UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

TESI DI LAUREA MAGISTRALE IN
INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

SVILUPPO DI TECNICHE STATISTICHE MULTIVARIATE PER LA PROGETTAZIONE DI PRODOTTO IN CONDIZIONI DI INCERTEZZA

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

Laureando: FILIPPO DAL PASTRO

ANNO ACCADEMICO 2012 – 2013

Riassunto

Uno degli obiettivi della progettazione di prodotto/processo è identificare le materie prime, le condizioni iniziali e i parametri di processo necessari ad ottenere un prodotto della qualità desiderata. A questo scopo possono essere utilizzati modelli multivariati di regressione a variabili latenti, i quali sono in grado di descrivere la relazione tra le caratteristiche delle materie prime, i parametri di processo, le condizioni iniziali (regressori) da una parte e la qualità del prodotto (variabile risposta) dall'altra. L'inversione di questi modelli permette, a partire da una qualità di prodotto desiderata, di ricavare la migliore combinazione di regressori necessari ad ottenere la risposta voluta. In questa Tesi viene proposta una procedura per caratterizzare l'incertezza insita nell'inversione di modelli a variabili latenti. La procedura può trovare applicazione nell'ambito della progettazione di nuovi prodotti e processi.

L'obiettivo principale della Tesi è caratterizzare come l'incertezza di predizione (Zhang e Garcìa-Muñoz, 2009) si propaghi dalla qualità di prodotto desiderata alle caratteristiche delle materie prime o ai parametri di processo calcolati dall'inversione del modello. Innanzitutto sono valutati alcuni approcci per la stima dell'incertezza in fase di predizione, scegliendoli tra quelli proposti da Zhang e Garcìa-Muñoz (2009). Successivamente è studiato come l'incertezza si correli all'accuratezza e alla rappresentatività dei modelli costruiti.

Le metodologie sviluppate sono applicate a tre diversi casi di studio: la determinazione degli ingressi di un modello matematico, la progettazione di prodotto di una granulazione umida (Vemavarapu *et al.*, 2009) e la progettazione di prodotto e processo in una granulazione a secco di cellulosa microcristallina simulata.

I risultati mostrano come le procedure sviluppate caratterizzino appieno la propagazione dell'incertezza di predizione nell'inversione dei modelli a variabili latenti, e come l'incertezza sia moderatamente correlata alle statistiche di distanza dai valori medi e di rappresentatività del modello. Tali risultati risultano fondamentali nella prospettiva di un utilizzo robusto della modellazione a variabili latenti per lo sviluppo di nuovi prodotti e processi, in linea con le richieste degli enti regolatori.

Indice

INTRODUZIONE 9
CAPITOLO 1 – L'incertezza nei modelli di regressione lineare 11
1.1 TECNICHE DI ANALISI STATISTICA MULTIVARIATA 11
1.1.1 Analisi delle componenti principali 12
1.1.1.1 Selezione del numero di componenti principali 15
1.1.1.2 Limiti di controllo 15
1.1.2 Proiezione su strutture latenti 16
1.2 METODI PER LA STIMA DELL'INCERTEZZA DELLE PREDIZIONI IN UN MODELLO PLS 17
1.2.1 Approcci per il calcolo della deviazione standard dell'errore di predizione 18
1.2.1.1 Deviazione standard dal metodo OLS 18
1.2.1.2 Deviazione standard dal metodo U-deviation 20
1.2.2 Approcci per il calcolo dei gradi di libertà 21
1.2.2.1 Naïve 21
1.2.2.2 Pseudo gradi di libertà (PDF) 21
1.2.2.3 Gradi di libertà generalizzati (GDF) 21
1.3 INVERSIONE NEI MODELLI DI REGRESSIONE LINEARE 22
1.3.1 L'inversione diretta e lo spazio nullo 22
1.4 INCERTEZZA DI PREDIZIONE E INVERSIONE DI MODELLO 24
CAPITOLO 2 – Descrizione dei dati 27
2.1 CASO DI STUDIO 1: DATI DA MODELLO MATEMATICO 27
2.1.1 Descrizione dei dati del caso 1 27
2.2 CASO DI STUDIO 2: DATI DI GRANULAZIONE UMIDA 28
2.2.1 Il processo di granulazione ad umido 28
2.2.2 Descrizione dei dati del caso 2 29
2.3 CASO DI STUDIO 3: DATI DI GRANULAZIONE DA UNA SIMULAZIONE CON *gSolids™* 30
2.3.1 Il processo di compattazione a rulli 30
2.3.2 La modellazione della compattazione a rulli 32

2.3.2.1 Calcolo dell'angolo di *nip* 32
2.3.2.2 Calcolo della pressione massima 33
2.3.2.3 Calcolo della densità e della porosità in uscita 33
2.3.3 Descrizione dei dati del caso di studio 3 34
CAPITOLO 3 – Analisi esplorativa dei modelli predittivi della qualità 37
3.1 CASO DI STUDIO 1: ANALISI ESPLORATIVA DEI DATI DA MODELLO MATEMATICO 37
3.2 CASO DI STUDIO 2: ANALISI ESPLORATIVA DEI DATI DI GRANULAZIONE UMIDA 39
3.3 CASO DI STUDIO 3: ANALISI ESPLORATIVA DEI DATI DI GRANULAZIONE MEDIANTE COMPATTAZIONE A RULLI SIMULATA 42
3.3.1 Modelli dello scenario 1 42
3.3.2 Modelli dello scenario 2 44
3.3.3 Modelli dello scenario 3 45
3.4 CONCLUSIONI SUI MODELLI PREDITTIVI DELLA QUALITÀ 47
CAPITOLO 4 – Caratterizzazione dell'incertezza nella progettazione di prodotto 49
4.1 RISULTATI PER IL CASO DI STUDIO 1 49
4.1.1 Applicazione dei metodi di calcolo dell'incertezza 49
4.1.2 Applicazione dell'inversione diretta 52
4.1.3 Caratterizzazione del risultato 56
4.1.4 Risultati 58
4.1.5 Correlazioni tra le metriche 61
4.2 RISULTATI PER IL CASO DI STUDIO 2 65
4.2.1 Applicazioni ai dati di granulazione umida 65
4.2.2 Correlazioni tra le metriche 66
4.3 RISULTATI PER IL CASO DI STUDIO 3 66
4.3.1 Risultati per lo scenario 1 67
4.3.2 Correlazioni tra le metriche per lo scenario 1 68
4.3.3 Risultati per lo scenario 2 70
4.3.4 Risultati per lo scenario 3 71
4.3.5 Correlazioni tra le metriche per gli scenari 2 e 3 72
4.4 CONCLUSIONI 72

CONCLUSIONI 75
NOMENCLATURA 77
RIFERIMENTI BIBLIOGRAFICI 81

Introduzione

La recente introduzione del concetto di *Quality-by-Design* da parte della *Food and Drug Administration* (FDA, 2004) e di altri enti regolatori promuove l'utilizzo da parte delle industrie di approcci scientifici sistematici, sia nelle fasi di sviluppo di nuovi prodotti e processi produttivi, che in quelle di ottimizzazione e controllo di processo. Per questo motivo le industrie sono incoraggiate ad utilizzare, tra le altre, le metodologie basate sull'analisi dei dati di processo. In letteratura si trovano diversi esempi industriali nei quali sono state applicate con successo tecniche di modellazione basate su metodi statistici multivariati. L'utilizzo di questi metodi, come ad esempio la modellazione a variabili latenti (LVM, *latent variable modeling*) permette di aumentare la conoscenza sul processo (Tomba *et al.*, 2013), ottimizzare il processo (Garcìa-Muñoz *et al.*, 2003), monitorare il sistema di produzione (Nomikos e MacGregor, 1995) e predire la qualità del prodotto basandosi solo su dati raccolti *on-line/in-line/at-line*, ed evitando dispendiose analisi di laboratorio *off-line* (Facco *et al.*, 2009).

Nel proprio utilizzo diretto, le tecniche LVM permettono di predire la qualità di prodotto (variabile risposta) a partire da un *dataset* di regressori, che possono comprendere le qualità delle materie prime e i parametri di processo. Nell'applicazione inversa, invece, queste tecniche vengono impiegate con efficacia per lo sviluppo di nuovi prodotti e processi (Jaeckle e MacGregor, 1998; Garcìa-Muñoz *et al.*, 2006). Grazie all'inversione del modello, è possibile infatti trovare le migliori combinazioni di regressori che sono necessarie per ottenere la risposta desiderata.

