, i dati simulati sono stati corrotti con rumore bianco gaussiano, la cui deviazione standard è basata su prove sperimentali condotte appositamente o su dati di letteratura (Mejdell e Skogestad, 1991).

Tabella 3.2: *tipologia di rumore addizionato ai dati simulati.*

Grandezza	Deviazione standard	Errore assoluto
Temperatura dei piatti	0.07 [°C]	/
Densità delle correnti	0.225 [kg/m^3]	/
Composizione del distillato	/	2 %
Composizione del residuo	/	0.15 %

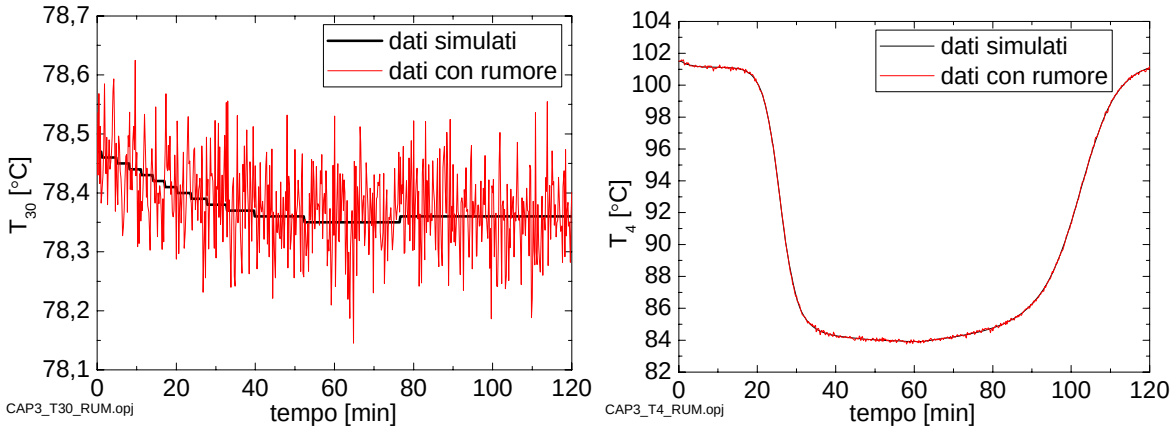
Il simulatore calcola il valore di temperatura per tutti i 30 piatti della colonna; tuttavia, di seguito verranno considerati solo i piatti corrispondenti alle posizioni per le quali sull'impianto reale è disponibile la misura. Inoltre non si utilizzeranno i dati termici relativi al piatto 8 (di alimentazione), poiché la simulazione di fenomeni di mescolamento tra correnti calde e fluidi freddi non è attendibile. A differenza di altre analisi effettuate sullo stesso set di dati (Del Bosco, 2005), si è deciso di mantenere la temperatura del piatto 30 nella simulazione, in quanto tale misura ha una duplice caratteristica:

1. è una variabile molto importante per la previsione della composizione del distillato, in quanto misura la temperatura del piatto di testa immediatamente sotto il condensatore;
2. è una variabile prevalentemente costante in tutti gli esercizi, in quanto i disturbi di portata di alimentazione e vapore vengono molto smorzati prima di giungere in testa alla colonna. Di conseguenza, il rumore aggiunto ai dati simulati rende questa misura estremamente variabile e con un rapporto segnale/rumore molto basso.



(a) *(b)*

Figura 3.3. *Profili simulati con e senza rumore di (a) densità e (b) composizione.*



(a) *(b)*

Figura 3.4. *Profilo di temperatura con e senza rumore dei piatti (a) 30 e (b) 4.*

È interessante osservare come i diversi metodi di selezione delle variabili si comportano nei confronti della temperatura del piatto 30, per capire i motivi della sua possibile inclusione od esclusione.

Per altre informazioni sui dati di simulazione utilizzati, sulle tipologie di disturbo introdotte in colonna e sul rumore addizionato ai dati si rimanda a Del Bosco (2005).

L'obiettivo dell'analisi è ottenere uno stimatore *software* che possa predire le composizioni di testa e di fondo colonna in tempo reale, utilizzando solo misure secondarie di temperatura e densità. Questo stimatore potrebbe essere applicato all'interno di un sistema

di regolazione e controllo di una colonna di distillazione, sostituendo la regolazione classica basata sul controllo della temperatura di *set point* su un piatto di riferimento che agisce a cascata sulle variabili manipolabili, portate di riflusso e di vapore.
Il modello PLS impostato costruisce uno stimatore valido sia per il calcolo della frazione molare di fondo (residuo) che per il calcolo della frazione molare di testa (distillato) utilizzando le stesse LV e gli stessi coefficienti di regressione; questo significa che il metodo di regressione ricerca la direzione di massima variabilità rispetto a entrambe le variabili dipendenti, accorpandole in un'unica nuova variabile latente. Se le variabili dipendenti, però, non variano allo stesso modo rispetto ad una stessa variabile predittiva, il modello perde la sua capacità di sintesi delle informazioni, perché dovrà forzatamente individuare una traiettoria di compromesso tra le due variabilità per la costruzione della nuova variabile latente: questo significa la perdita di ogni capacità di previsione accurata nei confronti delle frazioni molari di residuo e distillato.
Dato che i dati provengono da differenti simulazioni, si è proceduto, inizialmente, a normalizzare i dati in modo diverso, trovando un unico valore di media e di varianza per tutti i *sets* oppure calcolando singolarmente questi valori per ogni *set* di dati. Visto che la differenza tra i due metodi è minima, si sono utilizzati i dati a cui è stata applicata la normalizzazione globale dei dati.

3.2 Modello PLS a 12 variabili predittive

Prima di applicare i metodi di selezione delle variabili si procede con un'analisi complessiva dei dati a disposizione, per capire la reale dimensione latente del processo ed avere indicazioni sull'importanza relativa delle variabili predittive nella costruzione dello stimatore.
Per decidere il numero di LV ottimale per il modello si osserva l'andamento della varianza spiegata nelle matrici **X** e **Y** (R_x^2, R_y^2) e il valore dello scarto quadratico medio di calibrazione (MSEC) e di *cross - validation* (MSECV) all'aumentare del numero di variabili latenti.

Dall'analisi della varianza spiegata si ha l'indicazione ad utilizzare un modello a 2 LV, perché permette di spiegare quasi il 90% della varianza della matrice delle frazioni molari utilizzando circa il 75% della varianza dei dati di temperatura e densità. Questo risultato era atteso, perché i modelli PLS costruiti su ognuna delle 6 singole prove simulate avevano sempre indicato chiaramente che il modello a 2 LV era superiore agli altri.
In questo caso, però, la varianza totale spiegata delle variabili dipendenti, a parità di variabili latenti, è inferiore ai valori ottenuti dalle singole prove: questo risultato può essere motivato dal fatto che i diversi *sets* di variabili predittive e dipendenti uniti tra loro

spazzano zone diverse di variabilità del processo, e quindi la correlazione tra variabili dipendenti e predittive è meno forte dopo aver "unito" i dati.

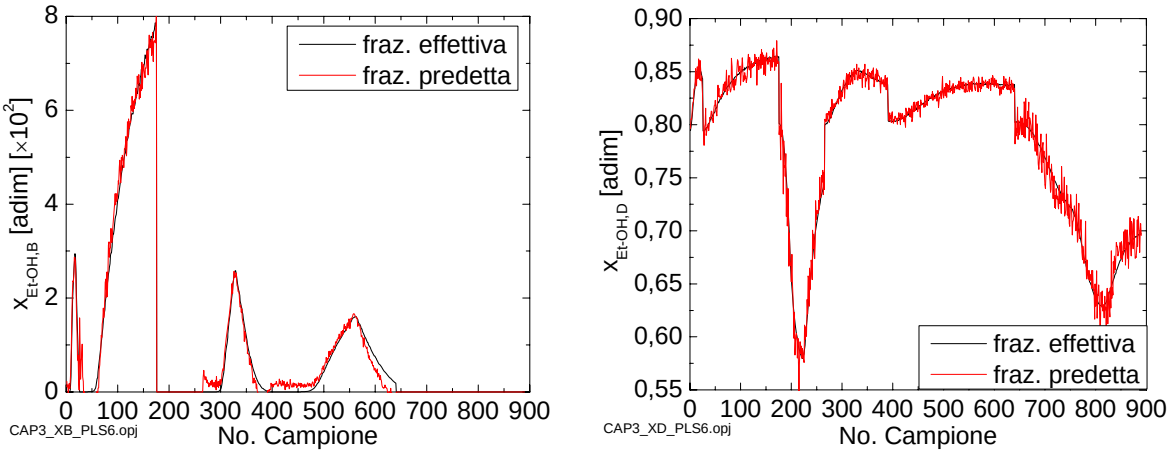
Per le prime 2 LV la varianza spiegata delle variabili dipendenti (89.31%) è molto maggiore di quella delle variabili indipendenti (74.53%), a causa della presenza, nella matrice **X**, di variabili predittive totalmente non correlate con le variabili dipendenti, ma che contribuiscono comunque alla variabilità dei dati di **X**: questo avviene perché l'algoritmo di costruzione delle LV ruota le direzioni principali della matrice **X** in modo da rendere lineare la relazione tra gli score delle matrici **X** e **Y**. In questo modo, però, una porzione di variabilità della matrice **X** contenente informazioni utili viene spiegata solo con le LV di ordine superiore, in particolare dalle LV di ordine superiore al numero di variabili dipendenti (in questo caso, dalla terza LV in poi).

La *cross - validation*, eseguita nella costruzione del modello *PLS* con il metodo dei blocchi adiacenti (*contiguous blocks*), non è in grado di indicare chiaramente il numero di variabili latenti ottimo per il modello, in quanto non individua un unico minimo, assoluto o relativo. L'assenza di un minimo, comunque, era già stato osservata e denunciata da Wold *et Al*. (2001), i quali attribuivano tale evento ad una possibile ridondanza di variabili molto correlate nella matrice **X**, alla presenza di variabili totalmente inutili nella matrice **X** nello spiegare la varianza di **Y** ed alla probabile non correlazione delle variabili dipendenti. Si individuano due minimi locali nei profili delle funzioni RMSECV e RMSEC per un numero 2 e 6 LV rispetto alla stima della frazione molare di residuo; le medesime due funzioni rispetto alla frazione molare di distillato, invece, non presentano un minimo definito né assoluto né locale, anche se per 3 LV si ottiene la riduzione massima relativa delle due funzioni. Si decide quindi di costruire cinque diversi modelli PLS, da 2 a 6 LV, contenenti tutte le variabili predittive misurate.

Prima di passare alla fase di modellazione dei PLS, si fa notare che non si è scelto un valore soglia della varianza spiegata di **Y** (R_Y^2) o di incremento della stessa per determinare il numero di LV, perché non è nota la quantità di varianza rappresentata dal rumore. All'aumentare del numero di LV non necessariamente migliora la rappresentazione del sistema: solo dopo aver stabilito la frazione di variabilità dipendente dal rumore può essere possibile creare un valore soglia di varianza che le LV devono coprire per poter costituire un modello corretto del sistema.

L'applicazione dei modelli costruiti con dati di convalida mostra con chiarezza la superiorità del modello a 6 LV sugli altri modelli; il profilo di composizione per la stima del fondo e della testa della colonna di questo modello è riportato in Figura 3.5.

La previsione ottenuta per il fondo colonna è meno rumorosa di quella ottenuta per il distillato; la rappresentazione del rumore cresce con il numero di LV nel modello.



(a) *(b)*

Figura 3.5: *Profili di concentrazione effettivo e predetto con un modello PLS a 23 variabili predittive a 6 LV del (a) fondo colonna e (b) distillato.*

E' importante notare che si ha miglioramento nella rappresentazione anche aggiungendo LV che spiegano poco più dell'1% della varianza sia di **X** che di **Y**, e ciò rende più difficile la distinzione tra variabili portatrici di informazione e di rumore.

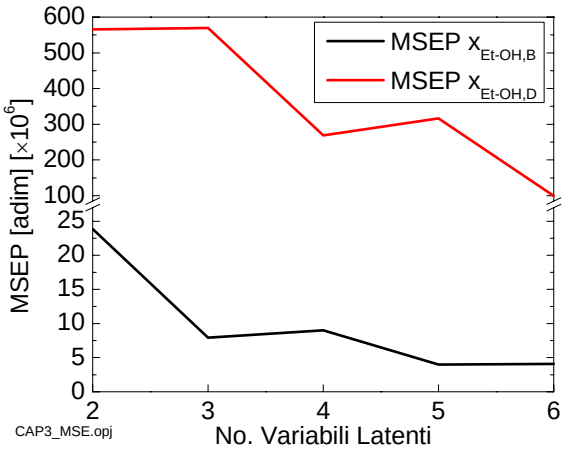


Figura 3.6. *Scarto quadratico medio di predizione calcolato con dati di convalida per le frazioni molari di residuo (nero) e distillato (rosso) al variare del numero di LV con il modello AS-PLS a 23 variabili predittive.*

Tabella 3.3. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello PLS con 23 variabili predittive.*

LV	$MSEP_B$ [×10^6]	$MSEP_D$ [×10^5]
2	23.858	56.563
3	7.9203	56.938
4	8.9881	26.917
5	3.9893	31.641
6	4.0742	9.9056

Analizzando l'andamento dello scarto quadratico medio in funzione del numero di variabili latenti, come in Figura 3.6 e Tabella 3.3, si ha l'evidenza che il metodo di costruzione delle LV segue alternativamente la modellazione prima di una e poi della seconda variabile dipendente, non potendo costruire una sola LV che univocamente rappresenti la variabilità di entrambe.

Osservando solo i grafici degli *scores* delle variabili dipendenti, cioè la proiezione dei dati di composizione lungo le direzioni indicate dei *loadings* delle sei LV, si vede come le LV rappresentino alternativamente la variabilità dei dati relativi alle due variabili dipendenti.

Il fatto più interessante è l'estrema similitudine esistente tra la prima e la quarta LV e tra la seconda e la quinta, in entrambi i casi in scala ridotta lungo l'ordinata. Si potrebbe ipotizzare che, in realtà, il modello PLS non riesce assolutamente a cogliere l'andamento temporale dei profili di composizione, tanto che, ancora alla quinta LV, dopo che cioè sono già state descritte quattro direzioni principali di variabilità dei dati, l'informazione residua è ancora in grado di rappresentare con molta precisione entrambi i profili delle variabili dipendenti, come se la varianza già spiegata dalle prime quattro LV fosse completamente inutile, o non riguardasse unicamente le traiettorie delle composizioni di testa e di fondo.

Il modello PLS costruito mostra gravi limiti, che rendono necessari ulteriori approfondimenti, volti ad accertare l'influenza del numero e del tipo di variabili predittive inserite nel modello, e a capire i possibili inconvenienti che nascono dalla calibrazione del modello rispetto a due variabili dipendenti poco correlate tra loro.

Un'altra caratteristica dei dati analizzati è la diminuzione costante del rumore nella rappresentazione della composizione del prodotto di fondo all'aumentare delle LV, mentre per il distillato avviene esattamente l'inverso, anche se accompagnato da una migliore precisione di modellazione.

L'ipotesi che si formula al riguardo è che le due variabili dipendenti siano suscettibili di rappresentazioni con gradi diversi di precisione, anche a causa del diverso intervallo di escursione: mentre la composizione di fondo varia in un intervallo massimo della frazione molare pari a 0.008 (per tre dei sei *sets* di dati l'intervallo è inferiore alla sesta cifra decimale), il prodotto di testa ha escursioni che vanno da qualità del prodotto in etanolo dello 0.55 molare fino a circa 0.87 molare. Nel caso del residuo, quindi, l'aggiunta di variabili latenti che modellino in parte anche il rumore è benefica, perché portano anche l'informazione utile ad una rappresentazione precisa fino alla quarta cifra decimale; per il distillato, invece, le nuove variabili latenti si limitano ad introdurre rumore nel modello.

Questo fatto contrasta con i benefici apportati dalla normalizzazione delle variabili: evidentemente, però, se una parte della varianza di una variabile dipendente è a priori correlata a una parte del rumore contenuto nelle variabili predittive, questa correlazione si manterrà anche dopo le operazioni di pre-trattamento dei dati.

Aggiungendo variabili latenti al modello si riesce a dimezzare l'errore assoluto per il calcolo del prodotto di fondo, mentre quello di testa rimane pressoché invariato; in nessun caso, comunque, si ottiene il risultato più importante, che è l'omogeneizzazione della dispersione dei residui (non rappresentati) in modo da rendere statisticamente significativa la regressione.

Solo i residui della frazione molare di distillato nel modello PLS a 6 variabili latenti mostrano una corretta distribuzione normale, ma ciò potrebbe non essere dovuto solo ad una corretta rappresentazione dei dati, ma anche al fatto che viene modellato sicuramente rumore, il quale aggiunge ai residui valori casuali distribuiti normalmente e a media nulla dello stesso ordine di grandezza dei residui stessi.

In definitiva, non è possibile ottenere una risposta univoca su quale sia il migliore modello PLS per la rappresentazione; sicuramente, la risposta dipende anche dalla tipologia di utilizzo cui deve essere sottoposto il modello: per una decisione più certa, comunque, sarà necessario attendere i risultati dei modelli ridotti, contenenti un numero inferiore di variabili predittive, e dei modelli separati che rappresentano con due diversi stimatori la composizione di fondo colonna e quella di distillato.

Prima di passare all'applicazione degli algoritmi di selezione si utilizza un diagramma dei *loadings* (Figura 3.7) che riporta il contributo di ogni variabile alle prime due LV, per osservare il contributo di ogni variabile al modello.

Vengono individuati due *clusters* di variabili, il primo riguardante misure del tronco di esaurimento (densità e temperatura del residuo), il secondo contenente variabili del tronco di arricchimento (T_{12}, T_{18}, T_{22}, T_{26}, T_{30} e ρ_D), separati dal piano della seconda *LV*; questa distribuzione nel piano dei *loadings* evidenzia che questi due gruppi di variabili hanno legami di tipo opposto con l'andamento delle composizioni per quanto riguarda la quota di varianza spiegata dalla seconda *LV*.

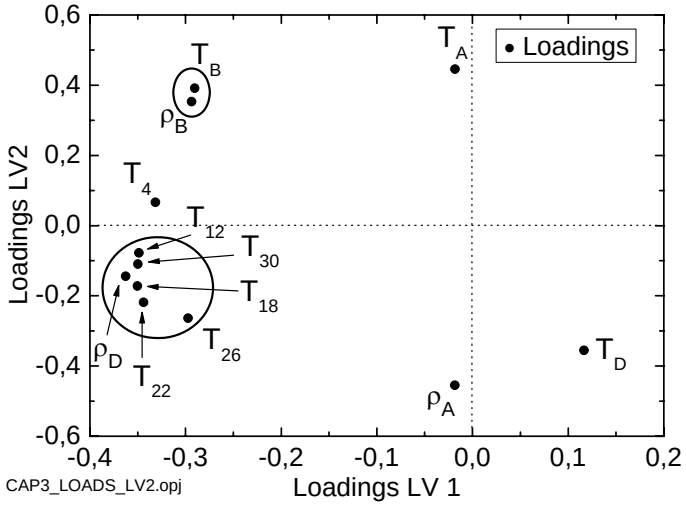


Figura 3.7. *Loadings di ogni variabile rispetto alle prime due LV.*

Leggermente scostati rispetto a questi due gruppi si trova T_4, mentre sono isolati la temperatura e la densità dell'alimentazione e la temperatura del riflusso di distillato.
Queste tre variabili sono inutili nel predire le variabili dipendenti, ma la loro posizione è anomala: ci si aspetterebbe, infatti, che i *loadings* fossero vicini all'origine e non speculari tra loro rispetto all'asse della prima variabile latente. Ciò significa che la seconda variabile latente tende a rappresentare la varianza di queste variabili: nel processo, però, queste variabili sono sempre costanti, e la loro variabilità dipende solo dal rumore di misura. Si può quindi supporre che già dalla seconda variabile latente il modello PLS rappresenti informazione inutile, perché le tre variabili T_A, ρ_A e T_D sono totalmente non correlate con i prodotti di testa e di fondo, come si osserva dalla matrice di covarianza delle variabili predittive e dipendenti in Tabella 3.4.

	T_{30}	T_{26}	T_{22}	T_{18}	T_{12}	T_4	T_A	T_B	T_D	ρ_A	ρ_B	ρ_D	x_B
T_{30}	1												
T_{26}	0.869	1											
T_{22}	0.894	0.822	1										
T_{18}	0.806	0.7	0.908	1									
T_{12}	0.734	0.611	0.797	0.921	1								
T_4	0.642	0.489	0.657	0.767	0.890	1							
T_A	0.020	-0.073	-0.114	-0.121	-0.084	-0.019	1						
T_B	0.527	0.327	0.442	0.51	0.594	0.727	0.296	1					
T_D	-0.399	-0.053	-0.183	-0.138	-0.154	-0.205	-0.314	-0.408	1				
ρ_A	-0.049	0.213	0.149	0.228	0.177	0.077	-0.594	-0.199	0.812	1			
ρ_B	0.527	0.367	0.468	0.541	0.614	0.690	0.276	0.964	-0.319	-0.105	1		
ρ_D	0.948	0.832	0.912	0.872	0.814	0.729	-0.03	0.561	-0.210	0.134	0.568	1	
x_B	-0.453	-0.284	-0.372	-0.435	-0.512	-0.631	-0.35	-0.977	0.339	0.178	-0.972	-0.493	1
x_D	-0.91	-0.835	-0.909	-0.880	-0.822	-0.732	0.074	-0.530	0.081	-0.247	-0.549	-0.991	0.470

Tabella 3.4. *Coefficienti di correlazione delle variabili dipendenti e indipendenti.*

Le variabili appartenenti al primo gruppo mostrano valori del coefficiente di correlazione multipla tutti superiori a 0.7, e la maggior parte dei valori si trova sopra il valore 0.8. Un solo coefficiente è inferiore, $corr(T_{12},T_{26})$: si ritiene che ciò sia dovuto al particolare set di dati utilizzati e non rappresenti una costante del processo.
Anche il secondo gruppo di dati mostra alta correlazione, e alle due variabili T_B e ρ_B può essere aggiunta anche T_4. Le tre variabili rimanenti, T_A, ρ_A e T_D, mostrano tutte una scarsa correlazione sia con le variabili dipendenti che con quelle indipendenti, e ci si aspetta siano scartate da qualsiasi metodo di selezione. Si nota, inoltre, che la correlazione tra temperatura e densità del distillato è negativa, e questo spiega la posizione relativa nel grafico degli *loadings*. Per inciso, il segno della correlazione è facilmente intuibile, visto che fisicamente la densità di un liquido diminuisce all'aumentare della temperatura.
Visto che le tre variabili non correlate con le variabili dipendenti hanno i valori più piccoli di *loading* rispetto alla prima LV (Figura 3.8), ci si aspetta che quest'ultima spieghi la

variazione complessiva dei dati di composizione. Si nota infatti che sia i coefficienti di correlazione che i valori di *loading* della altre variabili hanno segno negativo: esiste cioè una forte relazione lineare inversa tra i dati di composizione e quelli di temperatura e densità.