Tuttavia, la modellazione LVM è solitamente affetta da diversi tipi di incertezza: incertezza sui parametri del modello di calibrazione (Martens e Martens, 2000), incertezza dei dati di calibrazione (Reis e Saraiva, 2005), incertezza di predizione (Fernandez Pierna *et al*., 2003).

In questa Tesi, il concetto di incertezza di predizione proposto da Zhang e Garcìa-Muñoz (2009) viene incluso nell'inversione dei modelli a variabili latenti. In particolare, questa Tesi mira a caratterizzare la propagazione dell'incertezza di predizione dalla qualità di prodotto desiderata (variabile risposta) alle qualità delle materie prime e ai parametri di processo ottenuti dall'inversione (regressori).

Le metodologie proposte sono applicate a tre diversi casi studio: un modello matematico, la progettazione di prodotto di una granulazione umida (Vemavarapu *et al.*, 2009) e la progettazione di prodotto e processo in una granulazione a secco di cellulosa microcristallina simulata.

La Tesi è organizzata in quattro Capitoli. Nel primo Capitolo vengono descritti i metodi matematici utilizzati nell'analisi, in particolare le tecniche di analisi statistica multivariata e i metodi utilizzati per il calcolo dell'incertezza relativa all'utilizzo di modelli di regressione lineare multivariata. Infine, si discute l'inversione di questi ultimi e l'applicazione di metodi per la stima dell'incertezza di predizione nel caso dell'utilizzo dell'inversione.
Il secondo Capitolo contiene la descrizione dei casi di studio presi in esame.
Il terzo Capitolo presenta un'analisi esplorativa dei casi di studio mediante modelli a variabili latenti che stimano la qualità finale del prodotto. In particolare, si descrive la struttura dei modelli, si analizzano le correlazioni tra le diverse variabili dei dati di ingresso e si studiano la rappresentatività dei modelli di stima delle variabili di qualità e la loro accuratezza.
Infine, nel quarto Capitolo vengono presentati i principali risultati ottenuti. Innanzitutto sono verificati gli approcci migliori per la stima dell'incertezza, tra quelli proposti da Zhang e Garcìa-Muñoz (2009), e poi sono caratterizzate nel problema di progettazione di prodotto/processo l'incertezza, l'accuratezza e la rappresentatività dei modelli a variabili latenti, con particolare attenzione alle correlazioni tra queste.
Alcune considerazioni conclusive chiudono la Tesi.

Capitolo 1

L’incertezza nei modelli di regressione lineare

In questo Capitolo si descrivono i metodi matematici utilizzati in questa Tesi. Dopo aver introdotto i metodi per l’analisi statistica multivariata, con particolare attenzione alla proiezione su strutture latenti (*projection on latent structures*, PLS, detta anche *partial least squares regression*), si descrivono i metodi utilizzati per il calcolo dell’incertezza nei metodi di regressione lineare multivariata. Infine si discute l’inversione nei modelli di regressione lineare e la metodologia per il calcolo dello spazio nullo.

1.1 Tecniche di analisi statistica multivariata

Nell’applicazione delle tecniche statistiche multivariate (TSM) possono essere perseguiti diversi obiettivi: aumentare la conoscenza del processo (Tomba *et al.*, 2013) ottenendo modelli per ottimizzare il processo (Garcìa-Muñoz *et al.*, 2003), monitorare il sistema di produzione (Nomikos e MacGregor, 1995), predire la qualità del prodotto basandosi solo su dati raccolti *on-line/in-line/at-line*, evitando dispendiose analisi di laboratorio *off-line* (Facco *et al.*, 2009). Le TSM permettono di costruire modelli che rappresentano il sistema in esame a partire da dati a disposizione, attraverso l’individuazione di nuove variabili (variabili latenti) in numero minore rispetto a quelle originali, ma che catturano in maniera ottimale la struttura di variabilità e correlazione dei dati. I più comuni fra questi metodi sono l’analisi delle componenti principali (*principal component analysis*, PCA; Jackson, 1991) che permette di studiare le correlazioni tra le variabili misurate, e la PLS (Wold *et al.*, 1983; Höskuldsson, 1988), che permette di costruire modelli di regressione per lo studio della correlazione tra due *set* di dati: le variabili in ingresso al sistema (regressori) e le variabili di risposta (Figura 1.1).

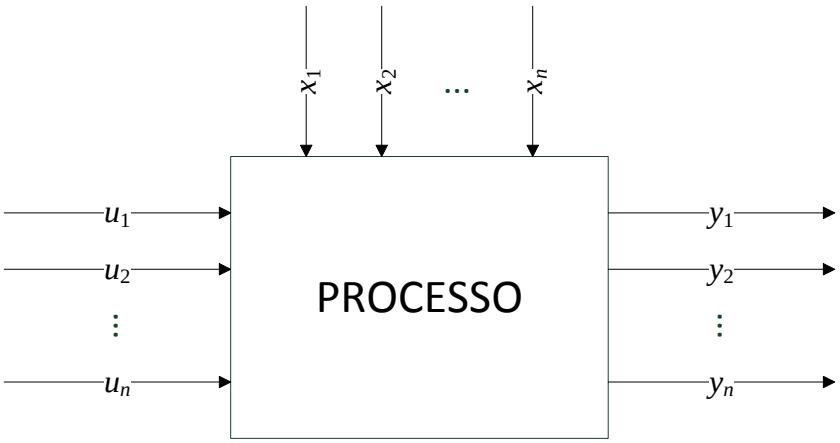


Figura 1.1: *Rappresentazione schematica di un processo, con ingressi u e x, e uscite y.*

È possibile distinguere diversi tipi di ingressi: le variabili u_n sono le grandezze misurabili e manipolabili (sono ad esempio proprietà delle materie prime o parametri di processo), mentre le variabili x_n sono misurate ma non manipolabili. Entrambe queste grandezze vengono raccolte nella matrice **U** degli ingressi. Le variabili in uscita y_n vanno a costituire le risposte del sistema e vengono raccolte nella matrice **Y** (ad esempio: qualità del prodotto finale). Nella regressione, le variabili della matrice **U** sono definite predittori (o regressori) e le grandezze di **Y** variabili stimate (o predette).

1.1.1 Analisi delle componenti principali

La PCA è un metodo statistico multivariato molto efficace nel comprimere una serie di dati correlati ed estrarne le informazioni più rilevanti che descrivono la variabilità sistematica dei dati, modellandone le caratteristiche di correlazione (Jackson, 1991).
Spesso, infatti, nelle misure di processo ci sono informazioni molto correlate e per questo ridondanti. Con la PCA si possono estrarre informazioni di covarianza e correlazione, trovando le combinazioni lineari delle variabili originali che descrivono in modo ottimale la variabilità dei dati (Wise e Gallagher, 1996). A livello geometrico, PCA individua le direzioni di massima variabilità dei dati, dette anche componenti principali (PC). Le componenti principali sono combinazioni lineari delle variabili originarie e sono tra loro ortogonali. In Figura 1.2 è illustrato un esempio di riduzione di uno spazio tridimensionale a bidimensionale. Nello spazio originario, la variabilità dei dati si sviluppa su tre variabili, genericamente u_1, u_2, u_3. Essa può essere rappresentata in uno spazio fittizio bidimensionale di componenti principali, definite dalla PCA, che rappresentano le informazioni contenute nello spazio originario, senza perdita rilevante di informazioni, in quanto esiste una correlazione intrinseca tra le variabili iniziali. Il discorso può essere esteso ad uno spazio multidimensionale.