La seconda LV, invece, pur essendo distorta dai *loadings* di T_A, ρ_A e T_D, esalta le differenze tra i profili delle due composizioni, mostrando il comportamento delle variabili predittive nei confronti delle variabili dipendenti singolarmente prese. Si nota come T_4, la quale si trova nel diagramma dei *loadings* compresa tra i due gruppi di variabili, del tronco di esaurimento e di arricchimento, abbia valori simili del coefficiente di correlazione rispetto alle due variabili di composizione: questo a motivare il fatto che la seconda LV esprime le differenze tra le due variabili dipendenti.

Il coefficiente di correlazione tra le due composizioni x_B e x_D è molto basso: questo è un altro indizio che uno stimatore basato su un modello PLS che tenti di prevedere sia la frazione molare di residuo che quella di distillato con lo stesso stimatore sia inadeguato.

Si nota infine che la Figura 3.7 differisce marcatamente dalle analoghe figure costruite sugli stessi *sets* di dati singolarmente presi in Dal Bosco (2005), ad esempio in Figura 2.8.

In particolare, i valori di *loading* riferiti alla seconda variabile latente per le tre variabili inutili T_A, ρ_A e T_D, sono sempre prossimi al valore nullo sui singoli set di dati, mentre differiscono da zero nel caso in esame: questo può significare che l'importanza di queste variabili è vincolata al carattere dinamico dell'analisi, cioè al fatto che queste variabili assumono valori costanti, seppur diversi, nei *sets* di dati disponibili che vengono in questa Tesi uniti per costituire un unico blocco di calibrazione.

Si riportano ora, in Figura 3.8 e in Tabella 3.5, i valori dei *loadings* per tutte le variabili latenti. L'analisi di questo grafico mette in evidenza alcuni aspetti del modello:

1. la prima LV è costruita con i contributi di tutte le variabili realmente importanti del processo; le variabili che hanno il più basso valore di *loadings* per questa LV sono infatti T_A e ρ_A, costanti lungo tutto il processo e T_D, temperatura costante del riflusso, mentre le altre variabili contribuiscono in modo equivalente;
2. le variabili latenti da 2 a 5, che dovrebbero essere determinanti nella rappresentazione del prodotto di fondo e di testa, sono in realtà dominata dai *loadings* di T_A, T_D e ρ_A, che sono variabili per niente correlate con la frazione molare di leggero nel residuo e nel distillato;
3. nella sesta LV si ripristinano le esatte gerarchie di importanza, con T_{26} e ρ_D a fornire i valori più elevati dei *loadings*; anche in questo caso, però, il contributo di ρ_A è non trascurabile.

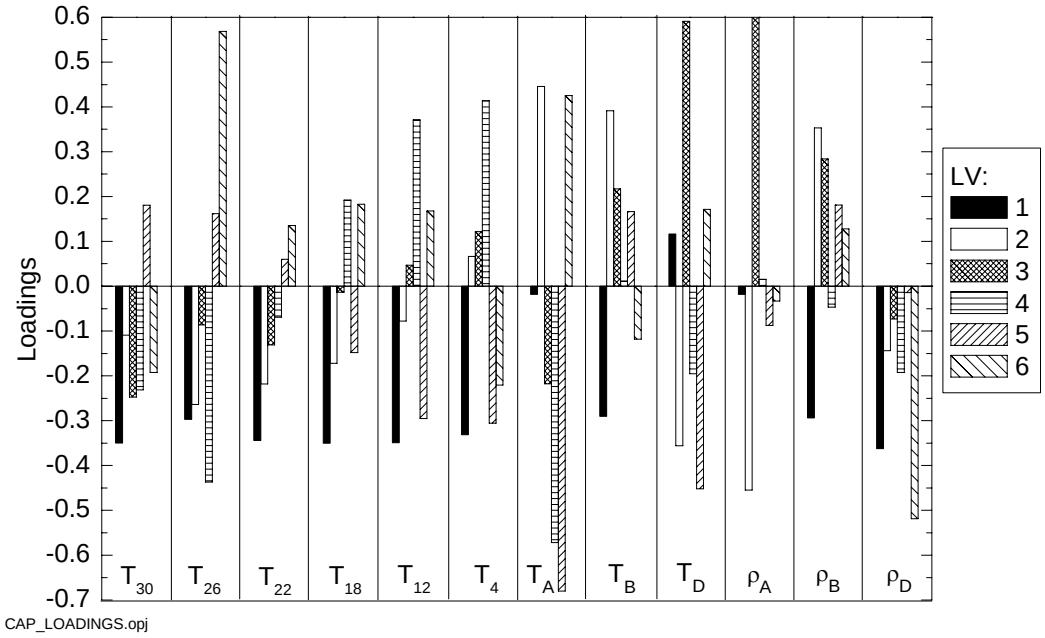


Figura 3.8. *Loadings delle variabili predittive rispetto alla prime 6 variabili latenti.*

Tabella 3.5. *Valori numerici dei loadings di ogni variabile.*

	LV1	**LV2**	**LV3**	**LV4**	**LV5**	**LV6**
T_{30}	-0.3499	-0.1093	-0.2472	-0.2316	0.1808	-0.1924
T_{26}	-0.2974	-0.264	-0.0864	-0.437	0.1617	0.5686
T_{22}	-0.344	-0.2184	-0.1307	-0.0694	0.0599	0.1352
T_{18}	-0.3504	-0.1719	-0.0135	0.1919	-0.1482	0.1831
T_{12}	-0.3489	-0.0775	0.0471	0.3711	-0.2956	0.1683
T_4	-0.3315	0.0665	0.1217	0.4135	-0.3055	-0.2208
T_A	-0.0182	0.4456	-0.218	-0.5721	-0.6801	0.4251
T_B	-0.2905	0.3914	0.2171	0.0114	0.1661	-0.1181
T_D	0.1168	-0.3557	0.591	-0.1952	-0.4519	0.1713
ρ_A	-0.0184	-0.455	0.6058	0.0152	-0.0873	-0.0333
ρ_B	-0.2938	0.3531	0.2841	-0.047	0.1811	0.1281
ρ_D	-0.3626	-0.1442	-0.0733	-0.1924	-0.0141	-0.519

Si era detto, nell'introduzione del § 2, che spesso le variabili inutili dal punto di vista del processo possono essere determinanti per la costruzione del modello di regressione; in questo caso, però, si ritiene che l'importanza di T_D, T_A e ρ_A sia artificiosa e dovuta al solo rumore di misura. Infatti, nella pratica, tali misure non mostrano nessuna variazione effettiva ma rimangono sempre costanti: la variabilità da loro spiegata, di conseguenza, non può essere che rumore di misura, il quale copre e nasconde l'informazione utile portata dalle altre variabili; qualsiasi metodo di selezione variabili dovrebbe essere in grado di eliminare queste variabili dal *subset* selezionato, pena la non validità di tale metodo.

3.2.1 Selezione di variabili con il metodo VIP

Il metodo VIP (Chong e Jun, 2005) seleziona per ogni variabile dipendente il sottoinsieme di variabili predittive più opportuno per la costruzione del modello PLS ridotto. Le variabili selezionate hanno valore del criterio VIP_j superiore a 1.

Il metodo VIP sceglie 8 delle 12 variabili iniziali per la composizione di testa, nell'ordine: ρ_B, T_B, ρ_D, T_{30}, T_{18}, T_{22}, T_{12}, T_4. Le variabili scelte per predire la composizione del fondo colonna sono soltanto tre: T_B, ρ_B, ρ_D.

La bontà della selezione è palese dall'esclusione delle variabili già considerate inutili, cioè T_A, ρ_A, T_D.

Come nel paragrafo precedente, si procede con la presentazione e il confronto simultaneo di tutti i modelli PLS aventi diverso numero di variabili latenti, sulla bare dei criteri MSECV e MSEC, R_x^2 ed R_y^2.

Si possono individuare due possibili modelli PLS, con 2 e 5 LV. Anche in questo caso la *cross validation* non individua un minimo assoluto da cui poter dedurre il numero di LV utili per la costruzione del modello PLS, ma gli andamenti dei criteri MSECV e MSEC sono più facilmente interpretabili rispetto ai modelli PLS contenenti tutte le variabili predittive.

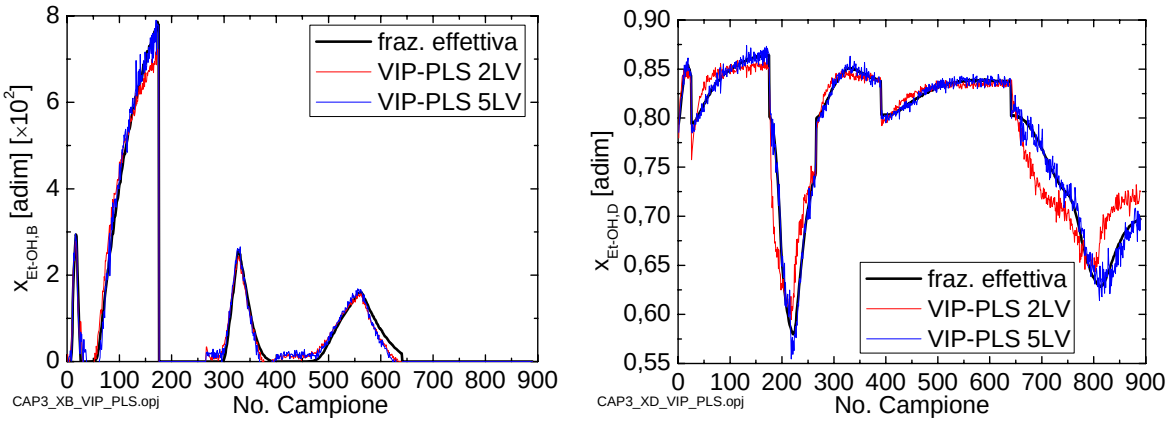
La selezione delle variabili ha l'effetto di incrementare il valore di varianza spiegata totale di **X** nelle prime variabili latenti: ciò significa che le variabili eliminate erano quelle che meno rappresentavano l'andamento delle variabili dipendenti.

Dopo la selezione, la quantità di varianza totale di **X** e **Y** corrisponde nelle diverse variabili latenti. In particolare, con 2 variabili latenti si spiega il 90.01% della matrice **Y** e il 93.57% della matrice **X**.

Come nell'analisi effettuata con il modello PLS contenente tutte le variabili predittive, anche in questo caso l'indicazione per il modello migliore è di 2 variabili latenti.

Il confronto tra i profili di distillato e di residuo effettivo e predetto si ha nella Figura 3.9. Il modello a 5 variabili latenti è sicuramente più preciso di quello a 2 variabili latenti, anche se la rappresentazione è più affetta da rumore; molto probabilmente il modello PLS a 2 variabili latenti serve a descrivere la maggior parte della varianza di entrambe le variabili dipendenti, mentre il modello a 5 variabili latenti, pur descrivendo parte del rumore di misura, aiuta a migliorare la rappresentazione del prodotto di fondo, che come si è visto necessita di un numero maggiore di variabili latenti per poter essere descritto.

Non si nota nessun miglioramento palese utilizzando un modello PLS costruito su un *set* ridotto di variabili rispetto al modello PLS costruito sull'intero *set* di variabili predittive, a parità di numero di variabili latenti. Questo è un fatto importante, perché potrebbe indicare che è inutile ridurre il *set* di variabili predittive.



(a) *(b)*

Figura 3.9. *Profili di concentrazione effettivo e predetto con un modello PLS a 2 LV e con un modello PLS a 5 LV del (a) fondo colonna e (b) distillato.*

Lo scarto quadratico medio calcolato utilizzando dati di convalida (MSEP) può essere utile per compiere ulteriori confronti tra gli stimatori e viene riportato in Tabella 3.6.

Tabella 3.6. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello VIP-PLS a 2 e 5 variabili latenti.*

LV	$MSEP_B$ [$\times 10^6$]	$MSEP_D$ [$\times 10^5$]
2	4.7311	57.782
5	3.8477	5.7644

Il miglioramento tra i due modelli “ridotti”, definiti modelli VIP-PLS, che graficamente è molto evidente, non è invece palese dall’analisi dello scarto quadratico medio, almeno per la stima della composizione di fondo, a causa dell’eccessivo oscillare delle variabili dipendenti calcolate.

Dal confronto dello scarto quadratico medio dei modelli PLS costruiti con il set completo e ridotto di variabili predittive (Tabelle 3.3 e 3.6), si ha però la conferma che i modelli VIP-PLS sono nettamente superiori ai modelli PLS: in particolare, il modello VIP-PLS a 2 LV mostra una riduzione di un ordine di grandezza del criterio MSEP per la previsione della composizione di fondo, mentre il modello VIP-PLS a 5 LV mostra un’analoga riduzione per la composizione del distillato. Si potrebbe proporre di utilizzare il modello a 2 LV per la stima del prodotto di fondo, e il modello PLS a 5 LV per la stima del prodotto di testa; da questo punto di vista, però, la conclusione più opportuna rimane quella della costruzione di 2 modelli PLS separati, per la testa e il fondo colonna.

Un altro vantaggio ottenuto nell’utilizzo di modelli PLS calibrati su un set ridotto di variabili latenti è la diminuzione del rumore, come si può controllare confrontando le

Figure 3.5 e 3.9. Ciò significa che la variabile T_{30} non è la sola responsabile della maggiore oscillazione del profilo predetto rispetto a quello effettivo.

I primi due profili degli *scores* del modello VIP-PLS sono identici ai profili delle prime due LV del modello *PLS*: questo era un risultato atteso per la prima LV, mentre è inaspettato per la seconda, perché nel modello PLS la seconda LV riceveva i contributi maggiori, in termini di *loadings*, proprio dalle variabili escluse dal modello VIP-PLS. Questo significa che, nonostante la presenza di variabili di processo non importanti, la regressione PLS è comunque in grado di estrarre solo l'informazione utile, prendendola dove ne trova in maggiore quantità.

Si può affermare, vedendo che i modelli migliori si hanno con 6 LV nel modello completo e con 5 LV nel modello ridotto, che la riduzione del numero di variabili predittive comporta un risparmio di una LV in termini di costruzione del modello PLS, senza perdere in capacità rappresentativa. Questo risultato era ancora più evidente nelle analisi, qui non riportate, che studiavano i singoli set di dati separatamente: la riduzione del numero di variabili predittive comportava il risparmio di 2 o 3 LV su ogni set di dati, a parità di capacità di modellazione. Evidentemente, la presenza di più set di dati nel modello, ognuno con differenti caratteristiche, rende più difficoltosa la riduzione del numero di LV, perché i dati mostrano una maggiore varianza.

Per capire il miglioramento tra la situazione prima e dopo la selezione di variabili conviene osservare il diagramma degli *scores* delle variabili predittive, proiettate rispetto alle prime due LV, in Figura 3.10. I punti appartenenti a diversi set di dati sono molto più vicini e compatti tra loro nel modello VIP-PLS: inoltre, la Figura 3.10 mostra lo stesso andamento ordinato di dati del diagramma degli scores delle variabili dipendenti, qui non riportato: la relazione interna che lega questi *scores*, quindi, sarà più precisa nel modello ridotto, dove i dati sono accorpati, piuttosto che nel modello completo.

Inoltre, si nota che la Figura 3.10 non può essere utilizzata per costruire intervalli di confidenza, perché rappresenta un processo dinamico: l'utilità di questo grafico è quella di mostrare come, in seguito alla creazione delle variabili latenti, sia possibile individuare un andamento lineare delle traiettorie delle variabili predittive e dipendenti esprimibile attraverso un'equazione di regressione lineare.

L'analisi dei residui mostra che:

1. i residui hanno una minore varianza rispetto ai modelli completi;
2. non vengono eliminati i *trends* per ogni *set* di dati, con l'eccezione della composizione di distillato del modello PLS a 5 *LV*; i modelli costruiti sono cioè non statisticamente significativi.

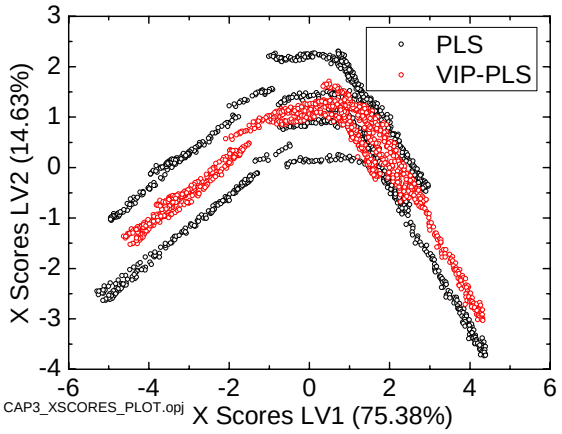


Figura 3.10: *diagramma degli scores delle variabili predittive; dopo l'eliminazione delle variabili inutili, i sets di dati mostrano profili più simili tra loro.*

Anche in questo caso la regressione PLS non ha colto completamente la variabilità dei dati, nonostante spieghi sempre più del 90% della varianza sia delle variabili predittive che dipendenti; l'unico miglioramento, rispetto al modello PLS "completo", si ha nella riduzione del numero di LV (da 6 a 5) per la rappresentazione del profilo di distillato, come già detto.

In conclusione, si analizzano i valori dei *loadings* per i modelli PLS a 2 e 5 variabili latenti in Figura 3.11 e Tabella 3.7.

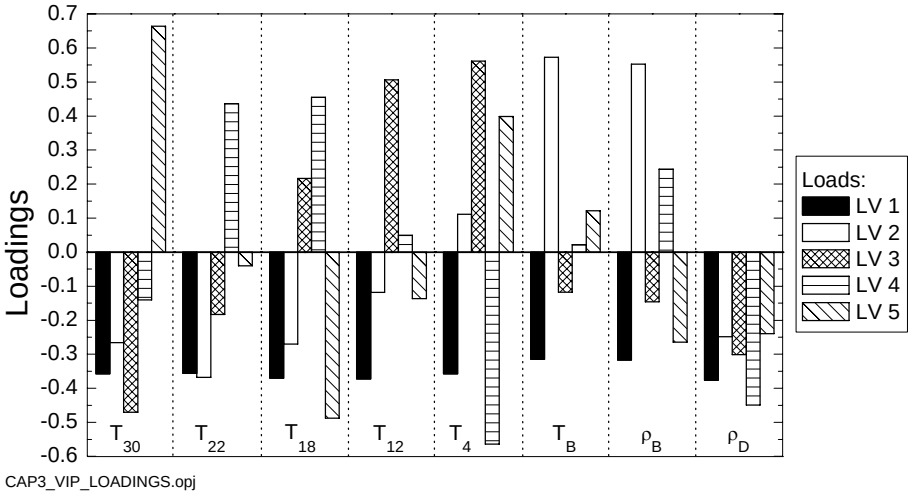


Figura 3.11. *Loadings delle 5 variabili latenti del modello VIP-PLS.*

Tabella 3.7. *Valori numerici dei loadings del modello VIP-PLS.*

	LV1	**LV2**	**LV3**	**LV4**	**LV5**
T_{30}	-0.3574	-0.2661	-0.4706	-0.1401	0.6637
T_{22}	-0.3562	-0.3680	-0.1830	0.4367	-0.0402
T_{18}	-0.3704	-0.2705	0.2171	0.4551	-0.4878
T_{12}	-0.3727	-0.1181	0.5067	0.0493	-0.1362
T_4	-0.3580	0.1114	0.5617	-0.5640	-0.3991
T_B	-0.3145	0.5724	-0.1175	0.0216	0.1219
ρ_B	-0.3172	0.5522	-0.1459	0.2439	-0.2642
ρ_D	-0.3764	-0.2486	-0.3011	-0.4495	-0.2395

I valori dei *loadings* di ogni variabile hanno un'interpretazione più facile rispetto al modello PLS contenente tutte le variabili predittive:

1. la prima LV descrive la variabilità generica di entrambe le variabili dipendenti; tutte le variabili predittive contribuiscono equamente alla costruzione di questa variabile, indipendentemente dalla loro appartenenza al tronco di arricchimento od esaurimento.
2. la seconda LV è caratterizzata da alti valori di *loadings* per le variabili T_B e ρ_B; ciò significa che questa LV migliora la rappresentazione del prodotto di fondo colonna. Si nota inoltre che i *loadings* del tronco di arricchimento sono tutti negativi, mentre quelli del tronco di esaurimento sono positivi: si divide così l'informazione di pertinenza della composizione di distillato da quella riferita alla composizione del residuo;
3. le ultime tre LV catturano la variabilità che distingue il profilo di distillato da quello di fondo colonna; in particolare, l'ultima LV riceve il maggior contributo da T_{30}: eesendo quest'ultima una variabile gravemente affetta da rumore, si ritiene che la quinta LV aggiunga rumore al modello.

3.2.2 Selezione delle variabili con il metodo Stepwise regression

L'algoritmo stepwise regression (Draper e Smith, 1998) seleziona un *set* di variabili nei confronti di una sola variabile dipendente alla volta. L'algoritmo è stato applicato quindi due volte prima con la frazione molare di residuo e poi con la frazione molare di distillato; nel primo caso sono state selezionate nell'ordine le variabili T_B, T_4, ρ_B, T_D, T_A, T_{18}, mentre nel secondo caso sono state selezionate tutte le variabili, tranne la densità dell'alimentazione. Dalla conoscenza del processo si può dedurre che il metodo *stepwise regression* non è utile per la selezione di variabili predittive da inserire in un modello *PLS*. Per cercare comunque di implementare la selezione in modo da ottenere un sottoinsieme di variabili predittive utile, si modificano la soglia di ingresso e di uscita per il test *F*, ovvero i coefficienti α_{in} e α_{out}; tuttavia la selezione rimane inalterata.

Si è provato anche selezionare le variabili dopo averle pre-trattate mediante centramento e normalizzazione, ma il risultato è rimasto inalterato. Si è quindi abbandonato il metodo di selezione *stepwise regression* per la costruzione di un insieme ottimo di variabili per la stima della composizione dei prodotti della distillazione.

3.2.3 Selezione delle variabili con l'algoritmo AS

L'algoritmo AS viene applicato facendo selezionare 8 delle 12 variabili predittive iniziali, ovvero nell'ordine ρ_B, ρ_D, ρ_A, T_D, T_B, T_{12}, T_{22}, T_4; la scelta appare sensata, anche se può preoccupare l'inserimento di ρ_A e T_D nel set ridotto di variabili. Si ricorda che l'algoritmo prevede la scelta del numero di variabili candidate ad entrare nel set ridotto: la scelta è stata di 8 in analogia con il metodo VIP e perché sembra assicurare stabilità al modello, in quanto contiene equamente variabili appartenenti al tronco di arricchimento ed esaurimento.