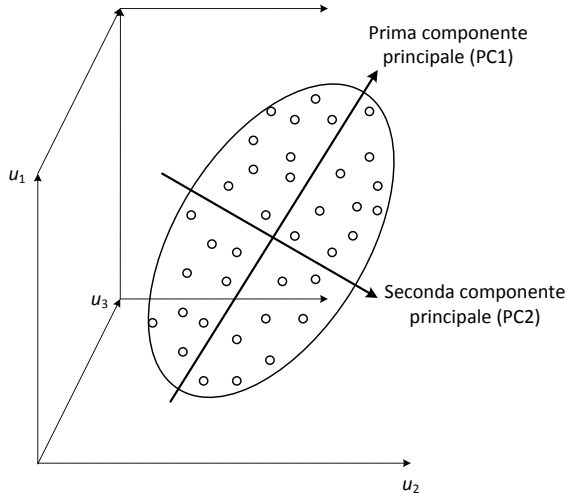


Figura 1.2: *Interpretazione geometrica della PCA, con esempio di riduzione da uno spazio tridimensionale a uno bidimensionale, in grado di descrivere la maggior parte della variabilità dei dati senza perdita rilevante di informazione.*

La PCA realizza una decomposizione delle variabili di processo in autovettori della matrice di covarianza di **U** [$N \times V_U$], matrice dei dati in esame in cui ogni riga rappresenta gli N campioni e ogni colonna le V_U variabili:

$$\mathrm{cov}(\mathbf{U}) = \frac{\mathbf{U}^{\mathrm{T}}\mathbf{U}}{N\text{-}1} \quad . \tag{1.1}$$

Dal punto di vista algebrico, la matrice **U** di rango R_{U} viene decomposta in una somma di R_{U} matrici $\mathbf{M}_r$ di rango unitario (Geladi e Kowalski,1986):

$$\mathbf{U} = \mathbf{M}_1 + \mathbf{M}_2 + ... + \mathbf{M}_r + ... + \mathbf{M}_{R_{\mathrm{U}}} = \sum_{r=1}^{R_{\mathrm{U}}} \mathbf{M}_r \quad . \tag{1.2}$$

La generica matrice $\mathbf{M}_r$ può essere espressa come il prodotto esterno di due vettori **t** e **p** (*score* e *loading*) di dimensione rispettivamente [$N \times 1$] e [$V_U \times 1$]:

$$\mathbf{U} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + ... + \mathbf{t}_r\mathbf{p}_r^{\mathrm{T}} + ... + \mathbf{t}_{R_{\mathrm{U}}}\mathbf{p}_{R_{\mathrm{U}}}^{\mathrm{T}} \quad . \tag{1.3}$$

PCA esegue dunque l'operazione di approssimazione:

$$\mathbf{U} = \sum_{a=1}^{A} \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}} + \mathbf{E}_{\mathrm{U}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{\mathrm{U}} \quad , \tag{1.4}$$

dove **T** è la matrice degli *score*, **P** la matrice dei *loading*, $\mathbf{E}_{\mathrm{U}}$ è la matrice dei residui e A il numero di componenti principali, che viene scelto per descrivere la variabilità sistematica dei dati.

In maggior dettaglio, gli *score* sono combinazioni dei dati originali secondo:

$$\mathbf{t}_i = \mathbf{U}\mathbf{p}_i \quad . \tag{1.5}$$

La matrice degli *score* **T**, che ha per righe i vettori $\mathbf{t}_i$, ha dimensioni [N×A] e rappresenta le coordinate dei campioni sullo spazio fittizio individuato dalle componenti principali. Gli *score* contengono le informazioni su come i campioni si relazionano tra loro.
La matrice **P** dei *loading*, di dimensioni [A×V_U], ha per righe i vettori $\mathbf{p}_i$, ovvero gli autovettori della matrice di covarianza. Essa contiene informazioni su come le variabili si relazionano tra loro. È importante notare che, poiché gli *score* sono tra loro ortogonali e i *loading* ortonormali, le componenti principali sono tra loro non correlate. I residui raccolti in $\mathbf{E}_U$ corrispondono alle informazioni che non vengono incorporate nel modello e che, mediante un'opportuna scelta di A, dovrebbero contenere informazioni sul solo rumore. Disponendo le coppie $\mathbf{t}_i$ e $\mathbf{p}_i$ in ordine decrescente rispetto ai rispettivi autovalori, che sono misure della varianza spiegata dalla a-esima componente principale, è possibile rappresentare i dati con un numero inferiore di variabili (A<<V_U) rispetto a quelle originarie, senza perdere informazioni rilevanti, qualora A sia scelto in maniera opportuna.
Il pretrattamento dei dati è molto importante e avviene tramite le operazioni di bilanciamento al valor medio (*mean centering*) e riduzione a scala unitaria (*autoscaling*): il primo consiste nel sottrarre ai valori delle colonne di **U** la rispettiva media, in modo da ottenere per ogni colonna media pari a 0; il secondo consiste nell'operare il *mean centering* e dividere ogni elemento della colonna per la deviazione standard della stessa.
La mancanza di accuratezza statistica nel rappresentare i dati è rappresentata dall'errore quadratico medio Q_n. Q_n è la somma dei quadrati di ciascun campione di $\mathbf{E}_U$, ovvero per il campione n-esimo:

$$Q_n = \mathbf{e}_n\mathbf{e}_n^{\mathrm{T}} \quad . \tag{1.6}$$

La statistica Q_n indica quanto bene ogni campione viene rappresentato dal modello e, in termini geometrici, il valore di $\sqrt{Q_n}$ rappresenta la distanza euclidea dell' n-esimo punto dall'iperpiano di dimensioni ridotte costituito dalle componenti principali.
Per quantificare quanto un'osservazione è lontana dalla media (l'origine del sistema delle componenti principali) si introduce la statistica T^2 di Hotelling. Essa è la somma al quadrato degli *score* normalizzati secondo la varianza spiegata ed è definita come:

$$T_n^2 = \mathbf{t}_n\boldsymbol{\lambda}^{-1}\mathbf{t}_n^{\mathrm{T}} \quad , \tag{1.7}$$

dove $\mathbf{t}_n$ è la n-esima riga della matrice degli *score* **T**, $\boldsymbol{\lambda}^{-1}$ è la matrice diagonale inversa degli autovalori (Wise e Gallagher, 1996).

1.1.1.1 Selezione del numero di componenti principali

Nella (1.4) il residuo $\mathbf{E}_U$ deve contenere solo informazioni sul rumore, mediante un'opportuna scelta di *A*, in modo da massimizzare la rappresentatività del modello. La scelta del numero di componenti principali risulta quindi fondamentale. Le metodologie utilizzate in questa Tesi sono due: la convalida incrociata (*cross validation;* Mosteller e Wallace, 1963; Stone, 1974) e regola dell'autovalore (Martens e Naes, 1989).

La convalida incrociata è una procedura iterativa in cui la matrice **U** dei dati viene suddivisa in segmenti, costituiti da uno o più campioni (righe di **U**), e un modello PCA viene costruito sulla matrice privata di un segmento. Sul modello si utilizza poi il segmento rimosso dalla calibrazione per la convalida. La procedura è applicata su tutti i segmenti e ad ogni iterazione si valuta l'errore medio quadratico di convalida incrociata (RMSECV, *root-mean squared error of cross validation*):

$$\mathrm{RMSECV}_k = \sqrt{\frac{\mathrm{PRESS}_k}{N}} \quad , \tag{1.8}$$

in cui:

$$\mathrm{PRESS}_k = \sum_{i=1}^{N} \left(u_{i,k} - \hat{u}_{i,k}\right)^2 \quad , \tag{1.9}$$

dove PRESS indica la somma dei quadrati dell'errore di predizione (*predicted error sum of squares*) e $\hat{u}_{i,k}$ è l'elemento stimato dal modello dato dal prodotto tra gli *score* e i *loading*.

L'aggiunta di componenti principali al modello solitamente fa decrescere l'errore di stima di $\hat{u}_{i,k}$ nel *set* di calibrazione. Quando però il numero di componenti principali è eccessivo si descrive una parte della variabilità dei dati non sistematica del *set* di calibrazione (ad esempio, rumore), e l'RMSECV cresce. Il minimo di RMSECV al variare del numero di componenti principali individua il numero di PC ottimale per costruire il modello.

Il secondo criterio è la regola dell'autovalore λ e si basa sul fatto che l'autovalore λ_a dell' *a*-esima componente principale è una stima indiretta del numero di variabili originarie rappresentate dalla stessa (Martens e Naes, 1989). Per questo motivo, gli autovalori con valore superiore all'unità rappresentano più di una variabile. Dunque, il numero di componenti principali *A* da adottare per costruire il modello corrisponde al numero di autovalori che soddisfano $\lambda_A \geq 1$.

1.1.1.2 Limiti di controllo

È possibile definire dei limiti di controllo sulla rappresentatività del modello, con un determinato margine di fiducia (1-α).

Secondo Jackson (1991), i limiti nello spazio degli *score* sono definiti da un'ellissoide di fiducia che ha come semiassi:

$$l_a = \sqrt{\lambda_a T^2_{A,N,\alpha}} \quad \forall a = 1,2,...,A \quad , \tag{1.10}$$

dove $T^2_{A,N,\alpha}$ è:

$$T^2_{A,N,\alpha} = \frac{A(N-1)}{N-A} F_{A,(N-A),\alpha} \quad , \tag{1.11}$$

in cui compare la distribuzione *F*, che dipende dal numero di componenti principali *A*, dal numero di campioni *N* e dal limite di fiducia (1-α).

1.1.2 Proiezione su strutture latenti

La proiezione su strutture latenti (PLS, Wold *et al.*, 1983; Höskuldsson, 1988), è un metodo di regressione utilizzato per correlare due matrici di dati tra loro, con scopo in genere predittivo. PLS trova la relazione tra i dati contenuti in una matrice **U** e le variabili risposta di una matrice **Y**, attraverso la costruzione di un modello che, noto il valore delle variabili in **U**, stimi il valore delle relative variabili **Y**.

In generale, il metodo PLS è utilizzato in chemiometria e nel controllo statistico di processo soprattutto nei casi in cui, come nella PCA, si devono trattare molti dati correlati tra loro. In questo senso, l'analisi PLS trova i fattori che colgono la parte della varianza nelle variabili in **U**, maggiormente correlata alla variabilità delle variabili in **Y**.

Definito un set di dati **U** [$N \times V_U$] di *N* campioni/osservazioni in cui V_U predittori sono misurati (ad esempio variabili di processo) e un set di dati **Y** [$N \times V_y$] in cui sono raccolte le V_y risposte per gli *N* campioni (ad esempio, qualità di prodotto), un modello di regressione lineare a variabili latenti (*latent variable regression model*, LVRM) trova le direzioni di massima variabilità della matrice **U** che meglio spiegano la matrice **Y**. In altri termini trova le principali forze agenti sul processo che più sono correlate alle qualità del prodotto (Wise e Gallagher, 1996).

Dopo un appropriato pretrattamento dei dati, PLS proietta i predittori e le risposte in uno spazio comune (lo spazio delle variabili latenti) in cui la correlazione tra **U** e **Y** è massimizzata.

Le matrici dei predittori **U** e **Y** delle variabili predette sono decomposte come:

$$\mathbf{U} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{\mathrm{U}} \tag{1.12}$$

$$\mathbf{Y} = \mathbf{J}\mathbf{Q}^{\mathrm{T}} + \mathbf{E}_{\mathrm{Y}} \tag{1.13}$$

$$\mathbf{T} = \mathbf{U}\mathbf{W} \quad , \tag{1.14}$$

dove **T** [$N \times lv$] e **J** [$N \times lv$] sono le matrici degli *score* di **U** e **Y** rispettivamente, $\mathbf{P}^{\mathrm{T}}$ [$V_U \times lv$] e $\mathbf{Q}^{\mathrm{T}}$ [$V_y \times lv$] quelle dei *loading*, $\mathbf{E}_{\mathrm{U}}$ [$N \times V_U$] e $\mathbf{E}_{\mathrm{Y}}$ [$N \times V_y$] sono le matrici dei residui per rendere conto degli errori di ricostruzione del modello e **W** [$V_U \times lv$] è la matrice dei pesi (*weights*)[1]. *lv* è il numero di variabili latenti utilizzate nella costruzione del modello e corrisponde alla dimensione dello spazio del modello che viene determinato mediante convalida incrociata o regola dell'autovalore.

I vettori degli *score* **t** sono calcolati per ogni dimensione del modello PLS, in modo che la combinazione lineare delle variabili in **U**, attraverso gli opportuni pesi (i *weights*), e la combinazione lineare delle variabili in **Y**, massimizzino la covarianza tra **U** e **Y**, spiegata da ciascuna LV (variabile latente, *latent variable*). I pesi **w** e **q** sono introdotti per mantenere l'ortogonalità degli *score*. L'analisi correla gli *score* della matrice **U** con gli *score* della matrice **Y** attraverso una relazione interna lineare:

$$\mathbf{j}_a = b_a \mathbf{t}_a \quad , \tag{1.15}$$

essendo b_a un elemento del vettore dei coefficienti di regressione della relazione interna **b** (Geladi e Kowalski, 1996).

1.2 Metodi per la stima dell'incertezza delle predizioni in un modello PLS

Nel caso in cui i modelli costruiti con le TSM vengano utilizzati per predire delle qualità di prodotto basandosi solo su dati raccolti *on-line/in-line/at-line*, vi è la necessità di fornire alcune metriche per misurare la qualità e l'affidabilità di una predizione, cioè bisogna stimare l'incertezza sulla predizione.

Alcune delle metriche che si impiegano per la costruzione e la convalida di un modello PLS non sono direttamente traducibili in stime per l'incertezza delle predizioni del modello. Tra queste vi è il SEC (errore standard di calibrazione, *standard error of calibration*), definito dalla:

$$\mathrm{SEC} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_{cal,i} - \hat{y}_{cal.i} \right)^2}{N - df}} \quad , \tag{1.16}$$

dove $\hat{y}_{i,cal}$ è la risposta *i*-esima calcolata dal modello in calibrazione, $\left(y_{cal,i} - \hat{y}_{cal.i} \right)$ è il residuo in fase di calibrazione e *df* è il numero di gradi di libertà utilizzati dal modello PLS.

[1] Le matrici degli *score* **T** e dei residui $\mathbf{E}_{\mathrm{U}}$ definite da PLS differiscono da quelle utilizzate da PCA, ma con abuso di notazione si è usata la medesima simbologia.

La procedura generale per valutare l'incertezza per una risposta $\hat{y}_{N+1}$ di una nuova osservazione $N+1$ con predittore il vettore riga $\mathbf{u}_{N+1}^{\mathrm{T}}$ è stata proposta da Zhang e Garcìa-Muñoz (2009) e consiste di due passaggi:

1. si stima la deviazione standard *s* dell'errore di predizione;
2. assumendo che gli errori seguano la distribuzione *t* di *Student*, si calcola l'incertezza in termini di intervallo di fiducia (*confidence interval*, CI):

$$\mathrm{CI} = \hat{y}_{N+1} \pm \left(t_{\alpha/2,N-df}\right)s \quad , \tag{1.17}$$

dove *t* è il valore della *t* di *Student* e con la significatività dell'intervallo di incertezza α.
Nel calcolo dell'intervallo d'incertezza è dunque necessario stimare la deviazione standard dell'errore di predizione *s* e il numero di gradi di libertà. Zhang e Garcìa-Muñoz (2009) considerano quattro metodi per il calcolo della deviazione standard *s* dell'errore di predizione nei metodi di regressione lineare multivariata:

- espressioni dal metodo ai minimi quadrati ordinari (*ordinary least squares*, OLS);
- metodi basati sulla linearizzazione;
- approcci basati sul ricampionamento;
- metodo empirico *U-deviation*;

e tre metodi per il calcolo dei gradi di libertà:

- Naïve;
- pseudo gradi di libertà (*pseudo degrees of freedom* , PDF);
- gradi di libertà generalizzati (*generalized degrees of freedom*, GDF).

Una discussione di questi metodi verrà proposta nel Capitolo 4.
È importante notare che tutti questi metodi sono applicabili ad una risposta **y** monovariata.

1.2.1 Approcci per il calcolo della deviazione standard dell'errore di predizione

In questo Sottoparagrafo vengono presentati i metodi per il calcolo della deviazione standard dell'errore di predizione basati sul metodo ai minimi quadrati ordinari e l'approccio empirico *U-deviation*.