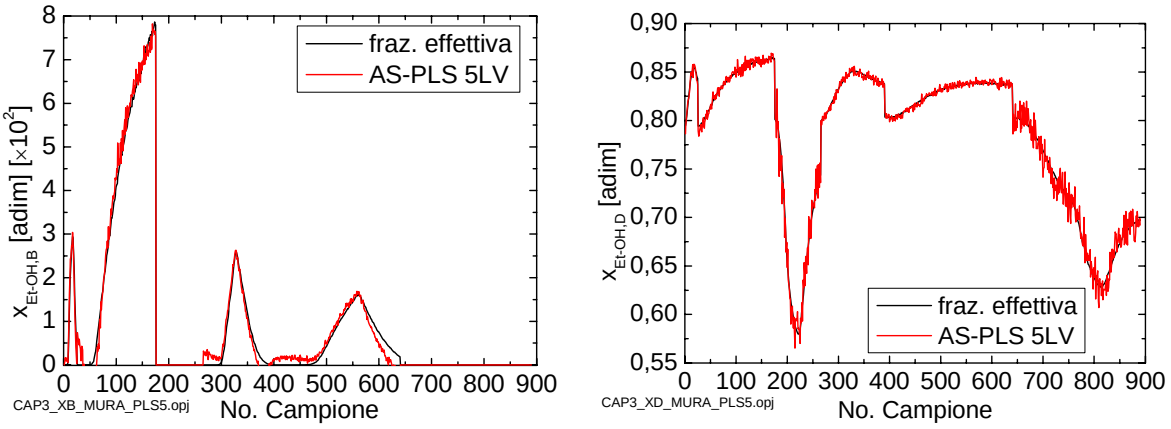
Va notato il fatto che viene esclusa la variabile T_{30}, probabilmente perché l'informazione residua contenuta in questa variabile dopo l'inserimento di ρ_D nel set ridotto è minima; infatti ρ_D e T_{30} sono molto correlate tra loro e l'inclusione di una di queste variabili comporta l'esclusione automatica dell'altra. Ciò è dovuto al fatto che l'algoritmo è di tipo *stepwise*, ed aggiorna le variabili dopo ogni iterazione.

Si procede con la ricerca del modello PLS più idoneo in base alla *cross-validation* e alla varianza spiegata (R_x^2 ed R_y^2). Vengono considerati tre possibili modelli, a 2, 3 e 5 variabili latenti: il modello a 2 LV è suggerito sia dalla varianza spiegata di $\mathbf{Y}$ che da un minimo locale del criterio MSECV della frazione molare di distillato, mentre i modelli a 3 e 5 LV sono indicati dal minimo assoluto dello stesso criterio rispetto alle frazioni molare di residuo e distillato, rispettivamente. Si preferisce comunque trattare anche il modello a 4 LV, per avere un quadro complessivo della situazione.

Si confrontano, in figura 3.12, i profili di distillato e residuo effettivi e predetti del modello PLS che minimizza il criterio MSEP, cioè quello a 5 LV.

Si riportano, in Figura 3.13 e nella Tabella 3.8, l'andamento del criterio MSEP in funzione del numero di variabili latenti, e i corrispondenti valori numerici; si vede come sia inutile utilizzare più di 3 variabili latenti per modellare il profilo di composizione del residuo.

Il modello a 2 variabili latenti è leggermente peggiore del modello PLS a 23 variabili predittive, mentre passando a 5 LV si vede l'effetto della selezione delle variabili: il modello AS-PLS è superiore al modello PLS soprattutto per la stima della composizione di testa, dove mostra un dimezzamento del criterio MSEP per 5 variabili latenti. Sia la stima del distillato che del residuo sono buoni quanto la stima ottenuta con il modello VIP-PLS.



(a) *(b)*

Figura 3.12. *Profili di concentrazione effettivo e predetto con un modello PLS a 5 LV del (a) fondo colonna e (b) distillato.*

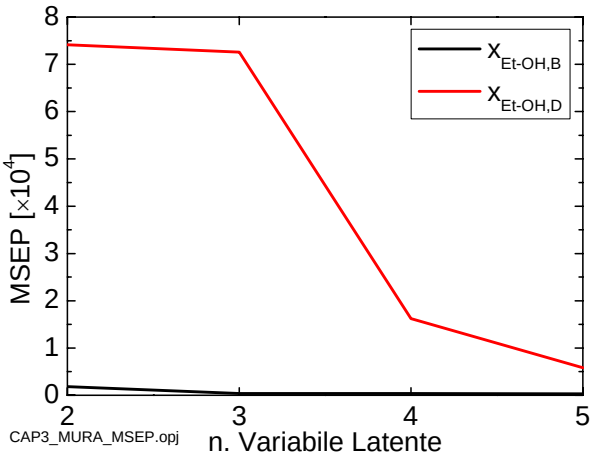


Figura 3.13. *Scarto quadratico medio di predizione per le frazioni molari di residuo (nero) e distillato (rosso) al variare del numero di LV con il modello AS-PLS.*

Tabella 3.8. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello AS-PLS.*

LV	$MSEP_B$ [×10^6]	$MSEP_D$ [×10^5]
2	18.241	74.141
3	3.8886	72.605
4	3.8319	16.258
5	3.4403	5.8368

A differenza di quanto visto in Figura 3.10, però, il grafico degli *scores* rispetto alle prime due variabili latenti delle variabili predittive per il modello AS-PLS non differisce rispetto al grafico per il modello PLS a 23 variabili predittive. Gli *scores* non sono accorpati come nel modello VIP-PLS: questo significa che non viene colta completamente la struttura

correlativa delle variabili. Ciò non ha effetto sulle capacità predittive del modello AS-PLS, ma potrebbe indicare che la selezione non è stata accurata come nel modello VIP-PLS.
Anche in questo caso i residui non sono statisticamente significativi, anche se la distribuzione dei dati residui ha minore varianza rispetto al modello PLS a 23 variabili predittive ed è confrontabile con i residui ottenuti dal modello VIP-PLS.
In conclusione, si analizzano i valori dei *loadings* in Figura 3.14 e Tabella 3.9.

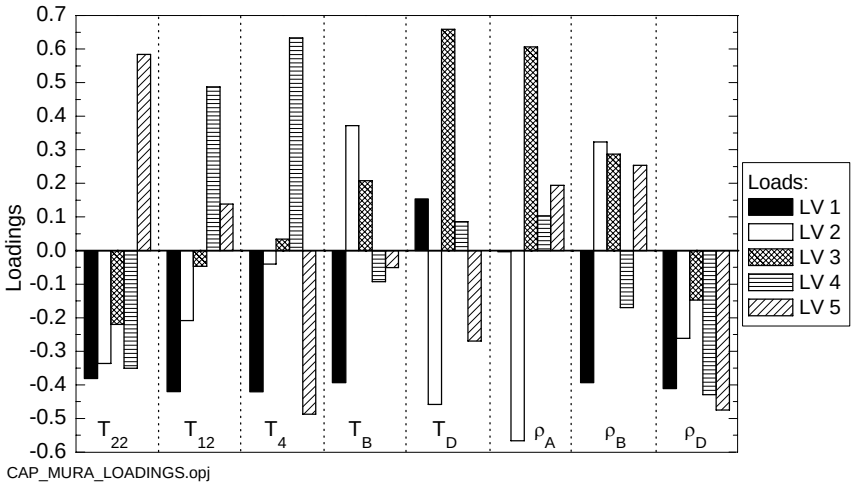


Figura 3.14. *Loadings delle 5 variabili latenti del modello AS-PLS.*

Si può notare che:

1. la seconda e la terza variabile latente sono dominate dal contributo delle due variabili predittive inutili T_D e ρ_A, ma questo non sembra influire sulle capacità predittive del modello;
2. la quarta e quinta variabile latente sono caratterizzati da alti valori dei *loadings* delle variabili utili, ed infatti i modelli PLS corrispondenti mostrano un grande miglioramento del criterio MSEP per la stima della frazione molare di distillato.

Tabella 3.9. *Valori numerici dei loadings del modello AS-PLS.*

	LV1	**LV2**	**LV3**	**LV4**	**LV5**
T_{22}	-0.3813	-0.3362	-0.2198	-0.3503	0.5838
T_{12}	-0.4196	-0.2090	-0.0470	0.4873	-0.1388
T_4	-0.4205	-0.0403	0.0337	0.6329	-0.4871
T_B	-0.3933	0.3723	0.2081	-0.0923	-0.0509
T_D	0.1536	-0.4582	0.6585	0.0858	-0.2698
ρ_A	-0.0031	-0.5661	0.6058	0.1032	0.1940
ρ_B	-0.3931	0.3233	0.2875	-0.1698	0.2536
ρ_D	-0.4109	-0.2612	-0.1474	-0.4290	-0.4747

Rispetto al modello PLS a 23 variabili predittive, nessun modello PLS “ridotto” è dovuto ricorrere a 6 variabili latenti per trovare un minimo del criterio MSEP: ciò significa che

entrambi gli algoritmi di selezione delle variabili apportano un miglioramento rispetto alla situazione senza selezione delle variabili.

I profili di composizione nelle Figure 3.5, 3.9 e 3.12 e i relativi scarti quadratici medi di convalida permettono di confrontare i modelli costruiti e di stabilire quale di essi sia il più opportuno per la modellazione in questo esempio.

La diminuzione di variabili predittive permette sicuramente un miglioramento nella fase di costruzione del modello PLS, in quanto riduce il rumore portato all'interno del modello e minimizza il criterio MSEP sia in valore assoluto sia per quanto riguarda il confronto a parità di variabili latenti. In particolare:

1. l'algoritmo AS permette di minimizzare il rumore perché esclude dalla selezione la variabile T_{30};
2. l'algoritmo VIP minimizza il criterio MSEP e riesce a massimizzare la varianza spiegata con il minimo numero di variabili latenti.

L'algoritmo AS è stato utilizzato in un'altra Tesi (Del Bosco, 2005) per selezionare variabili in un *set* di dati differente (prova ad onde quadre) ma riferito allo stesso processo. È interessante confrontare l'ordine di selezione delle variabili nei due casi, riportato in Tabella 3.10.

Tabella 3.10. *Lista delle temperature e delle densità ordinate per varianza spiegata dall'algoritmo di selezione AS. Le variabili selezionati dal primo modello riguardano la prova ad onde quadre (OQ), mentre le variabili selezionate dal secondo modello riguardano i sei sets di dati (6S) utilizzati in questa Tesi.*

posizione		**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
OQ	Variabile	T_B	ρ_D	T_4	ρ_B	T_{26}	T_{18}	T_{12}	T_{22}
	Var. Cum. Y [%]	68.45	92.96	96.49	97.77	98.19	98.99	99.10	99.15
6*S*	Variabile	ρ_B	ρ_D	ρ_A	T_D	T_B	T_{12}	T_{22}	T_4
	Var. Cum. Y [%]	67.45	89.08	94.20	97.72	99.01	99.07	99.13	99.19

Non c'è accordo né sulle variabili scelte, né sull'ordine di ingresso, nonostante entrambi i *sets* di dati esplorino la stessa variabilità di processo. In particolare, preoccupa la selezione delle due variabili inutili ρ_A e T_D nel set 6*S* che non era avvenuta con la prova ad onde quadre. La maggiore importanza era intuibile osservando il diagramma dei *loadings* in Figura 3.7, dove le variabili inutili si trovano lontano dall'origine dei due assi.

Nonostante ciò, il modello AS-PLS si dimostra competitivo nella stima delle composizioni di testa e di fondo rispetto al modello VIP-PLS a al modello PLS.

3.3 Stima della composizione di residuo (x_B)

Si analizzano ora i modelli PLS calibrati per stimare le due variabili di qualità separate.
Come avvenuto per la precedente analisi, si procede al confronto dei modelli PLS costruiti sia con il *set* completo di variabili predittive, sia con i sottoinsiemi ottenuti applicando gli algoritmi di selezione di variabili VIP e AS.
Il modello PLS contenente tutte e 12 le variabili predittive per la stima della composizione del residuo individua un minimo del criterio MSECV basato sulla *cross-validation* per 2 variabili latenti, mentre la varianza spiegata della variabile dipendente suggerisce di usare 2 o 4 variabili latenti. Si decide comunque di costruire tutti i modelli da 2 a 5 variabili latenti per poi confrontare la loro capacità predittiva.
Il modello a 2 variabili latenti spiega solo il 72.69% della varianza di **X** e il 90.27% della varianza di **Y**: anche in questo caso la selezione delle variabili predittive dovrebbe diminuire la quantità di varianza di **X** inesplicata.
L'analisi del criterio MSEP mostra che, in questo caso, il modello PLS migliore si ottiene utilizzando 5 variabili latenti e smentendo ancora una volta la *cross-validation*.
Il profilo di composizione del modello ottimo si trova in Figura 3.15, mentre il valore dello scarto quadratico medio di predizione di tutti i modelli costruiti è riportato nella Tabella 3.11.

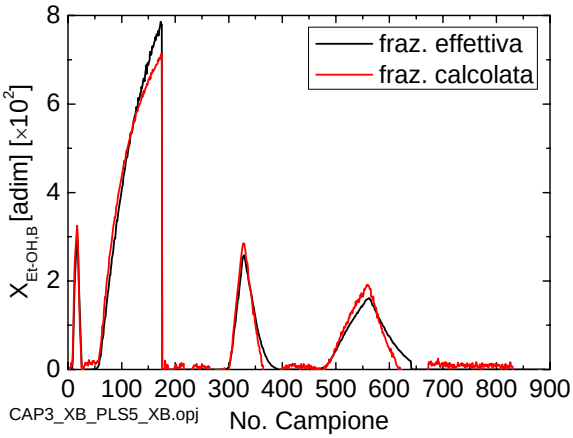


Figura 3.15. *Profilo di composizione del residuo del modello PLS calibrato con tutte le variabili predittive, ottenuto con dati di convalida.*

Tabella 3.11. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello PLS.*

LV	$MSEP_B$ [×10^6]
2	35.490
3	11.083
4	5.889
5	5.368

Si nota che la stima non è molto precisa, ma il rumore è sicuramente diminuito; ciò significa che questo era causato più da una rotazione errata delle variabili latenti che dalla presenza di variabili con un basso rapporto segnale/rumore.
Rispetto al modello PLS calibrato con 12 variabili predittive per la stima di entrambe le composizioni il valore dello scarto quadratico medio è leggermente più alto. Questo può significare che i modelli PLS costruiti per calibrare sia la frazione molare di distillato che di residuo minimizzavano il criterio MSEP, ma facendo questo rappresentavano anche il rumore di misura. Forse la selezione di variabili può migliorare la stima.
I residui, anche in questo caso, non sono statisticamente significativi secondo il test di Shapiro-Wilk.
L'algoritmo VIP seleziona 2 sole variabili, T_B e ρ_B. Si può notare come l'algoritmo VIP dipenda fortemente dal modo in cui è costruito il modello PLS: il modello calibrato per stimare contemporaneamente sia il prodotto di testa che quello di fondo, infatti, era tale da indurre il criterio VIP a selezionare tre variabili per la stima della composizione x_B, cioè T_B, ρ_B e ρ_D, mentre la calibrazione rispetto a una sola variabile dipendente modifica la rotazione degli *scores* e quindi anche l'importanza delle variabili predittive.
Nonostante ciò, il diagramma dei *loadings* rimane quasi inalterato rispetto al modello calibrato per la stima contemporanea delle due variabili dipendenti, e non viene qui riportato. Ciò significa che la rotazione delle variabili latenti non è stata elevata rispetto al modello PLS trattato nel § 3.2.
Si costruisce un solo modello VIP-PLS a due variabili latenti, che spiega il 100% della varianza di **X** e il 96.80% della varianza di **Y**. Questo modello stima la composizione di fondo con buona precisione, ma inferiore al modello PLS con tutte le variabili predittive: si ha infatti $MSEP_B = 1.0351 \times 10^{-5}$.
Il profilo di composizione è riportato in Figura 3.16.

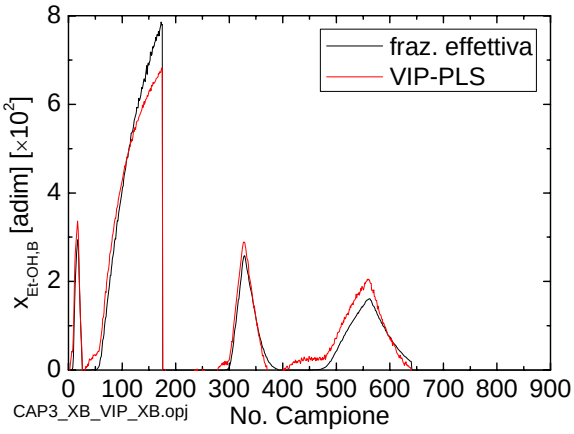


Figura 3.16. *Profilo di composizione della frazione molare di residuo per il modello VIP-PLS calibrato con due variabili predittive, T_B e ρ_B, ottenuto con dati di convalida.*

La selezione di variabili compiuta dal criterio VIP in questa situazione mette in luce un punto negativo di questo criterio: quando alcune variabili hanno valori di loadings nettamente superiori agli altri, l'algoritmo di selezione si focalizza su queste variabili predittive, scartando le altre che pur contribuiscono all'informazione contenuta nel modello. In altre parole, il criterio VIP tende a selezionare, in particolari occasioni, troppe poche variabili.

L'algoritmo AS viene indotto a selezionare due variabili predittive per poterlo confrontare con il modello VIP-PLS; le variabili selezionate sono T_B e ρ_D. La scelta è in completa analogia con quanto trovato in Dal Bosco (2005), in Tabella 3.7, in cui però vengono utilizzate tre misure per la calibrazione di un modello ai minimi quadrati classico.

Il modello PLS a 2 variabili latenti costruito con queste due variabili spiega sempre il 100% della varianza di **X** e il 96.90% della varianza di **Y**, superiore all'analogo valore ricavato dal modello VIP-PLS.

In Figura 3.17 è riportato il profilo stimato della composizione del residuo.

Il risultato è confrontabile con quello ottenuto sempre in Dal Bosco (2005): si confrontino, ad esempio, le Figure 3.16 e 3.17 di questa Tesi con la figura 3.14 in Dal Bosco (2005). Il modello PLS ottenuto selezionando due variabili predittive, sia con il metodo VIP che con il metodo AS, permette di costruire uno stimatore preciso altrettanto quanto lo stimatore costruito con un modello ai minimi quadrati calibrato con tre variabili predittive.

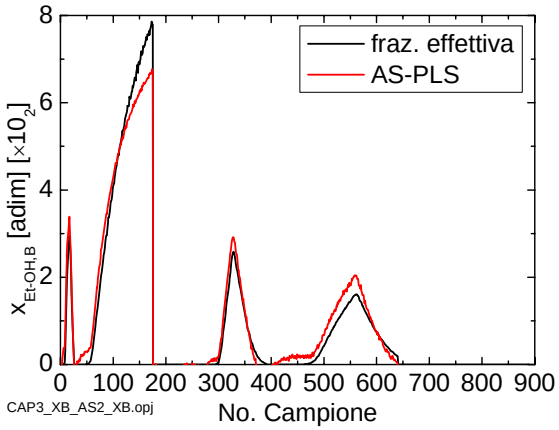


Figura 3.17. *Profilo di composizione del residuo per il modello VIP-PLS calibrato con due variabili predittive, T_B e ρ_D.*

Anche in questo caso il rumore di misura è molto meno forte, anche se lo scarto quadratico medio pari a 1.0256×10^{-5}, è superiore allo scarto quadratico medio trovato calibrando il modello PLS in funzione di entrambe le variabili dipendenti.

3.4 Stima della composizione di distillato (x_D)

Per la costruzione dello stimatore più opportuno per la stima della composizione di distillato, si analizzano il criterio MSECV e la varianza spiegata del modello PLS calibrato con tutte e 12 le misure delle variabili predittive a disposizione.
L'indicazione è di utilizzare 2 variabili latenti, secondo il criterio della quantità di varianza spiegata di **Y**, oppure 5 variabili latenti, dove si individua un minimo del criterio MSECV. Come al solito, si preferisce costruire tutti i modelli PLS compresi tra 2 e 5 LV.
Il modello a 5 variabili latenti minimizza lo scarto quadratico medio di predizione, e viene selezionato come ottimo per il modello PLS a 12 variabili predittive. Il profilo di composizione calcolato con questo modello è riportato in Figura 3.18.
In Tabella 3.12 si riportano i valori dello scarto quadratico medio di predizione ottenuti per tutti i modelli PLS costruiti. In analogia con quanto visto nel § 3.2, in Tabella 3.3 la stima della composizione di distillato mostra un grande miglioramento aumentando il numero di variabili latenti: in questo caso la diminuzione drastica si è avuta tra 4 e 5 LV, mentre nel § 3.2 un analogo risultato si era avuto tra 5 e 6 LV.
Nonostante sia ancora presente del rumore, il miglioramento rispetto al modello PLS calibrato per stimare entrambe le composizioni è netto, sia per quanto riguarda la capacità di riprodurre il profilo sia per la minimizzazione del rumore.

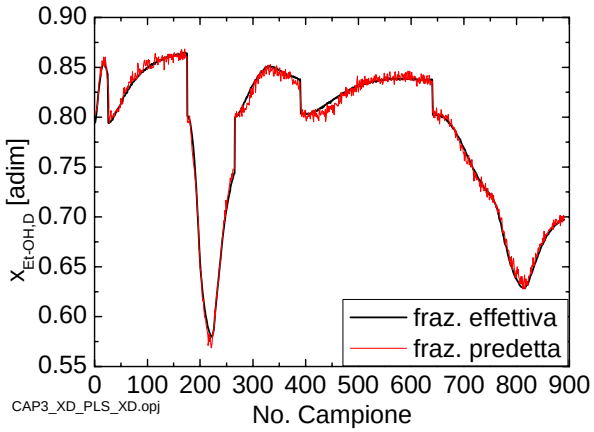


Figura 3.18. *Profilo di composizione del distillato per il modello PLS a 5 variabili latenti calibrato con tutte le variabili predittive.*

Tabella 3.12. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello PLS.*

LV	$MSEP_D$ [$\times 10^5$]
2	18.760
3	14.083
4	10.534
5	2.5505

Si possono confrontare le Tabelle 3.3 e 3.12: sebbene inizialmente i modelli contenenti poche variabili latenti siano migliori nel caso dei modelli PLS calibrati rispetto ad entrambe le variabili dipendenti, con 5 variabili latenti la situazione si inverte e si ottiene il risultato di un miglioramento delle stime.

Si passa ora all'analisi dei modelli contenenti un numero inferiore di variabili predittive.

L'algoritmo VIP seleziona 7 variabili, nell'ordine ρ_D, T_{30}, T_{22}, T_{18}, T_{26}, T_{12}, T_4. Anche in questo caso le variabili selezionate sono diverse da quelle selezionate per la stima di entrambe le variabili dipendenti.

Sia il criterio RMSECV ottenuto mediante *cross-validation* che la quantità di varianza spiegata della matrice **Y** indicano il modello a 4 variabili latenti come migliore. Ciò è confermato dallo scarto quadratico medio di predizione riportato in Tabella 3.13, mentre in Figura 3.19 si ha il profilo sperimentale ottenuto con il modello VIP-PLS per la stima della composizione di distillato.

Nonostante i valori del criterio MSEP siano superiori nel modello VIP-PLS rispetto al modello PLS calibrato su tutte e 12 le variabili predittive, lo scarto è minimo; inoltre il modello VIP-PLS consente di risparmiare quasi la metà delle variabili predittive e una variabile latente.

Tabella 3.13. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello PLS.*

LV	$MSEP_D$ [$\times 10^5$]
2	31.591
3	13.822
4	4.3664
5	3.1338

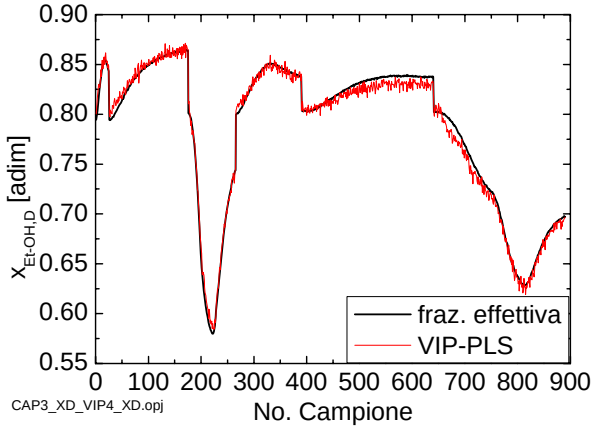


Figura 3.19. *Profilo di composizione del distillato per il modello VIP-PLS a 4 variabili latenti.*

Anche in questo caso si riesce a minimizzare il rumore, sebbene il profilo di composizione stimato non segua perfettamente il profilo effettivo della variabile dipendente.