1.2.1.1 Deviazione standard dal metodo OLS

Questo approccio per valutare l'incertezza nei metodi di regressione lineare parte da equazioni ricavate dal metodo ai minimi quadrati ordinari, che poi vengono estese alla PLS.
L'equazione che sta alla base dei modelli lineari è la seguente (Zhang e Garcìa-Muñoz, 2009):

$$\mathbf{y} = \mathbf{U}\hat{\boldsymbol{\beta}} + \mathbf{E}_{\mathrm{y}} \quad , \tag{1.18}$$

dove $\hat{\boldsymbol{\beta}}$ è la matrice dei coefficienti del modello di regressione lineare e **y** e **U** sono state scalate rispetto la media.
Mentre il modello per predire una risposta di un oggetto $N+1$ da un'osservazione $\mathbf{u}_{N+1}^{\mathrm{T}}$ è:

$$\hat{y}_{N+1} = \bar{y} + \mathbf{u}_{N+1}^{*\mathrm{T}}\hat{\boldsymbol{\beta}} + \mathbf{e}_{N+1} \quad , \tag{1.19}$$

dove $\bar{y}$ è la risposta media misurata e $\mathbf{u}_{N+1}^{*\mathrm{T}}$ è il predittore centrato rispetto la media.
La stima dei parametri di regressione $\hat{\boldsymbol{\beta}}$ è data da:

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y} \quad . \tag{1.20}$$

Dato che $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ è una funzione lineare di **y**, la sua covarianza può essere calcolata come:

$$\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right) = \left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}\left[\mathrm{var}(\mathbf{y})\right]\mathbf{U}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1} = \left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\sigma^2 \quad . \tag{1.21}$$

L'operatore var è riferito al calcolo della varianza se applicato ad uno scalare e della covarianza se usato per una matrice. Nella (1.19) la covarianza di **y** è pari a $\mathbf{I}\sigma^2$.
A questo punto la varianza della predizione può essere ottenuta utilizzando le (1.17) e (1.19):

$$\begin{aligned} s^2 &= \mathrm{var}\left(\bar{y} + \mathbf{u}_{N+1}^{*\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \mathbf{e}_{N+1}\right) = \mathrm{var}\left(\bar{y}\right) + \mathrm{var}\left(\mathbf{u}_{N+1}^{*\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right) + \mathrm{var}\left(\mathbf{e}_{N+1}\right) \\ &= \mathrm{var}\left(\bar{y}\right) + \mathbf{u}_{N+1}^{*\mathrm{T}}\,\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)\mathbf{u}_{N+1}^{*} + \mathrm{var}\left(\mathbf{e}_{N+1}\right) \\ &= \frac{\sigma^2}{N} + \left[\mathbf{u}_{N+1}^{*\mathrm{T}}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\mathbf{u}_{N+1}^{*}\right]\cdot\sigma^2 + \sigma^2 \quad . \end{aligned} \tag{1.22}$$

La (1.22) può essere così semplificata:

$$s^2 = \sigma^2 \cdot \left(1 + h_{N+1} + \frac{1}{N}\right) \quad , \tag{1.23}$$

in cui $h_{N+1} = \mathbf{u}_{N+1}^{*\mathrm{T}}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\mathbf{u}_{N+1}^{*}$ è il leveraggio (*leverage*) per l'osservazione *N+1*.
Estendendo le formule alla PLS, i coefficienti di regressione posso essere scritti come:

$$\hat{\boldsymbol{\beta}} = \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y} \quad , \tag{1.24}$$

dove **W** è la matrice dei pesi.
Ignorando la dipendenza di **W** da **y** si possono ottenere dei risultati simili alle Equazioni (1.21) e (1.22):

$$\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) \approx \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\left[\mathrm{var}(\mathbf{y})\right]\mathbf{U}\mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{W}\mathbf{W}^{\mathrm{T}}\sigma^2 \tag{1.25}$$

$$s^2 = \mathrm{var}\left(\bar{y} + \mathbf{u}_{N+1}^{*\mathrm{T}}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_{N+1}\right) = \mathrm{var}(\bar{y}) + \mathrm{var}\left(\mathbf{u}_{N+1}^{*\mathrm{T}}\hat{\boldsymbol{\beta}}\right) + \mathrm{var}(\mathbf{e}_{N+1})$$

$$= \mathrm{var}(\bar{y}) + \mathbf{u}_{N+1}^{*\mathrm{T}}\,\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{u}_{N+1}^{*} + \mathrm{var}(\mathbf{e}_{N+1}) = \frac{\sigma^2}{N} + \left[\mathbf{u}_{N+1}^{*\mathrm{T}}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{u}_{N+1}^{*}\right]\cdot\sigma^2 + \sigma^2$$

$$= \sigma^2 \cdot \left(1 + h_{N+1} + \frac{1}{N}\right) \quad . \tag{1.26}$$

La forma della (1.26) è stata ottenuta imponendo $h_{N+1} = \mathbf{t}_{N+1}^{\mathrm{T}}\mathbf{t}_{N+1} = \mathbf{u}_{N+1}^{*\mathrm{T}}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{u}_{N+1}^{*}$ dove $\mathbf{t}_{N+1}^{\mathrm{T}} = \mathbf{u}_{N+1}^{*\mathrm{T}}\mathbf{W}$ è lo *score* per l'osservazione N+1. All'Equazione (1.26) si farà riferimento come *Simple-Faber96* (SF96). La stima di σ è data dal SEC.

1.2.1.2 Deviazione standard dal metodo U-deviation

Il metodo *U-deviation* (UD) è basato sulla seguente relazione empirica (Zhang e Garcìa-Muñoz, 2009):

$$s = \sqrt{\frac{s_{y\_val}}{2}\left(h_{N+1} + \frac{s_{U,pr}}{s_{Utol,val}} + \frac{1}{N}\right)} \quad , \tag{1.27}$$

dove s_{y_val} è la varianza del residuo di **y** nel set di convalida, $s_{U,pr}$ è la varianza del residuo del predittore $\mathbf{u}_{N+1}$ e $s_{Utol,val}$ è la varianza media del residuo di **U** nel set di convalida. Se il predittore N+1 si comporta similmente ai dati del *set* di calibrazione, è lecito compiere la seguente approssimazione:

$$\frac{s_{U,pr}}{s_{Utol,val}} \approx 1 \quad . \tag{1.28}$$

Sostituendo la (1.28) nella (1.27) e confrontando con la (1.26) è possibile vedere come il metodo *U-deviation* utilizzi $s_{y\_val}/2$ come estimatore di σ^2. De Vries e Ter Braak (1995) studiarono come il metodo *U-deviation* sottostimasse l'incertezza sulla predizione di un modello PLS, suggerendo quindi di sostituire il fattore 2 nella (1.27) con $1/(1-(lv+1)/N)$, ottenendo:

$$s = \sqrt{s_{y\_val}\left(1 - \frac{lv+1}{N}\right)\left(h_{N+1} + \frac{s_{U,pr}}{s_{Utol,val}} + \frac{1}{N}\right)} \quad , \tag{1.29}$$

in cui *lv* è il numero di fattori del modello PLS. La stima di s_{y_val} è data dal SEC[2].

1.2.2 Approcci per il calcolo dei gradi di libertà

Il numero di gradi di libertà impiegati dal modello PLS è molto importante, in quanto è usato nella stima del SEC ed è coinvolto nel calcolo degli intervalli di fiducia nella *t* di *Student* (Zhang e Garcìa-Muñoz, 2009).

1.2.2.1 Naïve

Molto semplicemente, nell'approccio Naïve il numero di gradi di libertà è dato dal numero di variabili latenti *lv* utilizzate nella costruzione del modello PLS.

1.2.2.2 Pseudo gradi di libertà (PDF)

Un fattore del modello PLS può consumare più di un grado di libertà secondo van der Voet (1999). Il metodo PDF si basa sulla formula:

$$pdf = N\left(1 - \sqrt{\frac{(y_i - \hat{y}_i)^2_{FIT}}{(y_i - \hat{y}_i)^2_{CV}}}\right) \quad , \tag{1.30}$$

dove $(y_i - \hat{y}_i)^2_{FIT}$ è la somma dei quadrati degli errori di *fitting* del modello e $(y_i - \hat{y}_i)^2_{CV}$ è la somma dei quadrati degli errori di una convalida incrociata (*leave-one-out cross-validation*).

1.2.2.3 Gradi di libertà generalizzati (GDF)

L'approccio GDF definisce i gradi di libertà come la somma dei punti di leva (*leverage*), definiti come la sensitività di ogni risposta stimata alle perturbazioni sulle corrispondenti risposte osservate. La procedura coinvolge *k* iterazioni, ad ogni iterazione *k* la risposta osservata y_i viene perturbata con del rumore $\sigma_{i,k}$:

$$y^*_{i,k} = y_i + \sigma_{i,k} \qquad k = 1 \ldots K \quad . \tag{1.31}$$

Il modello PLS viene ricostruito tra le $y^*_{i,k}$ (*y i*-esima a cui è stato aggiunto il rumore $\sigma_{i,k}$) e le **U** di calibrazione in modo da ottenere le $\hat{y}^*_{i,k}$. I *leverage* $\hat{h}_i$ per l'oggetto *i*-esimo sono ottenuti dal fitting lineare tra le *K* coppie di $(\hat{y}^*_{i,k}, \sigma_{i,k})$ secondo l'equazione:

$$\hat{y}^*_{i,k} = \varphi + \hat{h}_i \sigma_{i,k} \qquad k = 1 \ldots K \quad . \tag{1.32}$$

Infine il numero di gradi di libertà è dato da:

$$gdf = \sum_{i=1}^{N} \hat{h}_i \quad . \tag{1.33}$$

Il rumore è rumore bianco normalmente distribuito con media 0 e deviazione standard $0.6\hat{\sigma}$, dove $\hat{\sigma}$ è dato dall'RMSECV di una *leave-one-out cross-validation* (come indicato dallo studio di Baumann e Stiefl, 2004).

1.3 Inversione nei modelli di regressione lineare

Nell'ambito della progettazione di prodotto l'inversione dei modelli di regressione lineare risulta essere uno strumento molto utile per lo sviluppo di nuovi prodotti e per la determinazione delle condizioni operative del processo che li produce (Tomba *et al.*, 2012).
L'obiettivo è impiegare il modello per trovare le migliori combinazioni di predittori (ad esempio, proprietà delle materie prime, parametri di processo, ecc.) che sono necessari per ottenere la risposta desiderata (ad esempio, qualità del prodotto).

1.3.1 L'inversione diretta e lo spazio nullo

Lo scopo dell'inversione dei modelli PLS di regressione lineare è utilizzare il modello per stimare un *set* di nuove condizioni di ingresso $\mathbf{u}_{NEW}$ (ad esempio proprietà delle materie prime nel caso di progettazione di prodotto o condizioni operative nel caso di progettazione di processo) corrispondenti al *set* di variabili di risposta desiderato $\mathbf{y}_{DES}$ (ad esempio qualità di prodotto, Figura 1.3).

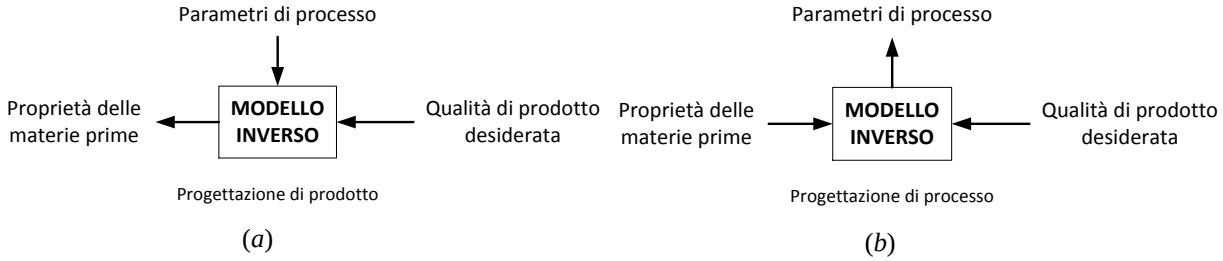


Figura 1.3: *Rappresentazione schematica dell'inversione di modelli di regressione lineari a) progettazione di prodotto b) progettazione di processo.*

Definendo con R_U il rango della matrice degli ingressi **U** e con R_Y il rango della matrice delle risposte **Y**, solitamente $lv=\max(R_U, R_Y)$, dove *lv* è il numero di variabili latenti selezionate. Il rango della matrice coincide con il numero delle variabili latenti che possono esser usate per descrivere la variabilità dei dati e può essere interpretato come il numero di forze agenti sul processo.
In base ai valori dei ranghi di **U** e **Y** è possibile distinguere tre casi (Jaeckle e MacGregor, 2000):

1. $R_U \leq R_Y$ (lo spazio latente di **U** è contenuto in quello di **Y**): tutte le variabili latenti di **U** hanno potenzialmente effetto sullo spazio di **Y**;

2. R_U=R_Y (la dimensione effettiva dello spazio di **U** è la medesima di quello di **Y**): se non esistono vincoli sulla soluzione $\mathbf{u}_{NEW}$, e se $\mathbf{y}_{DES}$ è completamente definita (mediante vincoli di uguaglianza), è possibile applicare l'inversione diretta del modello calcolando il vettore degli *score* collegato a $\mathbf{y}_{DES}$:

$$\mathbf{t}_{DES}^{T} = (\mathbf{Q}\mathbf{Q}^{T})^{-1}\mathbf{Q}^{T}\mathbf{y}_{DES}^{T} \quad . \tag{1.34}$$

Da $\mathbf{t}_{DES}$ si ricostruisce il vettore delle variabili di ingresso:

$$\mathbf{u}_{NEW} = \mathbf{t}_{DES}\mathbf{P}^{T} \quad . \tag{1.35}$$

Così definito, $\mathbf{u}_{NEW}$ appartiene allo spazio del modello e ha la stessa struttura della covarianza dei dati utilizzati per costruire il LVRM.

3. Se $R_U > R_Y$ (situazione che si verifica nella maggior parte dei casi), vi sono alcune variabili latenti nello spazio delle **U** che sono significative per la descrizione della variabilità sistematica in **U**, ma non contribuiscono alla spiegazione della variabilità nello spazio di **Y**. In pratica, queste direzioni tengono conto di una parte della variabilità dei dati **U** che non è collegata allo spazio di **Y**.

Queste variabili latenti formano il cosiddetto spazio nullo (*null space*), che rappresenta il luogo delle proiezioni di **U** che non hanno influenza sullo spazio delle risposte. Quindi, se esiste uno spazio nullo, la soluzione dell'inversione può teoricamente muoversi lungo esso senza interessare le qualità del prodotto; dunque, da *diverse* condizioni di processo che giacciono sullo spazio nullo si ottiene la *medesima* qualità finale di prodotto.

La procedura per il calcolo dello spazio nullo è stata descritta da Jaeckle e MacGregor (2000). Per garantire che $\mathbf{u}_{PRED}^{T} = \mathbf{u}_{NEW}^{T} + \mathbf{u}_{NULL}^{T}$ (dove $\mathbf{u}_{PRED}^{T}$ è la predizione del modello) mantenga la struttura della covarianza dei dati storici di **U** è necessario ricorrere alle proiezioni scalate (*score*) dello spazio di **U**:

$$\mathbf{u}_{PRED}^{T} = \left(\mathbf{t}_{DES}^{T} + \mathbf{t}_{NULL}^{T}\right)\mathbf{P}^{T} \quad . \tag{1.36}$$

Aggiungere $\mathbf{u}_{NULL}^{T} = \mathbf{t}_{NULL}^{T}\mathbf{P}^{T}$ a $\mathbf{u}_{NEW}^{T} = \mathbf{t}_{DES}^{T}\mathbf{P}^{T}$ non altera i valori di $\mathbf{y}_{DES}$, che sono determinati da $\mathbf{u}_{NEW}^{T}$.

Questo richiede che:

$$\mathbf{t}_{NULL}^{T}\mathbf{Q}^{T} = 0 \quad . \tag{1.37}$$

Qualsiasi $\mathbf{t}_{NULL}^{T}$ che giace sullo spazio nullo di $\mathbf{Q}^{T}$ è una soluzione della (1.35). La decomposizione ai valori singolari di $\mathbf{Q}^{T}$ rivela questo spazio:

$$\mathbf{Q}^{T} = \left[\mathbf{G}_1 \vdots \mathbf{G}_2\right]\mathbf{\Sigma}_{\mathbf{Q}^{T}}\mathbf{V}_{\mathbf{Q}^{T}}^{T} \quad . \tag{1.38}$$

$\mathbf{G}_2^{\mathrm{T}}$ definisce lo spazio in cui $\mathbf{t}_{\mathrm{NULL}}^{\mathrm{T}}$ può giacere:

$$\underset{(1\times lv)}{\mathbf{t}_{\mathrm{NULL}}^{\mathrm{T}}} = \underset{(1\times (lv\text{-}V_y))}{\delta^{\mathrm{T}}} \cdot \underset{((lv\text{-}V_y)\times lv)}{\mathbf{G}_2^{\mathrm{T}}} \quad . \tag{1.39}$$

Matematicamente, δ^{T} è una costante arbitraria in grandezza e direzione (Jaeckle e MacGregor, 2000).
A questo punto le $\mathbf{u}_{\mathrm{PRED}}^{\mathrm{T}}$ sono ottenute da:

$$\mathbf{u}_{\mathrm{PRED}}^{\mathrm{T}} = \left(\mathbf{t}_{\mathrm{DES}}^{\mathrm{T}} + \delta^{\mathrm{T}}\mathbf{G}_2^{\mathrm{T}}\right)\mathbf{P}^{\mathrm{T}} \quad , \tag{1.40}$$

e sono situate nell'intervallo dei dati storici di **U**.