Le prime sette variabili predittive selezionate dall'algoritmo AS sono, nell'ordine: ρ_D, T_D, T_4, T_{18}, T_B, ρ_B, T_{22}. A parte l'inclusione della temperatura del riflusso di distillato, considerata inutile, le altre variabili predittive sono significative e in accordo con le variabili selezionate dal metodo VIP.

Si costruiscono 4 modelli PLS con questo insieme ridotto di variabili, da 2 a 5 variabili latenti; il modello migliore è risultato essere quello a 5 variabili latenti, che minimizza il criterio MSEP.

Si riportano, in successione, il profilo calcolato della frazione molare di distillato con il modello AS-VIP a 5 LV in Figura 3.20 e lo scarto quadratico medio di tutti i modelli PLS elaborati in Tabella 3.14.

Il profilo ottenuto è molto buono, segno che la selezione operata con l'algoritmo AS è migliore di quella operata con l'algoritmo VIP. In articolare, si ritiene che tale miglioramento sia dovuto all'esclusione della variabile T_{30} dal sottoinsieme ridotto di variabili, a causa della correlazione con la variabile densità del distillato, ρ_D.

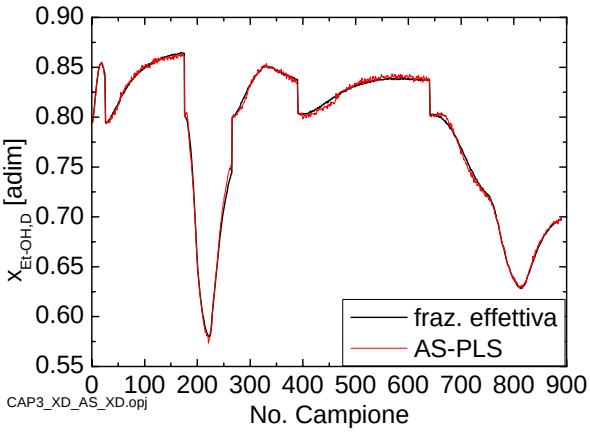


Figura 3.20. *Profilo di composizione del distillato per il modello AS-PLS a 5 variabili latenti.*

Tabella 3.14. *Valori numerici dello scarto quadratico medio di predizione calcolato su dati di calibrazione per il modello PLS.*

LV	$MSEP_D$ [$\times 10^5$]
2	36.388
3	7.0628
4	3.0383
5	0.9705

Confrontando i valori del criterio MSEP nelle Tabelle 3.12, 3.13, 3.14 si vede che il modello AS-PLS a 5 variabili latenti consente la minimizzazione assoluta di questo criterio,

e può quindi essere considerato il metodo di selezione ottimale in questa analisi per la ricerca delle variabili predittive migliori per la stima della frazione molare di distillato.
La stessa cosa non si può ripetere per il modello calibrato per la stima della composizione di residuo, forse a causa del ridotto numero di variabili incluse nel sottoinsieme; probabilmente, come mostra anche Del Bosco (2005), che nella sua Tesi seleziona 3 variabili predittive, l'inclusione di un'ulteriore variabile predittiva migliorerebbe la stima della composizione di fondo. Tuttavia, l'assenza di un criterio definito di stop per la selezione di variabili in questo metodo rende difficile la scelta della dimensione ottima del sottoinsieme di variabili predittive. L'utilizzo della varianza totale spiegata, inoltre, non fornisce buone indicazioni come criterio di stop: aggiungendo una variabile (ρ_B) al modello AS-PLS per la stima della composizione di fondo, infatti, si aumenta solo dello 0.36% la varianza spiegata totale della matrice **Y** per un modello a 2 variabili latenti, passando da 96.90% a 97.26%.

3.5 Conclusioni

Sulla base dei modelli analizzati, è difficile stabilire quale dei due algoritmi utilizzati con il *set* di dati simulati risulti migliore; si possono però fare alcune osservazioni utili, riservandosi di dare conclusioni più certe dopo l'analisi del secondo *set* di dati sperimentali nel Capitolo successivo.

1. Il metodo VIP seleziona sempre variabili di processo significative, mentre AS a volte seleziona variabili considerate poco importanti dal processo;
2. il metodo VIP riesce quasi sempre a massimizzare la varianza spiegata, ma spesso ciò lo porta a includere variabili rumorose nel modello;
3. il metodo AS, selezionando variabili tra loro ortogonali, riesce a cogliere la struttura dei dati senza inserire troppe variabili ridondanti o con un basso rapporto segnale/rumore;
4. il metodo AS non ha un criterio definito di stop e quindi non sempre è possibile sapere quante variabili è opportuno inserire nel modello ridotto.

Capitolo 4

Selezione delle variabili in un processo industriale di produzione di resina

In questo Capitolo verrà presentato un ulteriore esempio di selezione di variabili con le tecniche analizzate precedentemente. I dati provengono da un impianto industriale che produce resine di diverso tipo in reattori *batch* seguendo ricette stabilite. Lo scopo finale dell'analisi è quello di ottenere uno stimatore *software*, basato su un modello PLS costituito da un insieme ridotto di variabili, per mezzo del quale poter predire la viscosità e il numero di acidità della resina ad ogni istante del *batch*.

4.1 Introduzione

In questa sezione viene brevemente descritto l'impianto *batch* di produzione della resina, prestando particolare attenzione alle variabili monitorate *on-line* e alla misure di qualità effettuate *off-line*.

4.1.1 Descrizione del Processo

I dati sperimentali utilizzati per sviluppare il sensore software sono ricavati da un processo di produzione di resina poliestere attraverso la policondensazione di 1,6 esandiolo e acido dodecandioico che avviene all'interno di un reattore *batch* agitato, la cui capacità nominale è 12 m^3. La reazione avviene a circa 200°C, ed è necessario fornire calore al sistema mediante una serpentina esterna al reattore in cui scorre olio diatermico. La formazione del polimero provoca la formazione di acqua, che viene rimossa dal reattore per favorire la reazione diretta; la rimozione avviene in una colonna a corpi di riempimento, dove l'acqua in eccesso viene catturata sotto forma di vapore e raffreddata in un condensatore esterno. L'acqua residua viene eliminata in uno *scrubber*.

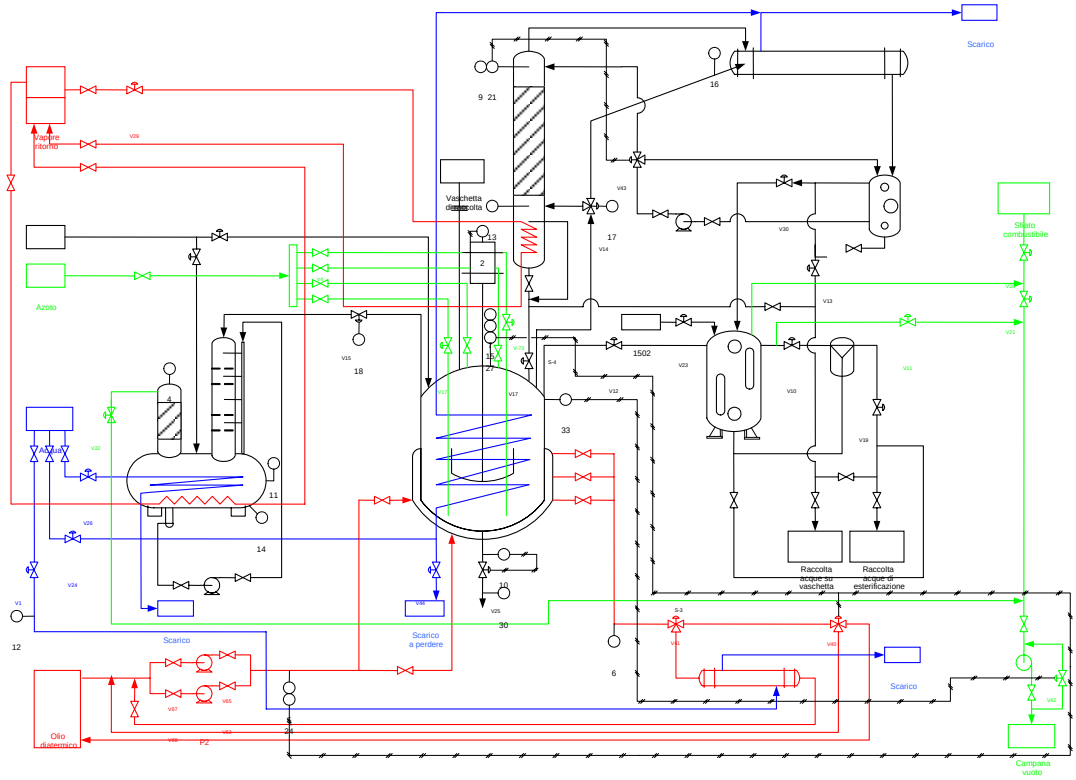


Figura 4.1: *Schema dell'impianto di produzione della resina.*

La qualità della produzione è regolata seguendo determinate ricette, che variano da resina a resina; solitamente, tutte le ricette comprendono una fase di carico, una di riscaldamento, in cui avvengono una o più reazioni, l'eventuale aggiunta di altre materie prime, un nuovo riscaldamento, un periodo di attesa, e il raffreddamento.

L'inizio di ogni fase è determinato da un *set point* (temperatura del reattore per la fase di riscaldamento) o dal raggiungimento di particolari condizioni (esaurimento della reazione principale). Non esistono invece regole precise per la durata di ogni fase.

Il principale problema nel seguire la ricetta consiste nell'assenza di misure in linea della qualità del prodotto: la conoscenza della situazione del processo può avvenire solo attraverso il campionamento del contenuto del *batch*, che viene inviato in laboratorio per analisi. L'analisi dei campioni fornisce due valori strettamente collegati alla qualità del prodotto: il numero di acidità (NA), che indica il grado di completamento della reazione, espresso in mg di KOH (idrossido di Potassio) necessari per neutralizzare un grammo di resina (è un indice indiretto della presenza di gruppi iniziatori di reazione –COH negativi elettron-attrattori all'interno della resina), e la viscosità (MU), espressa in Poise [P].

La ricetta prevede la fine della fase di riscaldamento e la fine della reazione per determinati valori del numero di acidità e della viscosità: la fase di riscaldamento non viene interrotta finchè non si raggiungono i valori di specifica.

Oltre all'assenza di indicazioni sulla durata delle fasi, ogni *batch* può prevedere ulteriori fasi non indicate nella ricetta, dette correzioni: ciò avviene quando la qualità del prodotto non è quella attesa secondo le analisi sui primi campioni. In questo caso si procede innescando il vuoto spinto nell'impianto (utile a far evaporare l'acqua prodotto di scarto della policondensazione) ed inserendo ulteriori reagenti che conducano la reazione in modo da avere una resa ottimale. La deviazione dalla ricetta standard introduce nuove fonti di variabilità nel processo: se complessivamente ogni *batch* è ottimale, perché mediante le correzioni si riesce comunque ad ottenere il prodotto desiderato, non è facile capire quali fattori portano a durate diverse di ogni *batch*, delle singole fasi e delle qualità finali.

I fattori che maggiormente influenzano il *batch* sono la qualità delle materie prime e la traiettoria seguita dalle variabili monitorate: si decide quindi di trascurare la variabilità dipendente dalle materie prime e di costruire uno stimatore in grado di prevedere la qualità del prodotto a partire dalle variabili monitorate.

A differenza degli stimatori finora costruiti (Duchesne e MacGregor, 2000) l'obiettivo non è la determinazione della qualità *finale* del *batch*, ma la determinazione istantanea del numero di acidità e della viscosità, in modo da poter immediatamente determinare la durata di ogni fase (o il grado di completamento della stessa) e l'eventuale necessità di apportare correzioni al *batch*. In questo modo ci si attende di ridurre la durata di ogni fase, diminuire il numero di campionamenti necessari ed eventualmente, in prospettiva, determinare *best practice* che migliorino la ricetta attuale.

L'impianto di produzione monitora complessivamente 34 variabili *on-line* con un intervallo di campionamento automatico di 30 s., per un totale di 4500÷7500 misure disponibili ad ogni *batch*; di queste, 23 sono state selezionate per costruire lo stimatore, escludendo i *set-point* ed altre variabili considerate poco influenti sul processo. Le variabili predittive sono elencate in Tabella 4.1.

Durante un esercizio *batch*, in media, vengono effettuati 20÷25 campionamenti di qualità del prodotto, ad intervalli di tempo non regolari.

Si costruisce prima uno stimatore che utilizza tutte le variabili predittive, poi si applicheranno i metodi di selezione di variabili esposte, si costruiranno i modelli PLS ridotti e si opereranno i confronti. Per costruire lo stimatore si ha a disposizione un *set* di 36 *batches*, di cui 9 verranno trascurati in fase di calibrazione per poter essere utilizzati nella fase di convalida. I *batches* utilizzati per la calibrazione sono considerati ottimi riferimenti in quanto non hanno subito troppe correzioni rispetto alla ricetta originale. Alcuni dei *batches* di convalida, invece, hanno subito forti correzioni o hanno avuto altri generi di problemi; si vuole capire se lo stimatore è in grado di identificare i *batches* anomali.

I dati di calibrazione sono contenuti nel file `DATI_CALIBRAZIONE.mat`, quelli di convalida nel file `DATI_CONVALIDA.mat`.

Tabella 4.1. *Elenco delle variabili monitorate nel processo, associate al rispettivo numero identificativo.*

Variabile monitorata	Simbolo	No. Identificativo
Giri reattore (OP, PV)	$G_{R,OP}$;$G_{R,PV}$	1,2
T linea vuoto	T_V	3
T olio in ingresso	$T_{Olio,in}$	4
T olio in uscita	$T_{Olio,out}$	5
T reattore	T_R	6
T testa colonna	T_{TC}	7
T vapore scrubber	T_{VS}	8
T acqua in ingresso	$T_{H20,in}$	9
T fondo colonna	T_{FC}	10
T fondo scrubber	T_{FS}	11
T reattore (2)	T_{R2}	12
T vapore via breve	T_{VVB}	13
T valvola colonna	T_{ValvC}	14
T valvola scrubber	T_{ValvS}	15
T vuoto reattore	T_{VuotoR}	16
T colonna (PV)	$T_{C,PV}$	17
T olio (OP, PV)	$T_{Olio,OP}$;$T_{Olio,PV}$	18,19
T reattore (OP, PV)	$T_{R,OP}$;$T_{R,PV}$	20,21
T vuoto reattore (OP, OV)	$T_{VR,OP}$;$T_{VR,PV}$	22,23

Da precedenti studi effettuati (Facco *et al*., 2007), per linearizzare la relazione tra le variabili dipendenti e predittive, si è deciso di separare il processo in tre fasi da studiare separatamente: la prima fase inizia dopo il carico dei reagenti e dura per tutta la prima fase di riscaldamento, la seconda inizia dopo la prima rottura del vuoto e copre diverse fasi del processo, la terza inizia quando il *set point* della temperatura del reattore viene innalzato da 202°C a 218°C per portare a termine la reazione ed eliminare l'acqua residua rimanente.
Anche l'analisi dei coefficienti di correlazione indica che la struttura delle relazioni tra le variabili si modifica durante il processo: la separazione in fasi diventa quindi una scelta obbligata per seguire meglio la traiettoria delle variabili.
Ogni fase ha durata diversa, ed anche tra i *batche*s la durata delle stesse fasi non è omogenea: si è provato ad effettuare un allineamento dei *batches* costruendo un modello PLS che utilizzi la variabile tempo come variabile dipendente, ma non si è individuata nessuna variabile predittiva correlata con la variabile tempo e capace di indicare il grado di avanzamento del processo. Questo è dovuto alle correzioni apportate ad ogni *batch* che impediscono di individuare traiettorie comuni di una stessa variabile all'interno di diversi *batches*, in modo da ottenere un indice dello stato del processo. Si è quindi rinunciato all'allineamento dei *batches*.

La singolarità di ogni *batch* rispetto agli altri ha anche conseguenze numeriche nella costruzione dello stimatore: come detto, la costruzione del modello PLS richiede la normalizzazione delle variabili sui dati disponibili in calibrazione. L'applicazione del modello sui dati di convalida necessita lo *scaling* anche dei dati di convalida, che però hanno valori di media e varianza molto diversi rispetto a quelli di calibrazione: si è dovuto decidere quale riferimento usare per scalare le variabili di convalida. Sebbene Sharmin *et al.* (2006) raccomandino l'utilizzo di valori diversi di media e varianza per ogni set di convalida, si ritiene che in questo caso sia appropriato utilizzare i valori ottenuti dalla calibrazione, perché i valori dei coefficienti di regressione si riferiscono esclusivamente ai valori di *scaling* di calibrazione. D'altra parte, se si riconoscesse l'unicità di ogni *batch* e si scalassero le variabili individualmente, non avrebbe neanche senso costruire uno stimatore valido per tutti i *batches*. Per lo stesso motivo, la fase di calibrazione è stata costruita scalando le variabili rispetto ad un unico valore di media e di varianza, invece che utilizzare i valori singoli di ogni *batch*.

Può essere comunque interessante cercare di capire il motivo di queste forti differenze anche tra *batches* che seguono la ricetta originale: a tale scopo, si dovrebbe cercare di includere la qualità delle materie prime all'interno del modello, per vedere come queste influenzino la correlazione tra le variabili e lo sviluppo del *batch*.

Prima di costruire il modello PLS con tutte le variabili predittive, si esegue una PCA per controllare il grado di correlazione tra le variabili dipendenti e capire se è possibile costruire uno stimatore unico valido sia per NA che per MU. Il modello PCA mostra una grande correlazione inversa tra MU e NA (-0.84 per la prima fase e leggermente inferiore per le altre due); di contro, però, il modello complessivo costruito per stimare contemporaneamente le due proprietà mostra un errore relativo medio molto più alto nella stima del numero di acidità. Di conseguenza, si decide di implementare due stimatori separati per la predizione della viscosità e del numero di acidità.

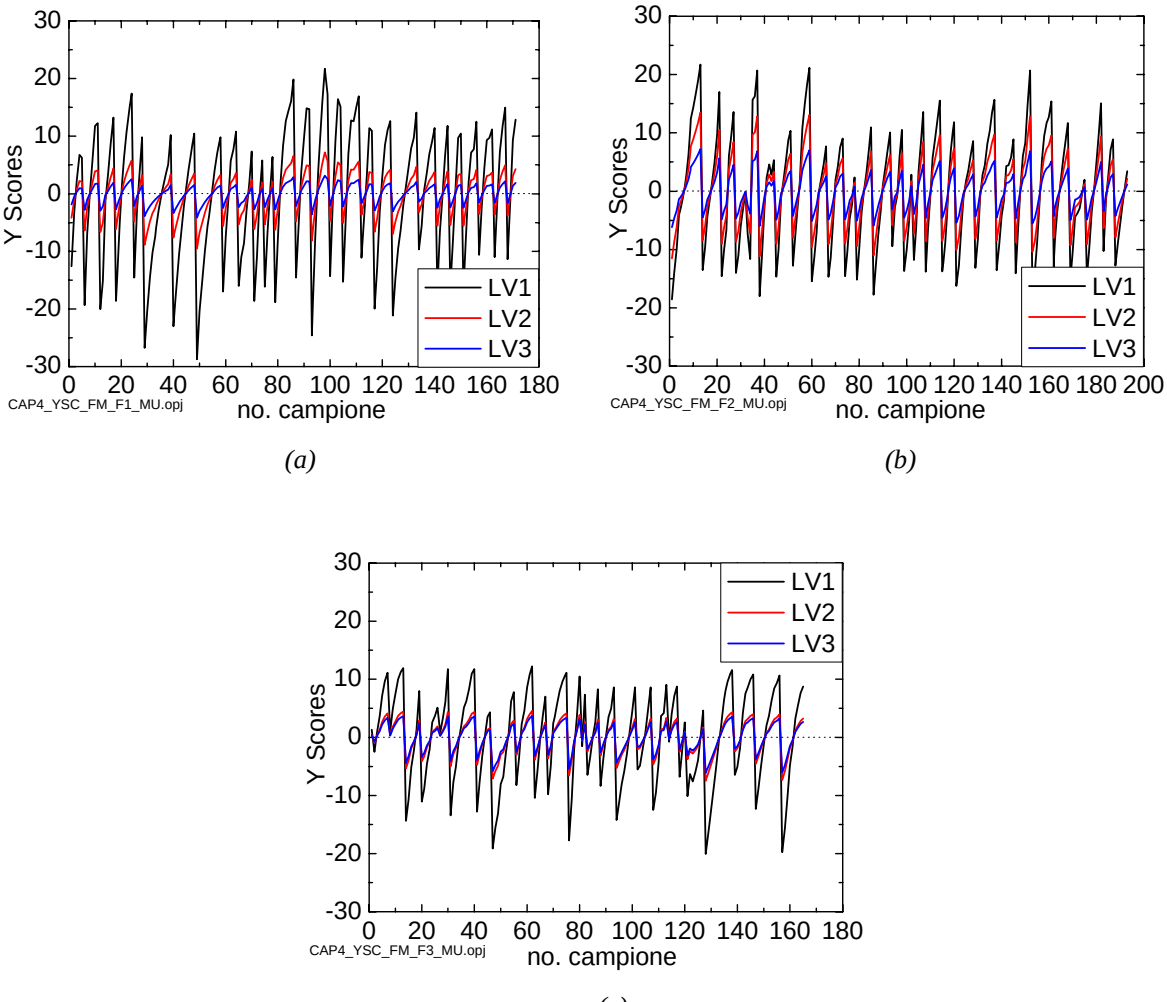
L'analisi del modello PCA mostra anche che la correlazione tra variabili predittive e dipendenti diminuisce passando dall'inizio del processo alla fine. Questo risultato era atteso, perché l'esercizio *batch* è un processo integrativo, in cui cioè le condizioni iniziali e/o precedenti hanno influenza determinante sulle fasi successive: ciò significa che le fasi finali di un *batch* dipendono più dagli eventi iniziali piuttosto che dalle condizioni attuali. A questo proposito, sarà interessante vedere se lo stimatore calibrato solo per essere utilizzato per la prima fase sarà utile anche nelle fasi successive e se la sua capacità predittiva sarà confrontabile con quella degli stimatori utilizzati sulla stessa fase sulla quale sono stati calibrati.

In conclusione, si dovranno costruire, inizialmente, sei modelli PLS: due per ogni fase, il primo per stimare NA, il secondo per stimare MU.