1.4 Incertezza di predizione e inversione di modello

L'elemento di novità presentato in questa Tesi è l'utilizzo simultaneo dei metodi per la stima dell'incertezza su una predizione e delle procedure per l'inversione di modelli PLS, in modo da fornire un supporto alla progettazione di prodotto.
Il risultato dell'inversione di un modello di predizione è determinare le variabili di ingresso (ad esempio, qualità delle materie prime o parametri di processo) in grado di fornire una risposta desiderata (ad esempio, qualità di prodotto) attraverso l'inversione del modello che lega i predittori alle risposte. A questo *framework* generale viene aggiunto il concetto di incertezza nella determinazione degli ingressi. Scopo della Tesi è cioè fornire degli strumenti per la progettazione di prodotto che tengano conto dell'incertezza con la quale gli ingressi calcolati possono portare all'ottenimento del prodotto desiderato.
Si intende in particolare sviluppare una procedura, riportata in Figura 1.4, che permetta di ottenere un set di variabili di ingresso in grado di restituire la risposta desiderata (y_{DES}) e calcolare l'incertezza sugli ingressi necessari per ottenerla[2].

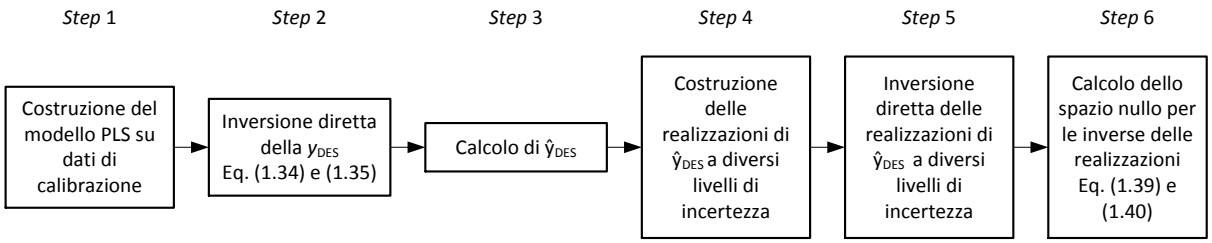


Figura 1.4: *Utilizzo dei metodi sull'incertezza con l'inversione diretta.*

[2] Ciò che viene invertito è il modello; questo modello inverso sfrutta poi come ingresso la variabile risposta desiderata y_{DES}. Qui e nel proseguo della Tesi, per brevità e leggerezza di forma, si farà riferimento a questo come 'inversione di una *y*' o 'applicazione dell'inversione ad una *y*'.

Vengono qui spiegati con maggior dettaglio i diversi passaggi nei quali la procedura si articola:

- *step* 1, costruzione del modello PLS: si costruisce il modello PLS sui dati di calibrazione in esame;
- *step* 2, inversione diretta: l'inversione del modello a partire da una y_{DES} selezionata, secondo le (1.34) e (1.35), restituisce le corrispondenti $\mathbf{u}_{NEW}$;
- *step* 3, calcolo di $\hat{y}_{DES}$: la $\mathbf{u}_{NEW}$ ottenuta nello *step* precedente viene proiettata nello spazio delle **y**, in modo da ottenere $\hat{y}_{DES}$, sulla quale è possibile calcolare la deviazione standard *s* dell'errore di predizione secondo la (1.26);
- *step 4*, applicazione dei metodi sull'incertezza: una volta stimata l'incertezza su $\hat{y}_{DES}$, si costruisce una distribuzione normale attorno al valore di $\hat{y}_{DES}$ con tale incertezza;
- *step* 5, inversione diretta della distribuzione di $\hat{y}_{DES}$: l'inversione di tutti i punti della distribuzione di $\hat{y}_{DES}$ produce una distribuzione per le variabili in ingresso $\mathbf{u}_{NEW}$, capace di restituire la $\hat{y}_{DES}$ con un'incertezza;
- *step* 6, calcolo dello spazio nullo: si calcola lo spazio nullo per tutte le $\hat{y}_{DES}$ della distribuzione (cioè per le realizzazioni di y_{DES} a diversi livelli di incertezza) secondo le procedure e i metodi descritti nel §1.3.1.

La procedura sopra descritta permette di ottenere delle realizzazioni delle variabili di ingresso **U** a diversi livelli di incertezza in grado di restituire la risposta desiderata. Le soluzioni ottenute si discostano dalla realtà sperimentale, in quanto vi è un'incertezza nella rappresentazione delle risposte **y** insita nel modello. L'obiettivo sarà dunque caratterizzare, attraverso l'incertezza, l'adeguatezza, la rappresentatività e l'accuratezza dell'inversione del modello.

Capitolo 2

Descrizione dei dati

Questo Capitolo contiene la descrizione dei dati che sono utilizzati nelle applicazioni dei metodi per la stima dell'incertezza e l'inversione. I casi di studio in esame sono tre: dati da un modello matematico, dati di una granulazione umida (Vemavarapu *et al*., 2009) e dati da un processo simulato di granulazione per compattazione a rulli.

2.1 Caso di studio 1: dati da modello matematico

Il primo caso di studio è costituito da un modello matematico, costruito artificialmente per verificare i metodi di valutazione dell'incertezza di una variabile di qualità di prodotto.

2.1.1 Descrizione dei dati del caso 1

Questo caso di studio è stato costruito per verificare le tecniche di valutazione dell'incertezza proposte da Zhang e Garcìa-Muñoz (2009) e per capire quali fossero gli approcci più promettenti. La matrice di ingresso **U** è costituita da 5 variabili, di cui solo le prime 2 indipendenti, secondo la seguente struttura:

$$\mathbf{U} = [\mathbf{u}_1; \mathbf{u}_2; \mathbf{x}_1; \mathbf{x}_2; \mathbf{x}_3];$$

$$\begin{cases} x_{n,1} = u_{n,1}^2 \\ x_{n,2} = u_{n,2}^2 \\ x_{n,3} = u_{n,1} \cdot u_{n,2} \end{cases} . \qquad (2.1)$$

$u_{n,1}$ e $u_{n,2}$ sono variabili manipolabili (indipendenti), generate secondo una distribuzione gaussiana di media 42 e 11 (rispettivamente per $u_{n,1}$ e $u_{n,2}$) e varianza 16 per la prima variabile e 3 per la seconda. Queste variabili vanno a costituire i vettori $\mathbf{u}_1$ [2000×1] e $\mathbf{u}_2$ [2000×1]. Le variabili $x_{n,1}$, $x_{n,2}$ e $x_{n,3}$ sono variabili non manipolabili e dipendenti dalle $u_{n,i}$, per questo collineari fra loro e con le $u_{n,i}$. Le $x_{n,i}$ vengono raccolte nei vettori $\mathbf{x}_1$ [2000×1], $\mathbf{x}_2$ [2000×1] e $\mathbf{x}_3$ [2000×1].
La risposta **y** è univariata e definita dal seguente modello:

$$\mathbf{y}_0 = k_0 + k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + k_3\mathbf{x}_1 + k_4\mathbf{x}_2 + k_5\mathbf{x}_3 \quad , \qquad (2.2)$$

con coefficienti $\mathbf{k} = [k_1; k_2; k_3; k_4; k_5; k_6] = [-21; 4.3; 0.022; 0.0064; 1.1; -0.12]$. I coefficienti **k** del modello sono stati scelti in maniera casuale visto che lo *scaling* dei dati rende la scelta dei coefficienti un semplice peso che viene assegnato alle variabili; dunque optare per altri *set* di parametri non modifica i risultati.