4.2 Modello PLS a 23 variabili predittive

Vengono trattati simultaneamente i sei modelli PLS calibrati su tutte e 23 le variabili predittive, evidenziandone similitudini e differenze.
I modelli della prima, seconda e terza fase vengono calibrati utilizzando rispettivamente 5, 4 e 3 LV: la diminuzione di LV è motivata con l'abbassarsi del contenuto informativo delle variabili predittive passando dall'inizio alla fine del *batch*; in questo modo si vuole evitare di modellare anche parte del rumore contenuto nelle fasi successive alla prima. La scelta del numero di LV non è stata indicata dalla *cross validation*, ma dalla conoscenza del processo.
L'analisi dei *loadings* delle prime due variabili predittive mostra che le variabili formano *clusters* che si mantengono anche nelle successive fasi del processo, anche se la posizione del *cluster* varia all'interno del diagramma; la struttura reciproca delle variabili è identica per i modelli PLS calibrati sia rispetto a MU che rispetto a NA. Variabili o *clusters* di variabili importanti nella prima fase del processo (alto valore di loadings per la prima e la seconda LV) possono diventare trascurabili nelle fasi successive, e viceversa; ad esempio, le variabili 1 e 2 diventano progressivamente sempre più importanti nel modello, mentre 4 e 5 sono utili solo nella prima fase.
Inoltre, molte variabili si trovano in tutte e tre le fasi vicino al centro del diagramma, sintomo che sono poco importanti nella modellazione dello stimatore.
Gli *scores* di **X** della prima e della seconda variabile latente mostrano i dati omogeneamente raggruppati in tutte e tre le fasi e rispetto ad entrambe le variabili dipendenti.
Gli *scores* della prima variabile latente della viscosità, rappresentati contro il tempo in Figura 4.2, mostrano, per tutte e tre le fasi, un profilo a dente di sega dovuto al fatto che i dati delle variabili dipendenti sono ottenuti unendo i profili esponenziali di ogni *batch* uno di seguito all'altro; questo profilo però peggiora e diventa meno riconoscibile e leggermente caotico passando dalla prima alla terza fase, in cui la correlazione tra variabili dipendenti e predittive è meno forte. Analoghi risultati si ottengono dall'analisi degli *scores* del numero di acidità.
Gli *scores* della seconda e terza variabile latente della viscosità, rappresentati contro il tempo, hanno lo stesso profilo a dente di sega con la stessa distanza tra un picco e l'altro, ma differiscono per l'altezza del profilo: la prima variabile latente coglie la struttura generale dei dati e ha il profilo più elevato, la seconda ha picchi meno accentuati e corregge le sbavature del modello per le predizioni dei valori minori di NA e MU, mentre la terza variabile latente, con picchi ancora meno accentuati, migliora la modellazione dei valori calcolati estremi.

Inoltre, il profilo degli *scores* per le prime tre variabili latenti tende a mescolarsi passando dalla prima fase alla terza, sovrapponendo gli effetti. Questo fatto è probabilmente dovuto al fatto che nella seconda e terza fase esistono *batches* aventi picchi di variabili misurate inferiori a valori centrali delle stesse variabili misurate in altri *batches*.



(a) *(b)*

(c)

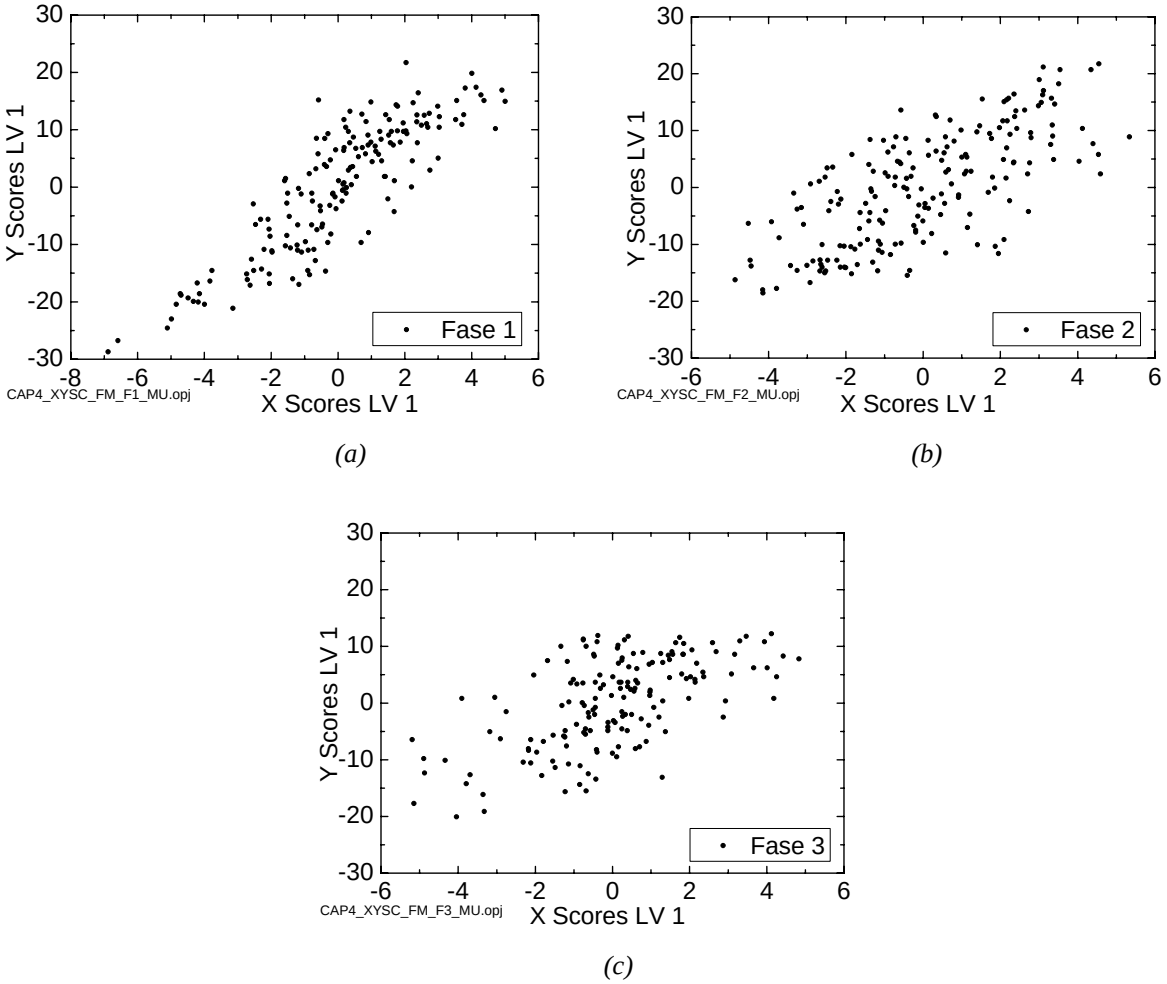
Figura 4.2. *Scores delle prime tre variabili latenti della variabile dipendente viscosità relativamente alla (a) prima, (b) seconda e (c) terza fase. Nelle seconda e terza fase le traiettorie sono più confuse e si sovrappongono, impedendo di distinguere il significato di ogni LV.*

Nelle figure 4.3 e 4.4 sono riportati i grafici degli *scores* della prima variabile latente di **X** e **Y** per le tre fasi di MU e solo della prima fase di NA.

Si possono fare le seguenti considerazioni:

1. il peggioramento della capacità predittiva dei modelli PLS relativi alle fasi 2 e 3 è evidente dall'analisi dei grafici degli *scores* delle prime variabili latenti di **X** e **Y**; mentre nella prima fase i punti sperimentali giacciono compatti su una retta, nelle successive fasi i dati si presentano sempre più dispersi, senza cogliere la relazione lineare tra gli scores delle variabili predittive e gli *scores* delle variabili dipendenti;

2. il maggiore errore relativo nella predizione di NA è dovuto al fatto che il numero di acidità è una funzione molto meno lineare rispetto alle variabili predittive di quanto non lo sia MU; questo si può evincere anche dall'analisi degli *scores* della prima variabile latente di **X** e **Y**, per la prima fase, della viscosità e del numero di acidità: nel caso di NA i punti giacciono su una traiettoria meno lineare rispetto a MU, pur rimanendo compatti.



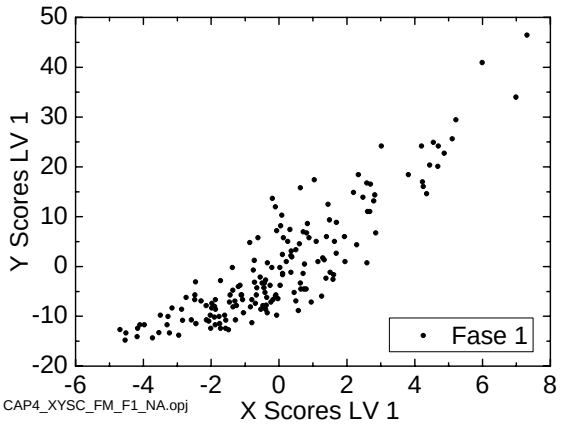
(a) *(b)*

(c)

Figura 4.3. *Scores dei dati di calibrazione della prima variabile latente di* **X** *e* **Y** *delle (a) prima, (b) seconda e (c) terza fase rispetto alla viscosità. La linearità dei dati decresce dalla prima alla terza LV, rendendo difficile la calibrazione.*

Tra i grafici osservati in fase di calibrazione, si sono considerati anche quelli relativi alle statistiche Q_{res}, T^2 e *leverage*: sebbene ogni modello mostrasse qualche possibile *outlier*, nessun punto è stato eliminato in fase di calibrazione. Tutte le misure raccolte sono sensate, ed il fatto che qualche punto abbia un valore di T^2 superiore agli altri è fisiologico, perché il processo non è stazionario e non individua necessariamente un *cluster* di dati omogenei, ma piuttosto una traiettoria nel piano degli *scores*. Inoltre, controllando il

contributo di ogni variabile ai possibili *outliers*, si è visto che le variabili maggiormente influenti sui valori di Q_{res} sono quelle meno importanti per la calibrazione del modello, meno correlate con la variabile dipendente e scartate da qualsiasi algoritmo di selezione.



(a)

Figura 4.4. *Scores dei dati di calibrazione della prima variabile latente di* **X** *e* **Y** *della prima fase rispetto al numero di acidità. I dati non sono disposti linearmente come per il modello di viscosità.*

Dopo aver analizzato la situazione di partenza, si passa ad un esame veloce di ognuno dei sei modelli costruiti, confrontando il modello costruito utilizzando tutte e 23 le variabili predittive originarie con i modelli costruiti con *set* ridotti. Si sono applicati i quattro metodi di selezione di variabili ai dati di calibrazione, ricavandone diversi *set* ridotti di variabili predittive. In particolare, si è selezionato:

1. un *set* per l'algoritmo VIP;
2. un *set* per l'algoritmo VIP in cui il limite soglia per l'inclusione delle variabili è stato abbassato fino a raccogliere tutte le variabili predittive con un coefficiente di importanza non nullo; si definisce questo *set* e i modelli costruiti utilizzando questo insieme di variabili con il codice VIP+;
3. un *set* per l'algoritmo *Stepwise regression*;
4. due *sets* per l'algoritmo SROV, il primo riguardante la selezione iniziale delle variabili, il secondo la selezione finale dopo la fase di rotazione e nuova selezione;
5. tanti *sets* selezionati secondo l'algoritmo AS quante sono le diverse dimensioni dei *sets* costruiti con gli altri sistemi di selezione; si avrà, cioè, un *set* AS di dimensioni uguali a VIP, un altro di dimensioni uguali a VIP+, SROV e *Stepwise Regression*. In questo modo si è potuto più facilmente confrontare gli algoritmi basandosi su un insieme paritetico di variabili selezionate. Spesso, inoltre, si è verificato il caso interessante in cui più algoritmi selezionavano lo stesso numero di variabili.

I modelli costruiti su un insieme ridotto di variabili sono stati valutati per diversi valori del numero di variabili latenti, perché non si è ancora in grado di distinguere il segnale dal

rumore. In ogni caso, il massimo numero di variabili latenti utilizzate non è mai stato superiore al numero utilizzato nel modello completo.

4.3 Prima fase

La modellazione nella prima fase del processo è molto buona, e migliora ancora selezionando opportunamente le variabili predittive. Nelle Tabelle 4.1 e 4.2 sono riportate le variabili selezionate nell'ordine di scelta da ogni algoritmo per la stima del numero di acidità e della viscosità:

Tabella 4.1. *Variabili selezionate da ogni algoritmo per la costruzione dello stimatore software per il numero di acidità. Le variabili selezionate con l'algoritmo AS nel primo set (S1) sono le stesse scelte con il metodo Stepwise Regression. Il set SROV S2 presenta le variabili in ordine numerico perché è il risultato di una rotazione e nuova selezione di variabili in cui tutte sono teoricamente selezionate come ultime.*

Algoritmo di selezione		Variabili Predittive Selezionate
AS	S1	10, 6, 17, 14, 13, 5, 20
	S2	10, 6 ,17, 14, 13, 5, 20, 2, 12
	S3	10, 6, 17, 14, 13, 5, 20, 2, 12, 19, 4, 3, 9
SROV	S1	10, 6, 17, 14, 13, 5, 19, 12, 9
	S2	5, 9, 10, 13, 14, 17, 19, 20, 21
VIP	VIP	10, 17, 7, 12, 21, 6, 20
	VIP+	10, 17, 7, 12, 21, 6, 20, 4, 19, 5, 11, 8, 18

Tabella 4.2. *Variabili selezionate da ogni algoritmo per la costruzione dello stimatore software per la viscosità. Le variabili selezionate con l'algoritmo AS nel primo set (S1) sono le stesse scelte con il metodo Stepwise Regression. Il set SROV S2 presenta le variabili in ordine numerico perché è il risultato di una rotazione e nuova selezione di variabili in cui tutte sono teoricamente selezionate come ultime.*

Algoritmo di selezione		Variabili Predittive Selezionate
AS	S1	10, 9, 14, 7, 4, 13
	S2	10, 9, 14, 7, 4, 13, 12
	S3	10, 9, 14, 7, 4, 13, 12, 11, 8, 2, 19, 5
	S4	10, 9, 14, 7, 4, 13, 12, 11, 8, 2, 19, 5, 22, 23
SROV	S1	10, 9, 14, 7, 4, 13, 12, 11, 8, 5, 19, 23, 18, 16
	S2	2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 19, 22, 23
VIP	VIP	10, 7, 17, 12, 6, 21, 11
	VIP+	10, 7, 17, 12, 6, 21, 11, 8, 20, 4, 19, 5

Si nota che:

1. l'algoritmo VIP e *Stepwise Regression* sono conservativi rispetto agli altri, cioè tendono a trattenere un numero minore di variabili predittive;

2. l'algoritmo SROV dipende dall'errore riportato; non potendo inserire l'errore percentuale reale (la misura sperimentale di NA e MU è affetta da un errore dell'8% e del 10% rispettivamente) perché l'algoritmo riteneva questo errore troppo elevato e selezionava una sola misura, si è selezionato un errore medio puntuale di 0.3×10^{-t}, dove *t* è l'ultima cifra decimale, come di *default* nel programma. Con questa scelta, l'algoritmo tende a selezionare più variabili degli altri algoritmi e ad allinearsi alla selezione compiuta con l'algoritmo AS;
3. l'algoritmo VIP+ seleziona variabili quasi sicuramente inutili, come la temperatura dello *scrubber* (8) e dell'olio OP (18);
4. le scelte compiute dall'algoritmo VIP sono simili nei due modelli PLS costruiti per la stima di NA e MU, mentre le variabili selezionate differiscono in modo più accentuato negli altri algoritmi di selezione;
5. la temperatura del fondo colonna (10) è riconosciuta in tutti i modelli come la più importante;
6. gli algoritmi AS, SROV e *Stepwise* sostituiscono le temperature del reattore e della colonna (6 e 17) con le temperature dell'acqua in ingresso al condensatore e della valvola della colonna (9 e 14) passando dalla stima di NA alla stima di MU; la scelta è molto strana, e fa pensare che, in questo caso, la selezione di variabili con questi metodi non sia affidabile.

Le variabili selezionate da ogni algoritmo, per la costruzione del solo stimatore del numero di acidità, sono riportate in Figura 4.5. La disposizione delle variabili nel grafico dei *loadings* per il modello calibrato rispetto alle misure di viscosità è identico.

Si può osservare, prima di tutto, che le variabili predittive tendono ad aggregarsi in *clusters* che indicano fasi o sezioni del processo. Si individuano 5 *clusters*:

1. variabili 6, 12, 21: sono tre misure ridondanti della temperatura del reattore;
2. variabili 7, 10, 17: temperature della testa, del fondo e PV della colonna;
3. variabili 4, 5, 19, 20: sono rispettivamente le temperature dell'olio in ingresso, in uscita, di nuovo dell'olio (PV) e del reattore (OP);
4. variabili 8, 11: temperature dei vapori dello *scrubber* e del fondo dello *scrubber*;
5. gruppo centrale: sono variabili secondarie, come il numero di giri del reattore, la temperatura di alcune valvole e variabili riconducibili al grado di vuoto che, essendo automatizzato, non sono soggette a variabilità.

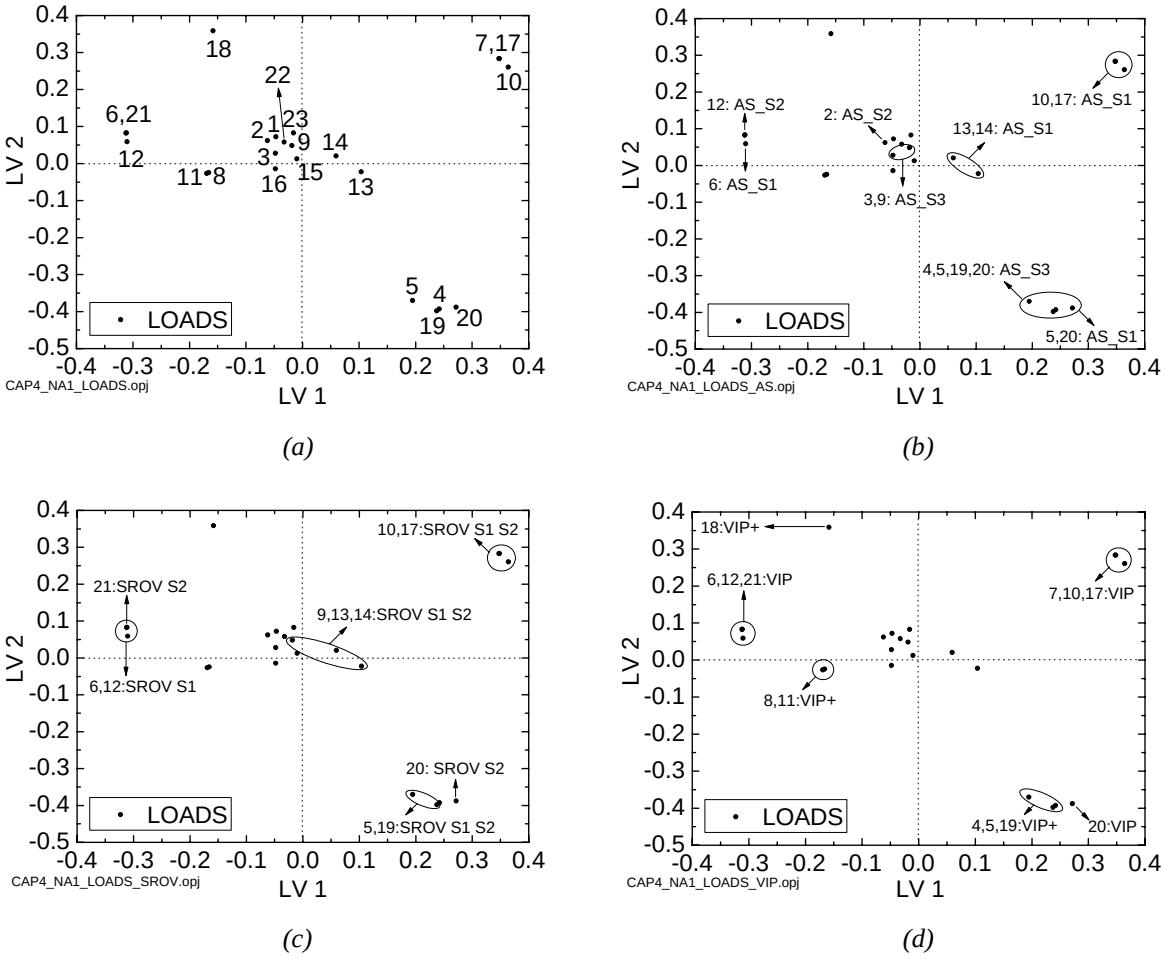


Figura 4.5. *Grafici dei loadings e indicazione (a) della posizione delle variabili predittive per il modello calibrato rispetto a NA e a MU; variabili selezionate con gli algoritmi (b) AS, (c) SROV e (d) VIP, solo per il modello calibrato rispetto a NA.*

Nonostante il valore elevato dei *loadings*, la variabile 18 è scelta solo dall'algoritmo VIP, e con importanza secondaria (è inclusa nel *set* ridotto selezionato da VIP+). Da questo punto di vista, si può dire che tutti gli algoritmi selezionano variabili in modo proficuo.

Vale la pena fare alcuni commenti sulla selezione di variabili operata da ogni algoritmo:

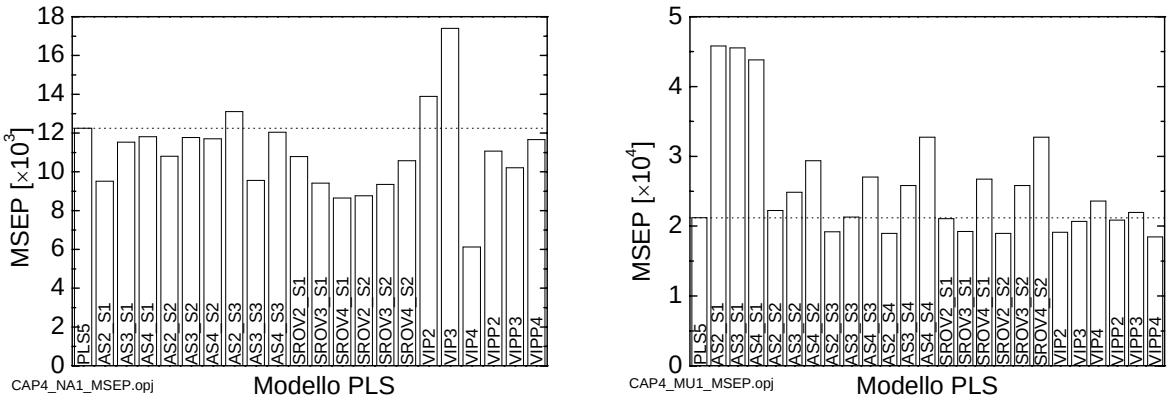
1. l'algoritmo AS (Muradore *et al*., 2006) seleziona le variabili omogeneamente da ogni *cluster*; questo era intuibile, perché ogni volta l'algoritmo aggiorna le variabili in modo da scegliere variabili ortogonali tra loro. L'algoritmo inizia a selezionare variabili in *overfitting* quando include le variabili del *set* 3: infatti seleziona contemporaneamente tutte le variabili del terzo *cluster* ed incomincia a introdurre variabili poste al centro del diagramma, poco predittive nei confronti della variabile dipendente;
2. l'algoritmo SROV (Shacham e Brauner, 2003) seleziona le variabili in modo molto simile al modello AS; in particolare, dopo la prima selezione, la rotazione delle

variabili esclude le variabili 6 e 12, ma le sostituisce con le variabili 21 (appartenente allo stesso *cluster*) e 20, significativa per il processo (temperatura del reattore);

3. l'algoritmo VIP (Chong e Jun, 2005), seleziona semplicemente le variabili aventi il più elevato valore di *loadings* lungo la prima LV; dato l'alto numero di variabili presenti, la scelta è molto selettiva (la somma dei quadrati di ogni vettore di *loading* è unitaria, e quindi più elevata è il numero di variabili, più basso è il valore di ogni singolo *loading*, rendendo la selezione più radicale).

Tutti gli algoritmi di selezione diminuiscono lo scarto quadratico medio di predizione (MSEP) e risultano particolarmente buoni i modelli contenenti il minor numero di variabili, con l'eccezione del primo set selezionato dall'algoritmo AS (a sei variabili predittive) per la stima della viscosità: confrontando le variabili scelte da questo *set* con le altre variabili incluse negli altri *sets*, si vede come la mancanza della seconda temperatura del reattore (variabile no. 12) impedisca la costruzione di un modello ottimo di regressione. Questo è un fatto importante perché mostra palesemente come la regressione PLS sia capace di sfruttare la sinergia di variabili correlate (come le due temperature disponibili per il reattore) per costruire uno stimatore preciso riducendo l'errore di calibrazione del modello.

È possibile effettuare un confronto visivo tra le capacità predittive dei modelli costruiti in Figura 4.6, mentre i valori numerici sono riportati in Tabella 4.2. Ogni modello è identificato da un codice alfa-numerico, che contiene, nell'ordine, la sigla dell'algoritmo utilizzato per la selezione delle variabili, un numero che indica il numero di variabili latenti utilizzato per questo modello e una eventuale successiva sigla (S1, S2, S3, S4) che si riferisce al *set* di variabili predittive utilizzate.



(a) *(b)*

Figura 4.6. *Scarto quadratico medio di predizione ottenuto con dati di convalida per tutti i modelli PLS elaborati rispetto (a) al numero di acidità e (b) alla viscosità. La prima colonna, PLS5, presenta lo scarto quadratico medio del modello a 5 LV e 23 variabili predittive e costituisce il riferimento, riportato con linea tratteggiata, per tutti gli altri modelli.*

Il valore del criterio MSEP aumenta o diminuisce all'aumentare del numero di variabili latenti, senza una regola precisa anche all'interno dello stesso algoritmo di selezione; probabilmente, l'inclusione di una o più variabili predittive nel modello separa il confine tra modellazione dell'informazione e modellazione del rumore; non tutte le variabili, infatti, hanno lo stesso rapporto segnale/rumore.
Si verifica anche il rapporto tra errore di predizione e varianza spiegata della sola variabile dipendente: i valori di R_Y^2 in funzione del modello sono riportati in Tabella 4.3.

Tabella 4.2. *Valori dello scarto quadratico medio di predizione per tutti i modelli PLS elaborati per la stima (a) del numero di acidità e (b) della viscosità.*

Modello PLS	MSEP		
PLS5	**0.01225**	SROV3 S1	0.00942
AS2 S1	0.00952	SROV4 S1	0.00865
AS3 S1	0.01153	SROV2 S2	0.00877
AS4 S1	0.01181	SROV3 S2	0.00935
AS2 S2	0.01081	SROV4 S2	0.01057
AS3 S2	0.01177	VIP2	0.01389
AS4 S2	0.01170	VIP3	0.01740
AS2 S3	0.01311	VIP4	0.00613
AS3 S3	0.00956	VIP+2	0.01107
AS4 S3	0.01204	VIP+3	0.01021
SROV2 S1	0.01079	VIP+4	0.01166

(a)

Modello PLS	MSEP [×10^4]		
PLS5	**2.1209**	SROV2 S1	2.1081
AS2 S1	4.5816	SROV3 S1	1.9261
AS3 S1	4.5531	SROV4 S1	2.6722
AS4 S1	4.3814	SROV2 S2	1.895
AS2 S2	2.225	SROV3 S2	2.5813
AS3 S2	2.4846	SROV4 S2	3.2756
AS4 S2	2.9359	VIP2	1.9136
AS2 S3	1.9204	VIP3	2.0696
AS3 S3	2.1296	VIP4	2.3592
AS4 S3	2.7023	VIP+2	2.0853
AS2 S4	1.895	VIP+3	2.1972
AS3 S4	2.5813	VIP+4	1.8453
AS4 S4	3.2756		

(b)

Confrontando le variabili selezionate con i valori del criterio MSEP e con la varianza spiegata si possono fare le prime conclusioni riguardanti la bontà dei modelli e gli algoritmi di selezione.

Tabella 4.3. *Valori di varianza spiegata in fase di calibrazione (R_Y^2) per tutti i modelli PLS elaborati per la stima (a) del numero di acidità e (b) della viscosità.*

Modello PLS	R_Y^2		
PLS5	**88.4997**	SROV3 S1	87.0969
AS2 S1	85.6226	SROV4 S1	88.4328
AS3 S1	87.6769	SROV2 S2	85.4479
AS4 S1	89.7765	SROV3 S2	87.2048
AS2 S2	86.2147	SROV4 S2	89.5718
AS3 S2	87.9089	VIP2	82.6899
AS4 S2	89.5845	VIP3	83.9372
AS2 S3	85.5848	VIP4	88.7808
AS3 S3	87.7499	VIP+2	84.5936
AS4 S3	89.4299	VIP+3	86.353
SROV2 S1	86.4682	VIP+4	87.8498

(a)

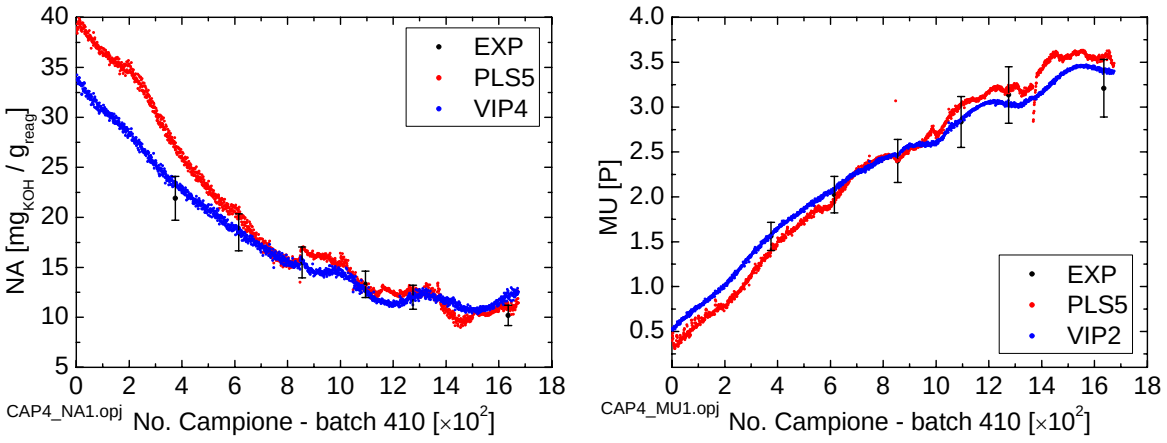
Modello PLS	R_Y^2		
PLS5	**86.5728**	SROV2 S1	84.6738
AS2 S1	84.9555	SROV3 S1	86.0258
AS3 S1	85.4912	SROV4 S1	86.444
AS4 S1	89.1128	SROV2 S2	84.7988
AS2 S2	85.9411	SROV3 S2	86.1925
AS3 S2	86.4104	SROV4 S2	86.9147
AS4 S2	87.3183	VIP2	81.7258
AS2 S3	84.9037	VIP3	82.2731
AS3 S3	85.986	VIP4	85.8497
AS4 S3	86.3839	VIP+2	82.7272
AS2 S4	84.7988	VIP+3	83.4954
AS3 S4	86.1925	VIP+4	84.8892
AS4 S4	86.9147		

(b)

1. non c'è una relazione tra varianza spiegata ed errore di predizione; ciò significa che la discriminante che determina un modello valido da uno non valido è l'informazione utile spiegata da ogni modello, maggiore dell'85% della varianza totale nel caso del numero di acidità e probabilmente minore dell'82% nel caso della viscosità;
2. non esiste una relazione lineare tra numero di variabili predittive e varianza spiegata, a parità di variabili latenti; infatti i tre *sets* selezionati con l'algoritmo AS mostrano che la varianza spiegata da un modello a 4 LV decresce da AS S1 ad AS S3. Questo può significare che l'aggiunta di variabili inutili rende più difficile la costruzione di variabili latenti significative per il modello, e questo effetto è tanto più marcato quanto più alto è il numero di variabile latente considerato.

3. per le prime 2/3 variabili latenti, un'oculata scelta delle variabili predittive migliora sempre la quantità di varianza spiegata. Il motivo è dovuto al fatto che il modello PLS riesce a sintetizzare l'informazione di più variabili predittive utili e quindi a costruire le direzioni principali con maggiore sicurezza;
4. i modelli che hanno il miglior rapporto varianza spiegata / MSEP sono VIP4, SROV4 S1 e SROV2 S2 per il numero di acidità e AS2 S4, AS3 S4, VIP2 e VIPP+4; in particolare, il modello VIP4 sopravanza nettamente tutti gli altri, nonostante i modelli analoghi a 2 e 3 variabili latenti fossero decisamente peggiori degli altri; probabilmente la selezione operata con il metodo VIP è così conservativa che si deve utilizzare tutta l'informazione contenuta nelle variabili predittive, e quindi tutte le variabili latenti utili per estrarre questa informazione, allo scopo di ottenere un modello ottimo. Inoltre il modello VIP4 supera anche il test di Shapiro-Wilk per la normalità dei residui e ciò permette di affermare, grossolanamente, che la regressione operata da questo modello è anche statisticamente significativa;
5. non esiste alcun legame tra la bontà della modellazione e il numero di variabili latenti selezionabili attraverso la *cross–validation*. Il confronto è stato effettuato utilizzando dati di convalida per MSEP e di calibrazione per RMSECV.

In conclusione, si riporta in Figura 4.7 il profilo di NA e MU per il modello PLS a 23 variabili e per il modello PLS-VIP a 4 LV:



(a) *(b)*

Figura 4.7. *Profilo di (a) numero di acidità e (b) viscosità per il batch di convalida no. 410 rispetto al modello PLS a 23 variabili predittive e al miglior modello ridotto rispetto alla prima fase, VIP-PLS. In nero sono riportati i punti sperimentali con un errore medio di misura del 10%.*

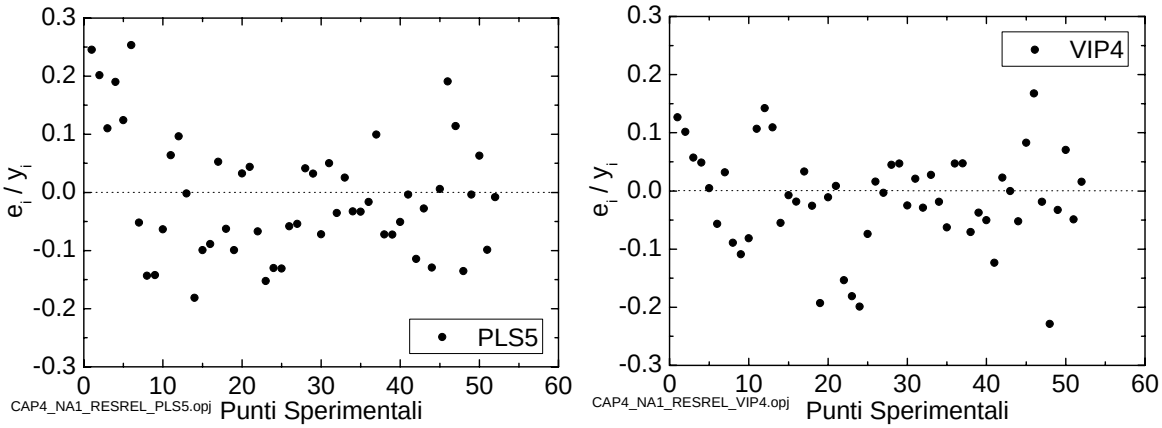
In tutti e due i casi si ottiene un miglioramento della predizione. Il miglioramento è più netto per alti valori del numero di acidità e piccoli valori di viscosità, come si può vedere

anche nel grafico, non riportato, del valore sperimentale di y_{NA} e y_{MU} contro i valori calcolati, sempre delle stesse variabili.

Si nota che la stima della viscosità necessita di un numero inferiore di LV rispetto alle variabili latenti necessarie per stimare il numero di acidità nei modelli ridotti: in Figura 4.6*b* è evidente che i modelli aventi un numero di LV inferiore sono migliori dei modelli con più LV. Ciò significa che la viscosità soffre molto di più del numero di acidità la presenza di rumore, e ciò influisce sul diverso errore relativo medio di predizione riscontrato nel modello a 23 variabili predittive. Si può notare come tutti i modelli PLS ridotti per la stima della viscosità riescano a migliorare la predizione indipendentemente dal numero di variabili predittive contenute: un modello a poche variabili latenti, infatti, non è influenzato dalle variabili predittive tanto quanto un modello a più variabili latenti, perché le prime direzioni principali scelte sono comunque le direzioni di massima varianza, che dipendono solitamente da un ristretto numero di variabili predittive che si trova in tutti i modelli. Con un modello a sole 2 LV è possibile inoltre ottenere stime a minore varianza: la dispersione dei punti calcolati per la stima della viscosità, in Figura 4.7*b*, è molto inferiore alla dispersione in figura 4.7*a*.

Inoltre, il profilo del numero di acidità è molto meno lineare del profilo di viscosità, e questo può significare che sono necessarie più variabili latenti lineari per rappresentarlo.

I grafici dei residui relativi sono riportati nella Figura 4.8; la distribuzione dei residui è normale, quindi il modello potenzialmente ha validità statistica.



(a) *(b)*

Figura 4.8. *Grafici dei residui relativi dei dati sperimentali del numero di acidità per i modelli (a) PLS a 5 LV e (b) PLS-VIP a 4 LV.*

Anche i grafici che riportano in ascissa i valori sperimentali e in ordinata i valori dei residui relativi (y_i vs e_i/y_i) mostrano una distribuzione omogenea.

I valori residui si riferiscono a tutti i *batches* utilizzati in fase di convalida, e quindi può essere opportuno fare alcune considerazioni: i primi 5 – 6 residui sono positivi per

entrambe le figure e si riferiscono allo stesso *batch* (no. 714); quindi non tutti i *batches*, singolarmente presi, hanno distribuzione normale (o almeno intorno allo zero) dei residui. Essendo il *batch* no. 714 "anomalo" questo risultato era atteso, anche se questo rende ancora più difficile la ricerca di un'equazione di regressione stabile rispetto a tutte le possibili evoluzioni di ogni singolo *batch*. Si deve però dire che la predizione, anche in questo *batch*, è molto buona, anche se evidentemente presenta tendenza a sovrastimare.
Non tutti i modelli ottenuti, inoltre, presentano distribuzione normale dei residui.
L'analisi dei grafici y_i vs e_i/y_i mostra che il modello PLS-VIP tende a sottostimare la predizione per valori bassi del numero di acidità e a sovrastimarla per valori elevati; questo *trend* è però appena accennato. Risultati analoghi si ottengono con i residui dei dati sperimentali di viscosità rispetto ai modelli PLS a 5 LV e PLS-VIP a 2 LV, che non vengono qui riportati.
Un ulteriore vantaggio nell'utilizzo di un sottoinsieme ridotto di variabili per la costruzione di modelli PLS si osserva costruendo un grafico degli *scores* della prima variabile latente di **X** e **Y** per il modello ridotto ottimo: rispetto alle Figure 4.3 e 4.4, i dati sono molto più compatti e allineati, segno che il modello coglie in modo più preciso l'andamento dei dati e riesce più efficacemente a separare il rumore dall'informazione utile.

4.4 Seconda Fase

La seconda fase presenta maggiori difficoltà rispetto a tutte le altre per ciò che riguarda la modellazione: anche l'errore medio di predizione è molto più alto rispetto alle altre fasi.
Si riporta in Tabella 4.4 il numero di variabili selezionate da ogni algoritmo. Rispetto alla prima fase, si nota che:

1. gli algoritmi tendono a scegliere le variabili in maggiore accordo, nonostante la seconda fase sia più difficile da modellare.
2. il metodo VIP+ seleziona esattamente le stesse variabili sia per la stima di NA che di MU, mentre VIP ha una variabile di differenza, statisticamente al limite della significatività (no. 23, temperatura del vuoto nel reattore; $VIP_{23} = 0.9$);
3. ci si aspetta che il secondo set selezionato con l'algoritmo SROV si comporti peggio del precedente, perché seleziona variabili inutili come il numero i giri del reattore (no. 1 e 2);
4. ancora una volta, gli algoritmi AS e *Stepwise* selezionano lo stesso tipo di variabili;
5. nel metodo VIP le variabili temperatura dell'acqua in ingresso (no. 9) e temperatura della "via breve" (no. 13) sostituiscono le tre temperature disponibili del reattore (6, 12 e 21). Quindi il processo in questa fase non è più dominato dalla reazione ma dal processo di depressurizzazione. L'algoritmo coglie bene questa situazione.
6. anche gli altri algoritmi di selezione includono le variabili 9 e 13.

Tabella 4.4. *Variabili selezionate da ogni algoritmo per la costruzione dello stimatore software per il numero di acidità. Le variabili selezionate con l'algoritmo AS nel primo set (S1) sono le stesse scelte con il metodo Stepwise Regression. Il set SROV S2 presenta le variabili in ordine numerico perché è il risultato di una rotazione e nuova selezione di variabili in cui tutte sono teoricamente selezionate come ultime.*

Algoritmo di selezione		Variabili Predittive Selezionate
AS	S1	10, 9, 23, 7, 13
	S2	10, 9, 23, 7, 13, 14
	S3	10, 9, 23, 7, 13, 14, 17, 5, 20, 18
	S4	10, 9, 23, 7, 13, 14, 17, 5, 20, 18,3, 12, 2, 1, 16, 15
SROV	S1	10, 9, 23, 7, 13, 14, 5, 20, 18, 12, 3, 16, 22, 6, 15, 4
	S2	1, 2, 3, 5, 7, 9, 10, 12, 13, 14, 15, 16, 18, 20, 22, 23
VIP	VIP	10, 7, 17, 9, 13, 23
	VIP+	10, 7, 17, 9, 13, 23, 22, 8, 11, 3

Tabella 4.5. *Variabili selezionate da ogni algoritmo per la costruzione dello stimatore software per la viscosità. Le variabili selezionate con l'algoritmo AS nel secondo set (S2) sono le stesse scelte con il metodo Stepwise Regression. Il set SROV S2 presenta le variabili in ordine numerico perché è il risultato di una rotazione e nuova selezione di variabili in cui tutte sono teoricamente selezionate come ultime.*

Algoritmo di selezione		Variabili Predittive Selezionate
AS	S1	10, 9, 23, 7, 14
	S2	10, 9, 23, 7, 14, 13, 4
	S3	10, 9, 23, 7, 14, 13, 4, 17, 3, 2
	S4	10, 9, 23, 7, 14, 13, 4, 17, 3, 2, 1, 18, 20, 21
SROV	S1	10, 9, 23, 7, 14, 13, 4, 2, 16, 18, 20, 12, 15, 3
	S2	2, 3, 7, 9, 10, 12, 13, 14, 16, 18, 19, 20, 21, 23
VIP	VIP	10, 7, 17, 9, 13
	VIP+	10, 7, 17, 9, 13, 23, 22, 8, 11, 3

Nel diagramma dei *loadings* in Figura 4.9 la maggior importanza di queste due variabili è palese, perché si portano dal centro del diagramma, dove si trovavano durante la prima fase, nel quarto quadrante; inoltre, essendo vicine, portano lo stesso contributo al modello.
Anche le variabili pressione del vuoto, temperatura dello *scrubber* e del fondo *scrubber* (3, 11 e 15) si spostano nella stessa zona del diagramma, mentre prima si trovavano al centro.
Viceversa, le variabili 6, 12 e 21 si spostano al centro del diagramma.
Invece di riportare i valori dello scarto quadratico medio di tutti i modelli calcolati, si preferisce indicare il migliore modello PLS per ogni metodo di selezione e poi confrontare tra loro questi modelli. Si precisa che non sempre il modello selezionato in base alla minimizzazione del criterio MSEP coincide con il modello indicato dalla *cross–validation*.

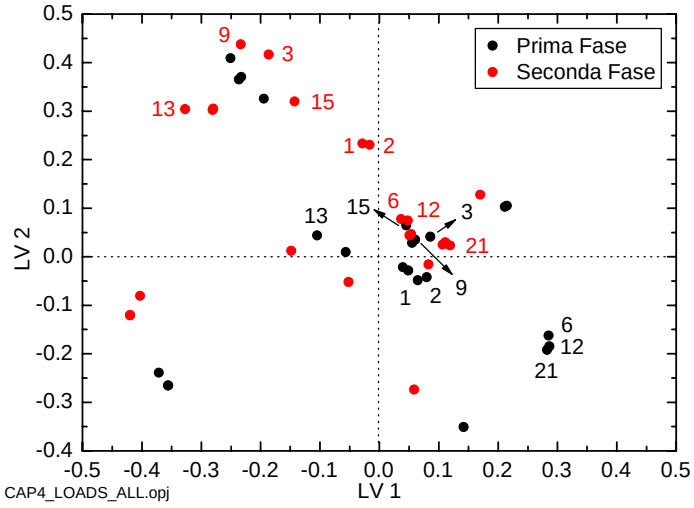


Figura 4.9. *Diagramma dei loadings delle prime due fasi. In nero sono indicate alcune variabili relative alla prima fase, in rosso alla seconda.*

Questa volta il modello migliore è determinato dal set di variabili AS S3 a 3 LV (stesso valore di MSEP di SROV S1 a 3 LV, ma molte variabili predittive in meno) per la stima del numero di acidità e dal set AS S4 a 3 LV per la stima della viscosità, anche se quasi tutti i modelli riescono a migliorare le prestazioni del modello PLS a 5 LV.

Se la correlazione tra variabili predittive e dipendenti non è molto forte, o la varianza spiegata della variabile dipendente non è elevata, anche con molte variabili latenti, diventa importante includere più variabili possibili nel modello, in modo da sfruttare tutta la capacità predittiva del modello.

Tabella 4.6. *Valori del criterio MSEP per tutti i modelli PLS elaborati per la stima (a) del numero di acidità e (b) della viscosità.*

Modello PLS	MSEP
PLS4	**0.00353**
AS2 S1	0.00315
AS2 S2	0.00313
AS3 S3	0.00309
AS3 S4	0.0032
SROV3 S1	0.00304
SROV3 S2	0.00312
VIP4	0.00322
VIP+3	0.00332

(a)

Modello PLS	MSEP
PLS2	**0.00263**
AS2 S1	0.00242
AS3 S2	0.00242
AS3 S3	0.00226
AS3 S4	0.00219
SROV3 S1	0.00245
SROV3 S2	0.00244
VIP2	0.00297
VIP+3	0.00243

(b)

Per questo stesso motivo i modelli costruiti sui *sets* di variabili selezionati con il metodo VIP non riescono a diminuire lo scarto quadratico medio come gli altri modelli: anche aumentando il numero di LV, i modelli costruiti utilizzando il metodo VIP spiegano solo una frazione troppo piccola della varianza di *y*, perché eliminano variabili predittive

importanti per la stima del numero di acidità e viscosità. In questo caso, un criterio meno restrittivo rispetto all'esclusione delle variabili predittive con valore $VIP_j < 1$ sarebbe consigliabile.

Tabella 4.7. *Valori di Varianza Spiegata (R_Y^2) per tutti i modelli PLS elaborati per la stima (a) del numero di acidità e (b) della viscosità.*

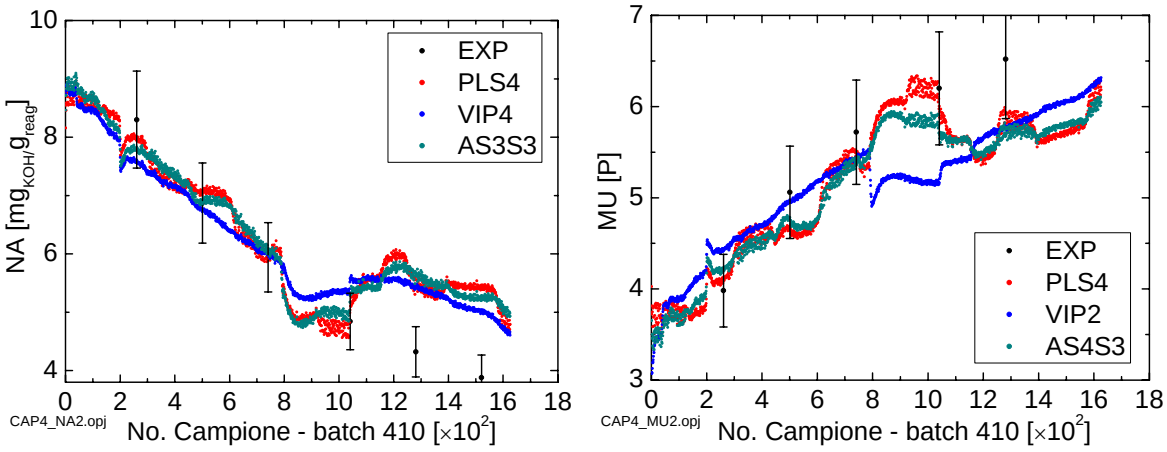
Modello PLS	R_Y^2 [%]
PLS5	**81.857**
AS2 S1	82.296
AS2 S2	82.611
AS3 S3	81.503
AS3 S4	81.467
SROV3 S1	83.113
SROV3 S2	82.87
VIP4	84.336
VIP+3	80.473

(a)

Modello PLS	R_Y^2
PLS4	**77.292**
AS2 S1	78.106
AS3 S2	78.962
AS3 S3	77.043
AS3 S4	76.436
SROV3S1	78.482
SROV3S2	78.555
VIP2	71.47
VIP+3	75.011

(b)

Si riportano, nella Figura 4.10, i valori calcolati per il *batch* 410 dei modelli AS S3 a 3 LV e VIP4 per la stima del numero di acidità e AS S4 a 3 LV e VIP2 per la stima della viscosità.



(a) *(b)*

Figura 4.10. *Profilo di (a) numero di acidità e (b) viscosità per il batch di convalida no. 410 rispetto al modello PLS a 23 variabili predittive e ai modello ridotti VIP-PLS e AS S3 rispetto alla seconda fase.*

Nonostante il maggiore errore, i dati calcolati rientrano ancora all'interno dell'errore sperimentale, tranne negli ultimi due punti di entrambe le figure, dove probabilmente una variabile predittiva assume maggiore influenza e costringe la traiettoria a mutare

andamento. In tutte e due le figure, intorno al punto 200, si assiste ad un "salto" dei valori predetti: la variabile associata a questo salto è la temperatura del fondo della colonna (no. 10), che ripete un analogo "salto" intorno al punto 800. E' così spiegato il motivo della variabilità nell'andamento delle variabili calcolate, anche se rimane da capire se questa variazione improvvisa di temperatura è plausibile nel processo o è dovuta solo ad errori di misura.

I residui relativi non sono statisticamente significativi, perché alcuni *batches* tendono a generare errori molto autocorrelati ed elevati in valore assoluto. Isolando i valori residui dei singoli *batches*, si vede che, nonostante la predizione non ottima, i residui dei dati sperimentali si distribuiscono omogeneamente intorno al valore nullo per alcuni *batches* buoni, come il no. 410.

Anche i valori dei residui relativi riportati contro i valori sperimentali (y_i vs e_i/y_i) mostrano una distribuzione non omogenea, dovuta sempre all'autocorrelazione tra i residui.

4.5 Terza Fase

La terza fase del processo è più facile da modellare rispetto alla seconda, perché l'andamento della viscosità e del numero di acidità è più lineare.

Le variabili selezionate da ogni algoritmo sono riportate nelle Tabelle 4.8 e 4.9: in questo caso si è rinunciato a selezionare un set di variabili con il metodo VIP+, perché non era possibile distinguere un valore soglia, inferiore a 1, per distinguere tra le variabili importanti e le variabili da scartare.

Tabella 4.8. *Variabili selezionate da ogni algoritmo per la costruzione dello stimatore software per il numero di acidità. Le variabili selezionate con l'algoritmo AS nel secondo set (S2) sono le stesse scelte con il metodo Stepwise Regression, mentre le variabili selezionate con l'algoritmo AS nel terzo set (S3) equivalgono a quelle selezionate dall'algoritmo SROV nel primo set. Il set SROV S2 presenta le variabili in ordine numerico perché è il risultato di una rotazione e nuova selezione di variabili in cui tutte sono teoricamente selezionate come ultime.*

Algoritmo di selezione		Variabili Predittive Selezionate
AS	S1	21, 17, 13, 10, 2
	S2	21, 17, 13, 10, 2, 9, 15, 5
	S3	21, 17, 13, 10, 2, 9, 15, 5, 19, 1, 8, 11
SROV	S2	1, 2, 5, 7, 8, 9, 10, 11, 13, 15, 19, 21
VIP	VIP	6, 12, 21, 17, 7

Tabella 4.9. *Variabili selezionate da ogni algoritmo per la costruzione dello stimatore software per la viscosità. Le variabili selezionate con l'algoritmo AS nel secondo set (S2) sono le stesse scelte con il metodo Stepwise Regression. Il set SROV S2 presenta le variabili in ordine numerico perché è il risultato di una rotazione e nuova selezione di variabili in cui tutte sono teoricamente selezionate come ultime.*

Algoritmo di selezione		Variabili Predittive Selezionate
AS	S1	21, 2, 17, 13, 8, 10
	S2	21, 2, 17, 13, 8, 10, 1, 11
	S3	21, 2, 17, 13, 8, 10, 1, 11, 14, 9, 18, 5, 23, 22
SROV	S1	21, 2, 17, 13, 8, 10, 1, 11, 14, 9, 18, 5, 23, 4
	S2	1, 2, 7, 8, 9, 10, 11, 13, 14, 18, 20, 21, 22, 23
VIP	VIP	6, 12, 21, 17, 7, 13

Anche in questa ultima fase le variabili selezionate da ogni algoritmo sono molto simili tra loro; a differenza delle altre fasi, però, la prima variabile selezionata non è più la temperatura del fondo colonna (no. 10), ma una delle temperature del reattore (no. 21), a significare che la situazione critica del processo, cioè la disidratazione della miscela polimerica, si è spostata dalla colonna a corpi di riempimento al reattore, mediante l'aumento della temperatura interna.

E' invece preoccupante il fatto che venga selezionata al secondo posto la variabile che riporta i giri al minuto del reattore (no. 2), la quale è totalmente insignificante rispetto al processo: questo può significare che ormai le caratteristiche di qualità del prodotto sono decise ed è quindi insignificante cercare di inferire tali qualità dalle misure attuali dei sensori del processo. Tale conclusione, inoltre, sarebbe in accordo con il comportamento integrativo del processo *batch*.

Per confermare questa ipotesi si è provato ad effettuare uno studio in cui si utilizzavano misure relative alla sola prima fase per calibrare uno stimatore valido per tutte le fasi del processo: la qualità delle previsioni si è rivelata inferiore a quella ottenuta dai singoli stimatori dedicati ad ogni fase, ma è importante notare che comunque veniva colto il trend presente in ogni *batch*. Probabilmente, la linearizzazione delle variabili dipendenti potrebbe far ottenere uno stimatore più preciso, nonostante durante il processo la correlazione tra variabili si modifichi.

Gli algoritmi di tipo *stepwise* sono ancora una volta in accordo sull'ordine di scelta delle variabili.

Si confrontano ora i migliori modelli ottenuti da ogni algoritmo attraverso lo scarto quadratico medio di predizione e la varianza spiegata, in Tabella 4.10 e 4.11.

Tabella 4.10. *Valori del criterio MSEP per tutti i modelli PLS elaborati per la stima (a) del numero di acidità e (b) della viscosità.*

Modello PLS	MSEP [$\times 10^3$]
PLS3	**1.2016**
AS3 S1	1.3125
AS3 S2	1.2564
AS3 S3	1.2591
SROV4 S2	1.2137
VIP3	1.2934

(a)

Modello PLS	MSEP [$\times 10^3$]
PLS3	**4.5926**
AS2 S1	7.8725
AS2 S2	4.2213
AS2 S3	4.499
SROV2S1	4.6816
SROV3S2	4.1935
VIP2	3.614

(b)

Tabella 4.11. *Valori della varianza spiegata (R_Y^2) per tutti i modelli PLS elaborati per la stima (a) del numero di acidità e (b) della viscosità.*

Modello PLS	R_Y^2 [%]
PLS4	**58.792**
AS3 S1	59.13
AS3 S2	59.6
AS3 S3	58.349
SROV4 S2	59.909
VIP3	51.397

(a)

Modello PLS	R_Y^2 [%]
PLS3	**47.713**
AS2 S1	47.186
AS2 S2	46.797
AS2 S3	45.363
SROV2S1	45.387
SROV3S2	50.336
VIP2	39.882

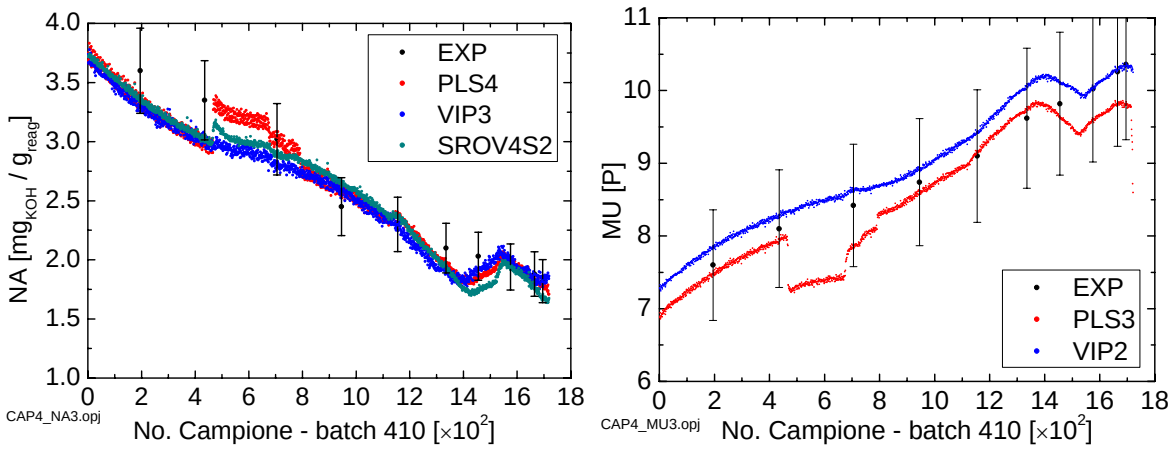
(b)

Si nota che:

1. la migliore predizione del modello SROV S2 a 3 LV per la stima della viscosità è dovuta alla maggior quantità di varianza spiegata, inclusa senza incorporare rumore nel modello;
2. il modello VIP a 2 LV è in grado di minimizzare il criterio MSEP nonostante includa meno varianza di tutti. Questo significa che è in grado di selezionare, in questa fase, proprio le variabili migliori; non a caso, quindi, esclude le variabili come la no. 2 e la no. 13, non più determinante in questa fase del processo. Anche nel modello PLS a 3 LV, nonostante non sia il migliore, si ha una buona performance con una quantità di varianza spiegata minima;
3. i modelli ridotti per la stima del numero di acidità non sono migliori del modello completo; viceversa, alcuni modelli ridotti per la stima della viscosità sono migliori del modello completo;

4. quando la varianza complessiva spiegata è bassa, come in questo caso, selezionare variabili per un modello ridotto è problematico e spesso conviene utilizzare un criterio cautelativo, includendo più variabili possibili, come nella seconda fase;
5. la maggiore importanza della temperatura del reattore è visibile anche nel diagramma dei *loadings* dove la variabile si sposta lontano dal centro.

Si riportano, infine, i valori calcolati dei modelli selezionati per la stima di ogni variabile dipendente in Figura 4.11: i modelli scelti sono SROV S2 a 4 LV e VIP a 3 LV per il numero di acidità e il solo VIP a 2 LV per la viscosità:



(a) *(b)*

Figura 4.11. *Stime dei valori di (a) numero di acidità e (b) viscosità per i modelli PLS selezionati.*

La stima della viscosità è più precisa e i dati presentano minore dispersione rispetto ai valori ottenuti per la stima del numero di acidità.

Anche in questo caso si assiste ad uno scarto improvviso delle variabili calcolate dovuto all'inclusione della variabile no. 10 (temperatura di fondo colonna) che bruscamente passa da 54°C a 64°C; il modello VIP, che non include questa variabile, presenta un andamento più lineare, anche se leggermente più disperso rispetto al modello ridotto SROV S2 a 4 LV. I residui sono statisticamente significativi e superano il test di Shapiro-Wilk: ciò significa che i modelli costruiti hanno buona possibilità di essere statisticamente significativi.

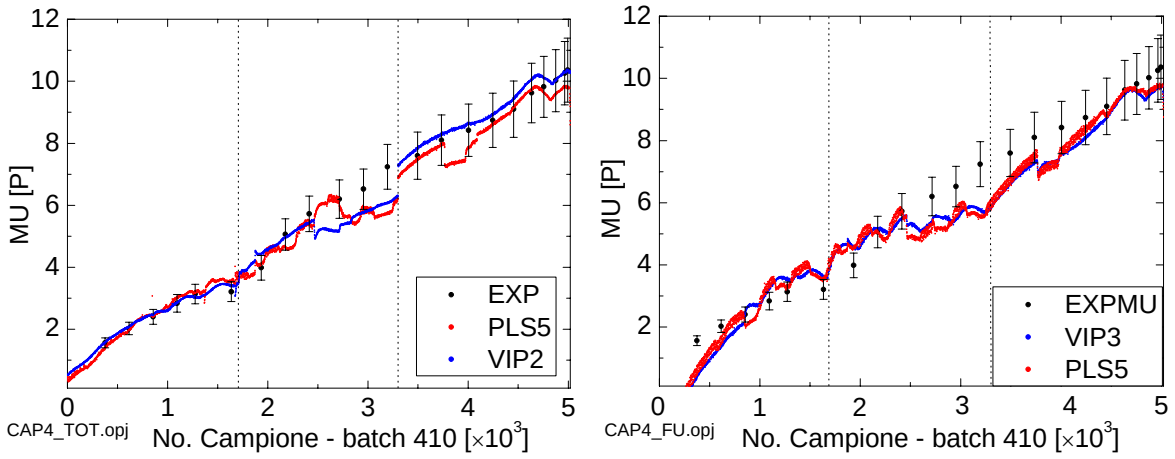
4.6 Conclusioni

Si è visto che la selezione di variabili è in grado di migliorare le prestazioni del modello PLS, se questo viene inizialmente costruito in modo opportuno mediante la divisione in fasi. La selezione di variabili, operata sul modello globale, non riesce a portare lo stesso miglioramento che si ottiene operando sulle singole fasi, così come visto nel § 3.

Per effettuare un confronto, si cerca di costruire uno stimatore basato su un unico *set* di variabili predittive per tutte le fasi del processo: tale *set* viene costruito utilizzando le informazioni ottenute dai *sets* ottenuti per le singole fasi utilizzando l'algoritmo VIP, perché l'applicazione diretta dell'algoritmo porta a includere troppe variabili troppo diverse da quelle selezionate in ogni fase, a causa della variazione della correlazione delle variabili nelle diverse fasi del processo.

Le variabili selezionate sono quindi: temperature del reattore (No. 6, 12, 21), presenti solo nella prima e nella terza fase; temperature della colonna, del fondo e della testa (No. 7, 10, 17), che compaiono in tutte e tre le fasi; temperatura del vapore nella via breve (No. 13), selezionate per le ultime due fasi.

I modelli che si ottengono unendo le stime per il modello trifasico e monofasico con una selezione empirica o manuale delle variabili sono riportati in Figura 4.12; si è scelto di rappresentare la sola viscosità perché ciò permette di utilizzare un unico modello PLS, selezionato come ottimo in tutte e tre le fasi, cioè VIP a 2 LV.



(a) *(b)*

Figura 4.12. *Confronto del modello (a) trifasico e (b) monobasico con selezione di variabili basata sull'algoritmo VIP.*

Il miglioramento che si ottiene selezionando le variabili è netto sia per quanto riguarda la diminuzione della varianza, sia per ciò che concerne la diminuzione della dispersione dei dati intorno alla traiettoria ideale. Tale miglioramento è più evidente se il processo viene diviso in fasi.

Si passa quindi al confronto tra gli algoritmi di selezione. Tutti e 4 gli algoritmi proposti sono stati in grado di selezionare un sottoinsieme di variabili predittive; inoltre, tutti e 4 gli algoritmi hanno consentito di ottenere un modello di regressione che diminuisse lo scarto quadratico medio di predizione in almeno una fase rispetto al modello calibrato su una delle variabili dipendenti.

Per quanto riguarda la minimizzazione del criterio MSEP e la scelta del numero e del tipo di variabili selezionate si possono fare delle considerazioni più approfondite:

1. l'algoritmo *Stepwise regression* nella prima fase ha uno scarto quadratico medio molto elevato a causa del mancato inserimento di una variabile; ciò impedisce di poter considerare questo algoritmo come affidabile in senso generale;
2. l'algoritmo SROV riesce spesso a minimizzare lo scarto quadratico medio, ma tende a inserire troppe variabili nel sottoinsieme ridotto. Molte di queste variabili, inoltre, sono inutili dal punto di vista del processo. Anche l'algoritmo SROV non può quindi essere utilizzato con sicurezza;
3. entrambi i modelli gli algoritmi, AS e VIP, riescono ad ottenere quasi sempre un miglioramento delle stime delle variabili di qualità. L'algoritmo AS, però, ottiene la riduzione del criterio MSEP utilizzando un numero di variabili predittive molto diverso da fase a fase e per le due variabili dipendenti: per la prima fase, ad esempio, la minimizzazione si ottiene con poche variabili predittive per stimare il numero di acidità (*set* AS S1, 6 variabili incluse) e molte per stimare la acidità (*set* AS S4,14 variabili incluse). L'algoritmo, cioè, soffre dell'assenza di un criterio per determinare il numero di variabili ottime da includere nel sottoinsieme;
4. l'unico algoritmo che si comporta in modo "univoco" in ogni fase e per ogni modello di stimatore e VIP: seleziona sempre un numero conservativo di variabili predittive, cioè inferiore rispetto agli altri algoritmi, e le variabili selezionate sono sempre significative per il processo. Inoltre, questo riesce quasi sempre a diminuire lo scarto quadratico medio rispetto al modello calibrato con tutte le variabili di processo; il minimo del criterio MSEP si ottiene quasi sempre con lo stesso numero di LV anche tra fasi diverse (4 LV per la stima del numero di acidità, 2 LV per la stima della viscosità).

L'algoritmo che si può ritenere più affidabile rispetto alla selezione di variabili predittive appropriate in numero e tipologia è, in questo caso, VIP.

Considerazioni conclusive

In questa Tesi ci si è occupati della selezione di variabili di processo significative per la costruzione di sensori *software* più robusti e meno soggetti all'avaria dei sensori di processo, in modo da ottenere stime più precise e continue della variabili di qualità non misurate in linea.

Gli algoritmi sono stati applicati su due *sets* di dati: il primo, ottenuto mediante simulazione, rappresentava un processo di distillazione binaria sottoposto a varie tipologie di disturbi di portata di riflusso, vapore ed alimentazione ed era corrotto da un rumore bianco addizionato ai dati prodotti mediante simulazione; il secondo, invece, era costituito da dati sperimentali ottenuti da un impianto industriale di produzione di una resina.

Entrambi i *sets* sono stati utilizzati per costruire stimatori inferenziali basati su modelli PLS contenenti diversi sottoinsiemi di variabili predittive selezionate con i 4 algoritmi considerati. Questi modelli sono stati poi confrontati tra loro e con il modello PLS contenente tutte le variabili predittive utilizzando dei dati di convalida.

I metodi di selezione di variabili possono essere divisi in conservativi, i quali tendono a selezionare un numero ristretto di variabili predittive (VIP), e non conservativi (*Stepwise regression*). Il metodo SROV e il metodo AS, le cui scelte dipendono fortemente dai parametri impostati dall'operatore, non sono stati classificati.

In particolare, la selezione di variabili effettuata con il test *F* (*Stepwise Regression*), si è dimostrata la meno affidabile in quanto non è in grado di distinguere il rumore di misura dal segnale e di selezionare di conseguenza le variabili predittive. Per questo la maggior parte delle volte conduce ad *overfitting*, oppure non è in grado di selezionare nessun sottoinsieme di variabili utili. Anche l'algoritmo SROV non è stato in grado di selezionare un sottoinsieme utile di variabili predittive con i dati simulati di distillazione.

Gli algoritmi AS, SROV e *Stepwise regression* hanno selezionato spesso le variabili nello stesso ordine, ma in numero diverso, nonostante il metodo AS non disponga di un criterio di stop, segno che il criterio di entrata basato sul coefficiente α per il test *F* è forse troppo basso se le variabili predittive sono affette da rumore.

Di contro, l'algoritmo SROV fornisce criteri troppo restrittivi, che lo rendono inapplicabile nella pratica industriale: infatti la sua corretta applicazione richiede la disponibilità di dati molto precisi per riuscire a costruire un'equazione di regressione adeguata, in quanto tale metodo tende a scartare troppe variabili a causa dell'elevato valore di soglia del criterio TNR, corretto dal punto di vista numerico, ma limitante in molti casi.

L'utilizzo pratico dell'algoritmo, inoltre, ha comportato alcune difficoltà: a volte la rotazione delle variabili dopo la prima selezione iterava in *loop* continuo, senza scegliere un modello finale definitivo. Inoltre, la rotazione delle variabili non sempre porta una ulteriore minimizzazione dello scarto quadratico medio di predizione; in generale, però, si può notare che il *set* scelto per primo dall'algoritmo trova un minimo di MSEP per un numero di LV superiore al *set* finale.

L'algoritmo VIP è stato l'unico a selezionare sempre un sottoinsieme di variabili tutte significative per il processo, in numero omogeneo per ogni stimatore costruito rispetto a fasi o variabili di qualità diverse; in un solo caso, nella stima della composizione del residuo della distillazione, non è stato in grado di costruire uno stimatore migliore del modello contenente tutte le variabili predittive, in quanto il numero di variabili selezionate (2) era probabilmente troppo piccolo. Ciò è dovuto al criterio di ingresso utilizzato, che è costretto ad operare con valori di importanza delle variabili predittive normalizzati, che non riflettono l'importanza assoluta della variabile ma solo l'importanza relativa rispetto alle altre variabili: se esistono variabili predittive molto più correlate di altre con la variabile dipendente, solo queste verranno selezionate, escludendo le variabili meno correlate ma comunque portatrici di informazione utile.

Si potrebbe quindi cercare di migliorare l'algoritmo VIP in modo da rendere i valori di importanza delle variabili predittive, cioè i *loadings*, non soggetti al vincolo di norma unitaria del vettore *loading*.

Infine, l'algoritmo AS è stato in grado di selezionare sempre variabili significative per il processo, anche se a volte includeva variabili inutili, a causa dell'algoritmo che tende ad includere variabili tra loro ortogonali nel modello ridotto. L'unico inconveniente di questo modello è l'assenza di un criterio affidabile di interruzione nell'inserimento di variabili: non potendo utilizzare la varianza totale spiegata della variabile dipendente, in quanto non è nota la quantità di rumore contenuta nei dati, diventa importante riuscire a sviluppare un criterio di ingresso delle variabili predittive basato sul rapporto segnale/rumore di ogni variabile (come nell'algoritmo SROV) o di ogni variabile latente (come per la *cross - validation*).

Indipendentemente dall'algoritmo utilizzato, si sono dimostrati migliori i sottoinsiemi di variabili predittive che riuscivano a spiegare, in ogni variabile latente del modello PLS, una quantità analoga di varianza di **X** e **Y**, in modo che l'algoritmo non costruisse LV dedicate esclusivamente alla modellazione di **X** o di **Y**. Inoltre, la quantità totale di varianza spiegata di **Y** non ha mai rappresentato un discriminante per la bontà dei modelli: anche se un modello spiegava al massimo solo l'80% della varianza di Y poteva essere migliore di un modello che, con lo stesso numero di LV, modellava quasi il 100%, includendo di fatto il rumore.

In conclusione, si è ritenuto preferibile rispetto agli altri l’algoritmo VIP, a causa della sua stabilità nel selezionare variabili di processo opportune per tipologia e numero, e per la sua capacità di costruire modelli PLS che ottengano stime affidabili di qualità del prodotto.

Appendice A

Figure contenute nella Tesi

Le Figure presenti nella Tesi sono reperibili nelle sottocartelle contenute nella cartella principale `\Tesi\Figure` .

A.1 Figure inserite nel Capitolo 3

Le Figure indicate in Tabella A.1 sono contenute nella cartella `\Tesi\Figure\Capitolo 3`.

Tabella A.1 *Figure del Capitolo 3.*

Figura	File grafico
Figura 3.1	Colonna_lab.vsd
Figura 3.2*a* Figura 3.2*b* Figura 3.2*c* Figura 3.2*d*	CAP3_PORTATA.opj CAP3_TEMP.opj CAP3_DENS.opj CAP3_FRAZMOLARI.opj
Figura 3.3*a* Figura 3.3*b*	CAP3_DENS_RUM.opj CAP3_FRAZDIST_RUM.opj
Figura 3.4*a* Figura 3.4*b*	CAP3_T30_RUM.opj CAP3_T4_RUM.opj
Figura 3.5*a* Figura 3.5*b*	CAP3_XB_PLS6.opj CAP3_XD_PLS6.opj
Figura 3.6	CAP3_MSE.opj
Figura 3.7	CAP3_LOADS_LV2.opj
Figura 3.8	CAP3_LOADINGS.opj
Figura 3.9*a* Figura 3.9*b*	CAP3_XB_VIP_PLS.opj CAP3_XD_VIP_PLS.opj
Figura 3.10	CAP3_XSCORES_PLOT.opj
Figura 3.11	CAP3_VIP_LOADINGS.opj
Figura 3.12*a* Figura 3.12*b*	CAP3_XB_MURA_PLS5.opj CAP3_XD_MURA_PLS5.opj
Figura 3.13	CAP3_MURA_MSEP.opj
Figura 3.14	CAP3_MURA_LOADINGS.opj
Figura 3.15	CAP3_XB_PLS5_XB.opj
Figura 3.16	CAP3_XB_VIP_XB.opj
Figura 3.17	CAP3_XB_AS2_XB.opj
Figura 3.18	CAP3_XD_PLS5_XD.opj
Figura 3.19	CAP3_XD_VIP4_XD.opj
Figura 3.20	CAP3_XD_AS_XD.opj

A.2 Figure inserite nel Capitolo 4

Le Figure riportate in Tabella A.2 sono contenute nella cartella `\Tesi\Figure\Capitolo 4`.

Tabella A.2 *Figure del Capitolo 2.*

Figura	File grafico
Figura 4.1	Imp_resina.vsd
Figura 4.2*a* Figura 4.2*b* Figura 4.2*c*	CAP4_YSC_FM_F1_MU.opj CAP4_YSC_FM_F2_MU.opj CAP4_YSC_FM_F3_MU.opj
Figura 4.3*a* Figura 4.3*b* Figura 4.3*b*	CAP4_XYSC_FM_F1_MU.opj CAP4_XYSC_FM_F2_MU.opj CAP4_XYSC_FM_F3_MU.opj
Figura 4.4	CAP4_XYSC_FM_F1_NA.opj
Figura 4.5*a* Figura 4.5*b* Figura 4.5*c* Figura 4.5*d*	CAP4_NA1_LOADS.opj CAP4_NA1_LOADS_AS.opj CAP4_NA1_LOADS_SROV.opj CAP4_NA1_LOADS_VIP.opj
Figura 4.6*a* Figura 4.6*b*	CAP4_NA1_MSEP.opj CAP4_MU1_MSEP.opj
Figura 4.7*a* Figura 4.7*b*	CAP4_NA1.opj CAP4_MU1.opj
Figura 4.8*a* Figura 4.8*b*	CAP4_NA1_RESREL_PLS5.opj CAP4_NA1_RESREL_VIP4.opj
Figura 4.9	CAP4_LOADS_ALL.opj
Figura 4.10*a* Figura 4.10*b*	CAP4_NA2.opj CAP4_MU2.opj
Figura 4.11*a* Figura 4.11*b*	CAP4_NA3.opj CAP4_MU3.opj
Figura 4.12*a* Figura 4.12*b*	CAP4_TOT.opj CAP4_FU.opj

Nomenclatura

a	=	numero generico di componenti principali considerate (-)
A	=	numero totale di componenti principali considerate (-)
AS	=	metodo di selezione di variabili Algoritmo di Selezione
$\mathbf{b}$	=	vettore dei coefficienti di regressione (-)
$\mathbf{b_{PLS}}$	=	vettore dei coefficienti di regressione della relazione interna nel modello PLS
b_j	=	elemento generico del vettore dei coefficienti di regressione (-)
corr	=	correlazione (-)
cov	=	covarianza (-)
diag	=	diagonale di una matrice (-)
e	=	errore o residuo (-)
$\mathbf{e}$	=	vettore degli errori nel caso di regressione lineare (-)
E	=	valore atteso (-)
$\mathbf{E}$	=	matrice bidimensionale degli errori nei metodi statistici multivariati per la matrice $\mathbf{X}$ (-)
$\|\|\mathbf{E}\|\|$	=	norma della matrice $\mathbf{E}$ (-)
$\mathbf{E}_h$	=	matrice dei residui per $\mathbf{X}$ all'iterazione h (-)
e_i	=	elemento del vettore $\mathbf{e}$ (-)
$\mathbf{e}_i$	=	vettore riga della matrice dei residui $\mathbf{E}$ (-)
$\mathbf{e}_i^T$	=	vettore colonna della matrice dei residui $\mathbf{E}$ (-)
$e_{i,j}$	=	elemento della matrice $\mathbf{E}$ (-)
f	=	funzione generica (-)
f_0	=	valore del test – F (-)
$\mathbf{F}$	=	matrice bidimensionale degli errori nei metodi statistici multivariati per la matrice $\mathbf{Y}$ (-)
$\|\|\mathbf{F}\|\|$	=	norma della matrice $\mathbf{F}$ (-)
$F_{(n,n-k-1,\alpha)}$	=	distribuzione statistica F (-)
$\mathbf{F}_h$	=	matrice dei residui per $\mathbf{Y}$ all'iterazione h (-)
h	=	generico pedice ad indicare un elemento o un passo iterativo (-)
h_0	=	ipotesi nulla
i	=	indicatore generico o pedice generico(-)
$\mathbf{I}$	=	matrice identità (-)
j	=	indicatore generico di una variabile di processo (-)

k = numero totale delle variabili di processo misurate (-)
LV = variabile latente (-)
MSEC = *Mean Square Error of Calibration*
MSECV = *Mean Square Error of Cross - Validation*
MSEP = *Mean Square Error of Prediction*
MSR = *Mean Square of Regression*
MU = Viscosità [P]
n = numero totale di elementi generici (-)
NA = Numero di Acidità [mg_{KOH} / g_{reag}]
NIPALS = algoritmo iterativo non lineare per la proiezione su strutture latenti (-)
OLS = *Ordinary Least Squares*
$\mathbf{P}$ = matrice dei *loading* (-)
$\mathbf{P}^T$ = matrice dei *loading* trasposta (-)
PC = componente principale (-)
PCA = metodo dell'analisi delle componenti principali (-)
PCR = *Principal Component Regression* (-)
$\mathbf{p}_j$ = vettore colonna della matrice dei *loading* $\mathbf{P}$
$\mathbf{p}_h^T$ = vettore riga della matrice dei *loading* $\mathbf{P}$ (-)
$||\mathbf{p}_h||$ = norma del vettore riga della matrice dei *loading* (-)
$\mathbf{p}_{h,new}$ = vettore $\mathbf{p}_h$ aggiornato all'iterazione h (-)
$\mathbf{p}_{h,old}$ = vettore $\mathbf{p}_h$ calcolato all'iterazione h (-)
$p_{i,j}$ = elemento della matrice dei *loading* $\mathbf{P}$ (-)
PLS = metodo della proiezione su strutture latenti (-)
$PRESS_j$ = errore di predizione sulla somma dei quadrati dei residui (-)
$\mathbf{Q}$ = matrice dei *loading* delle variabili di qualità del processo (-)
$\mathbf{Q}^T$ = matrice trasposta di $\mathbf{Q}$ (-)
$\mathbf{q}_h$ = vettore dei *loading* delle variabili di qualità del processo (-)
$\mathbf{q}_h^T$ = vettore trasposto di $\mathbf{q}_h$ (-)
$||\mathbf{q}_h||$ = norma di $\mathbf{q}_h$ (-)
Q_i = errore di predizione al quadrato (-)
r = rango di una generica matrice (-)
$RMSEC$ = *Root-Mean Square Error of Calibration* (-)
$RMSECV_j$ = *Root-Mean-Square Error of Cross-Validation* (-)
R^2 = varianza totale spiegata (-)
R^2_{adj} = varianza totale spiegata aggiustata (-)
s = stima della deviazione standard
SROV = metodo di selezione di variabili *Stepwise regression on Orthogonalized Variable* (-)

SSE = *Sum of Squares of Errors*

SSR = *Sum of Squares of Regression*

SSR = *Total Sum of Squares*

SVD = decomposizione ai valori singolari

$\mathbf{T}$ = matrice degli *score* sulle variabili di processo (-)

$\mathbf{T}^T$ = matrice trasposta di $\mathbf{T}$ (-)

$\mathbf{t}_h$ = vettore della *h*-esima colonna della matrice degli *score* $\mathbf{T}$ delle variabili di qualità del prodotto (-)

$\mathbf{t}_h^T$ = vettore trasposto di $\mathbf{t}_h$ (-)

$\mathbf{t}_{h,new}$ = vettore $\mathbf{t}_h$ aggiornato all'iterazione *h* (-)

$\mathbf{t}_{h,old}$ = vettore $\mathbf{t}_h$ calcolato all'iterazione *h* (-)

$\mathbf{t}_i$ = vettore degli score di $\mathbf{X}$ dell'*i*-esima variabile latente (-)

$\mathbf{t}_i^T$ = vettore trasposto di $\mathbf{t}_i$ (-)

T^2 = limite di confidenza per il diagramma degli score (-)

TNR = criterio di inserimento variabili vasato sul rapporto segnale/rumore

$\mathbf{U}$ = matrice degli *score* sulle variabili di qualità del prodotto (-)

$\mathbf{u}_h$ = vettore della *h*-esima colonna della matrice degli *score* $\mathbf{U}$ delle variabili di qualità del prodotto (-)

$\mathbf{u}_h^T$ = vettore trasposto di $\mathbf{u}_h$ (-)

$\mathbf{u}_j$ = vettore degli *score* di $\mathbf{Y}$ della *j*-esima variabile latente (-)

VIP = metodo di selezione di variabili *Variable Importance in Projection*

var = varianza (-)

$\mathbf{w}_h$ = pesi per mantenere ortogonali gli *score* nell'algoritmo di NIPALS (-)

$\mathbf{w}_h^T$ = vettore trasposto di $\mathbf{w}_h$ (-)

$\mathbf{w}_{h,new}$ = vettore $\mathbf{w}_h$ aggiornato all'iterazione *h* (-)

$\mathbf{w}_{h,old}$ = vettore $\mathbf{w}_h$ calcolato all'iterazione *h* (-)

$\mathbf{X}$ = matrice bidimensionale delle variabili di processo misurate (-)

$\mathbf{X}^{-1}$ = inversa della matrice delle variabili di processo $\mathbf{X}$ (-)

$\mathbf{X}$ = matrice tridimensionale delle variabili di processo misurate

$\mathbf{X}_h$ = matrice $\mathbf{X}$ aggiornata all'iterazione *h* (-)

$\mathbf{x}_i$ = vettore riga di $\mathbf{X}$ (-)

$\mathbf{x}_i^T$ = trasposto del vettore riga $\mathbf{x}_i$ (-)

$x_{i,j}$ = elemento della matrice bidimensionale delle variabili di processo $\mathbf{X}$ (-)

x_j = variabile di processo generica (-)

$\mathbf{x}_j$ = vettore colonna della matrice $\mathbf{X}$ o della matrice $\mathbf{X}$ (-)

x = variabile di processo generica (-)

y = variabile indipendente (-)

$\hat{y}_i$ = valore stimato della variabile dipendente y nel punto generico (-)

$\mathbf{Y}$ = matrice delle variabili di qualità del prodotto (-)
$\mathbf{Y}_h$ = matrice $\mathbf{Y}$ aggiornata all'iterazione h (-)
$y_{i,j}$ = elemento della matrice delle variabili di qualità del prodotto $\mathbf{Y}$ (-)
$\mathbf{y}_j$ = vettore colonna della matrice delle variabili di qualità del prodotto $\mathbf{Y}$ (-)

Lettere greche

α = percentile o grado di confidenza (-)
b_i = coefficiente di regressione parziale (-)
β = vettore dei coefficienti di regressione parziale (-)
ε_i = errore di misura generico (-)
ε = vettore degli errori di misura (-)
Λ = vettore che ha per elementi gli autovalori della matrice di covarianza (-)
λ_h = autovalore della matrice di covarianza di $\mathbf{X}$ associato alla h-esima componente principale
σ = deviazione standard (-)

Riferimenti bibliografici

Brauner, N. e M. Shacham, (1998). Role of Range and Precision of the Indipendent Variable in Regression of Data, *AIChE J*., **44**, 3, 603 – 611.

Brauner, N. e M. Shacham, (1999). Considering Error Propagation in Stepwise Polynomial Regression, *Ind. Eng. Chem. Res*., **38**, 4477 – 4485.

Burnham, A. J., J. F. MacGregor e R. Viveros (1999). Latent Variable Multivariate Regression Modeling, *Chem. Intll. Lab. Syst*., **48**, 167 – 180.

Centner, V., D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna (1996). Elimination of Uninformative Variables for Multivariate Calibration, *Anal. Chem.*, **68**, 3851 – 3858.

Chong, I. G. e C. H. Jun (2005). Performance of Some Variable Selection Methods when Multicollinearity is Present, *Chem. Intll. Lab. Syst*., **78**, 103 – 112.

Del Bosco E. (2005). Un Metodo Statistico per la Selezione Ottima di Sensori di Misura: Teoria e Sperimentazione. *Tesi di Laurea in Ingegneria Chimica*, DIPIC, Università di Padova.

Draper, N. R. e H. Smith (1998). *Applied Regression Analysis* (3rd ed.). J. Wiley & Sons, New York, (U.S.A.).

Duchesne, C. e J. F. MacGregor (2000). Multivariate Analysis and Optimization of Process Variable Trajectories for Batch Processes, *Chem. Intll. Lab. Syst*., **51**, 125 – 137.

Facco, P., M. Olivi, C. Rebuscini, F. Bezzo e M. Barolo (2007). Multivariate Statistical Estimation of Product Quality in the Industrial Batch Production of a Resin. *Proc. DYCOPS – 8th IFAC International Symposium on Dynamics and control of Process systems*, Cancùn (Mèxico), June 6 – 8th.

Forina, M., S. Lanteri, M. C. Cerrato Oliveros, C. Bizzarro Millan (2004). Selection of Useful Predictors in Multivariate Calibration, *Anal. Bioanal. Chem*., **380**, 397 – 418.

Furnival, G. M. (1971). All Possible Regressions with Less Computations, *Technometrics*, **13**, 2, 403 – 409.

Gauchi, J.-P. e P. Chagnon (2001). Comparison of Selection Methods of Explanatory Variables in PLS Regression with Application to Manufacturing Process Data, *Chem. Intll. Lab. Syst*., **58**, 171 – 193.

Geladi, P. e B. R. Kowalski (1986). Partial Least Squares Regression: a Tutorial, *Anal. Chim. Acta*, **185**, 1 – 17.

Gunst, R. F. e R. L. Mason (1977). Biased Estimation in Regression: An Evaluation Using Mean Squared Error, *J. Am. Stat. Ass.*, **72**, 359, 616 – 628.

Hocking, R. R. (1976). A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression, *Biometrics*, **32**, 1, 1 – 49.

Hoskuldsson, A. (1996). Dimension of Linear Models, *Chem. Intll. Lab. Syst.*, **32**, 37 – 55.

Kabe, D. G. (1963), Stepwise Multivariate Linear Regression, *J. Am. Stat. Ass.*, **58**, 303, 770 – 773.

Kano, M., K. Miyazaki, S. Hasebe e I. Hashimoto (2000). Inferential Control System of Distillation Composition using Dynamic Partial Least Squares, *J. Proc. Cont.*, **10**, 157 – 166.

Kourti, T. e J. F. MacGregor (1995). Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods, *Chem. Intll. Lab. Syst.*, **28**, 3 – 21.

Kresta, J. V., J. F. MacGregor e T. E. Marlin (1991), Multivariate Statistical Monitoring of Process Operating Performance, *Can. J. Chem. Engng.*, **69**, 35 – 47.

Lazraq, A., R. Cleroux, J.-P. Gauchi (2003). Selecting both Latent and Explanatory Variables in the PLS1 Regression Model, *Chem. Intll. Lab. Syst.*, **66**, 117 – 126.

Lindgren, F., P. Geladi, S. Rannar, S. Wold (1994). Interactive Variable Selection, IVS for PLS: 1. Theory and algorithms, *J. Chemometrics*, **8**, 349 – 363.

Lorho, G., F. Westad, e R. Bro (2006). Generalized Correlation Loadings. Extending Correlation to Congruence and to Multi-Way Models, *Chem. Intll. Lab. Syst.*, **84**, 119 – 125.

Marchi, C. (1999). Aggiornamento tecnologico e modellazione matematica di un impianto pilota di distillazione. *Tesi di Laurea in Ingegneria Chimica*, DIPIC, Università di Padova.

Mejdell, T. e S. Skogestad (1991). Estimation of Distillation Composition from Multiple Temperature Measurements using Partial-Least-Squares Regression, *Ind. Eng. Chem. Res.*, **30**, 2543 – 2555.

Montgomery, D. G. e E. A. Peck (1992). *Introduction to Linear Regression Analysis.* J. Wiley & Sons, New York (U.S.A.).

Montgomery, D. G. e G. C. Runger (2003). *Applied Statistic and Probability for Engineers* (3rd ed.). J. Wiley & Sons, New York (U.S.A.).

Muradore, R., F. Bezzo e M. Barolo (2006). Optimal Sensor Location for Distributed-Sensor Systems using Multivariate Regression, *Comp. Chem. Engng.*, **30**, 521 – 534.

Shacham, M. e N. Brauner, (1997). Minimizing The Effects of Collinearity in Polynomial Regression, *Ind. Eng. Chem. Res.*, **36**, 4405 – 4412.

Shacham, M., e N. Brauner (1999). Considering Precision of Experimental Data in Construction of Optimal Regression Models, *Chem. Eng. Proc.*, **38**, 477 – 486.

Shacham, M., e N. Brauner (2003). The SROV Program for Data Analysis and Regression Model Identification, *Comp. Chem. Engng.*, **27**, 701 – 714.

Shacham, M. e N. Brauner (2007). A New Procedure to Identify Linear and Quadratic Regression Models Based on Signal-to-Noise Ratio Indicators, *Math. Comp. Model*., **46**, 1- 2, 235 – 250.

Sharmin, R., U. Sundararaj, S. Shah, L. V. Griend e Y. J. Sun (2006). Inferential Sensor for Estimation of Polymer Quality Parameter: Industrial Application of a PLS-based Soft Sensor for a LDPE plant. *Chem. Eng. Sci*., **61**, 6372 – 6384.

Trygg, J. e S. Wold (2002). Ortogonalizated Projection to Latent Structure, O-PLS, *J. Chemom*., **16**, 3, 119 – 128.

Valle, S., W. Li e S. J. Qin (1999). Selection of the Number of Principal Components: the Variance of the Reconstruction Error Criterion with a Comparison to other Methods, *Ind. Eng. Chem. Res*., **38**, 4389 – 4401.

Wise, B.M. e N.B. Gallagher (1996). The Process Chemometrics Approach to Chemical Process Fault Detection and Supervision, *J. Proc. Cont.*, **6**, 329 – 348.

Wise, B.M. and N.B. Gallagher (2006). *PLS_Toolbox for Use with MATLAB®, Version 4.0*, Eigenvector Research, Wenatchee, WA.

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Model, *Technometrics*, **20**, 4, 397 – 405.

Wold, S., H. Annti, F. Lindgren and J. Ohman (1998). OSC of Near-Infrared Spectra, Chem. Intll. Lab. Syst., 44, 175-185.

Wold, S., M. Sjostrom, e L. Erikkson, (2001). PLS-regression: a Basic Tool of Chemometrics, *Chem. Intll. Lab. Syst.*, **58**, 109-130.

Xu, L. e W.-J. Zhang (2001). Comparison of Different Methods of Variable Selection, Anal Chim. Acta, **446**, 477-483.

Zamprogna, E., M. Barolo e D. E. Seborg (2004). Estimating Product Composition Profiles in Batch Distillation via Partial Least Squares Regression, *Cont. Engng. Prac*., **12**, 917 – 929.

Zamprogna, E., M. Barolo e D. E. Seborg (2005). Optimal Selection of Soft Sensor Inputs for Batch Distillation using Principal Component Analysis, *J. Proc. Cont*., **15**, 39 – 52.

Zarzo, M. e A. Ferrer, (2004), Batch Process Diagnosis: PLS with Variable Selection versus Block-Wise PCR, *Chem. Intll. Lab. Syst.* , **73**, 15 – 27.

Zhai, H. L., X. G. Chen, Z. D. Hu (2006). A new approach for the identification of important variables, *Chemom. Intll. Lab. Syst*., **80**, 130-135.