Al fine di studiare la sensitività dei risultati al rumore, in questo studio sono stati utilizzati diversi vettori **y**, ottenuti aggiungendo a $\mathbf{y}_0$ del rumore bianco con media 0 e con varianza di ampiezza pari a diverse percentuali della deviazione standard del valore originario di $\mathbf{y}_0$: 10%, 40% e 70%. Nel seguito della Tesi ci si riferirà alle risposte corrotte da diversi livelli di rumore con $\mathbf{y}_{10}$, $\mathbf{y}_{40}$ e $\mathbf{y}_{70}$.

Si sono generati 2000 campioni sia per la matrice **U** che per il vettore **y**, suddivisi in un *set* di calibrazione (primi 1000 campioni) e in uno di convalida (secondi 1000 campioni).

2.2 Caso di studio 2: dati di granulazione umida

Il secondo caso di studio affrontato in questa Tesi riguarda dati di un processo di granulazione umida, riportati nel lavoro di Vemavarapu *et al*. (2009). Si è scelto questo *set* di dati in quanto tratta un caso reale di interesse farmaceutico, con uno studio che coinvolge un ampio spettro di principi attivi ed eccipienti. Questo caso di studio raccoglie dati associati ad un esempio di progettazione di prodotto, con un numero ridotto di campioni disponibili, che rispecchiano quindi una sperimentazione reale.

2.2.1 Il processo di granulazione ad umido

La granulazione ad umido prevede l'utilizzo di una fase bagnante (acqua o solventi organici) e di processi di lavorazione atti a stabilizzare il granulato finale (ad esempio, essiccamento, sinterizzazione). Le particelle si uniscono le une alle altre per mezzo di forze capillari e viscose, che si sostituiscono poi con veri e propri legami solidi durante l'essiccamento o la sinterizzazione. In generale le fasi della granulazione ad umido si possono suddividere in (Giannetti, 2012):

- pretrattamento: consente di uniformare le caratteristiche dei materiali di partenza attraverso operazioni come setacciamento e macinazione. Non sempre è necessario, dipende dal tipo di materiale;
- miscelazione: le polveri da processare vengono miscelate;
- bagnatura: consiste nell'umidificare e impastare le polveri con un'adeguata quantità di liquido, in modo da conferire alle particelle le corrette proprietà di adesione;
- accrescimento: si formano i granuli e dipende dalla tecnica utilizzata;
- essicamento: si ha il consolidamento dei granuli formati; avviene a temperatura controllata, per salvaguardare il prodotto da processi degradativi;

- calibrazione: è la setacciatura dei granuli secchi, per uniformare le dimensioni del prodotto desiderato.

La granulazione ad umido è ampiamente utilizzata in ambito farmaceutico (McCormick, 2005) ed è particolarmente utile quando è necessaria una distribuzione uniforme di un componente (ad esempio, principio attivo, leganti, agenti bagnanti, ecc.) presente in minor quantità (ad esempio, <5% in peso) all'interno del *bulk* della formulazione per garantire le prestazioni desiderate del farmaco.

2.2.2 Descrizione dei dati del caso 2

I dati disponibili sono raccolti in due *set*: un *set* **U** [25×7] di proprietà di ingresso (sono proprietà delle 25 materie prime in ingresso al processo), per ciascuno delle quali esistono 7 caratteristiche misurate, e un *set* **Y** [25×7] di proprietà della qualità del granulato prodotto, che comprende 25 prodotti corrispondenti ai materiali di ingresso con 7 diverse variabili di qualità misurate. Le materie prime sono state caratterizzate e processate fissando le condizioni operative a valori costanti, in modo da fornire le informazioni necessarie per studiare l'effetto delle proprietà degli ingressi sul prodotto finale (i dettagli sono disponibili nel lavoro originale). Le prove sono state condotte in modo da coprire un ampio spettro di variabilità delle proprietà di interesse, garantendo allo stesso tempo la processabilità dei materiali nella granulazione. Le proprietà degli ingressi e le qualità misurate per caratterizzare il prodotto sono riportate in Tabella 2.1.

Tabella 2.1: *Proprietà misurate per i materiali di ingresso e per i prodotti ottenuti*.

$\mathbf{u}_i$	$\mathbf{y}_i$
1 Solubilità in H_2O (mg/mL)	1 Perdita per essicamento (%)
2 Angolo di Contatto (gradi)	2 *Oversize* (%)
3 Capacità di ritenzione dell'H_2O (guadagno % in peso)	3 Δ(*Flodex*) (mm)
4 *D* [3,2] (μm)	4 Δ(Compattabilità) (kPa/MPa)
5 *D*90/*D*10	5 *D* [3,2] (μm)
6 Area superficiale (m^2/g)	6 *D*90/*D*10
7 Volume dei pori (cm^3/g)	7 Rapporto di crescita

Le variabili *D* [3,2] (diametro medio di Sauter) e *D*10/*D*90 (ampiezza di distribuzione, *distribution span*) sono rappresentative della distribuzione di dimensione delle particelle (*particle size distribution*, PSD) e sono state raccolte sia per i materiali di ingresso che per le uscite in quanto vengono modificate dal processo di granulazione ad umido. La variabile *oversize* indica la percentuale di granuli aventi dimensione maggiore di una certa quantità prefissata. *Flodex* è un indice rappresentativo della capacità di flusso di polveri, misurato con appositi *tester*. Le variazioni dovute alla granulazione sono state misurate confrontando le

proprietà del granulato con quelle della miscela di polveri che non ha subito il processo di granulazione ad umido (*preblend*) e quindi avente medesima composizione chimica del granulato, ma diverse proprietà di flusso e meccaniche. Tali grandezze sono indicate con Δ in Tabella 2.1, indice della differenza tra le proprietà del granulato e della *preblend*. Il rapporto di crescita è invece dato dal rapporto tra il *D* [3,2] del prodotto granulato e quello di un materiale preso come riferimento.

2.3 Caso di studio 3: dati di granulazione da una simulazione con *gSolids™*

Il terzo caso di studio trattato in questa Tesi riguarda un processo di granulazione a secco di cellulosa microcristallina mediante un compattatore a rulli (*roller compactor*). I dati disponibili sono stati ottenuti da simulazioni effettuate con il *software gSolids™* (v. 3.0, *Process System Enterprise Ltd*., Londra, U.K.).

2.3.1 Il processo di compattazione a rulli

La compattazione a rulli è un processo di granulazione a secco continuo largamente utilizzato nell'industria farmaceutica, chimica, mineraria e alimentare per la produzione di agglomerati (Guigon *et al.*, 2007). La compattazione a rulli è relativamente semplice: l'alimentazione in polvere è fatta fluire attraverso due cilindri contro-rotanti; questa polvere è soggetta ad alta pressione nello stretto spazio tra i cilindri, consentendo la realizzazione di un prodotto compattato nella forma di striscia continua o di mattonelle separate. Questo può rappresentare il prodotto finale, ma spesso, nell'industria farmaceutica, l'agglomerato viene ridotto di dimensione mediante macinazione e setacciamento per ottenere una polvere che fluisca facilmente nelle successive operazioni di processo (ad esempio, formazione delle compresse - *tableting*). La compattazione viene solitamente condotta in maniera continua, ma è adattabile a processi *batch* o *semi-batch*, a differenza delle altre tecniche di granulazione che non prevedono processi continui.

Wennerstrum (2000) elenca i vantaggi del processo:

- miscele uniformi: si producono granuli di consistenza uniforme minimizzando i problemi di segregazione dovuti a differenze di dimensione delle particelle, forma e densità;
- particelle di dimensioni uniformi: la compattazione a rulli aiuta nella produzione di agglomerati di dimensioni specifiche;
- miglioramento delle proprietà di flusso: il granulato presenta proprietà migliori rispetto la polvere di alimentazione;
- controllo delle polveri fini: riduce la produzione di scarti e aumenta la sicurezza per gli operatori esposti;

- aumento della densità di *bulk*: l'aumento di densità facilità il trasporto e lo stoccaggio;
- sostanze sensibili all'umidità: il processo non richiede leganti liquidi ed è quindi adatto a trattare sostanze sensibili all'umidità;
- sostanze termosensibili: la granulazione a secco non necessita di successivi stadi di essicamento e può essere utilizzato per sostanze termosensibili.

Il meccanismo chiave nella compattazione a rulli è che l'agglomerazione delle particelle deriva solamente dalle forze di compressione; per questo, la scelta della polvere da processare diventa un elemento critico. In questo caso un ruolo fondamentale è giocato dagli eccipienti. Gli eccipienti sono ingredienti farmacologicamente inerti (ovvero tutto ciò che non è il principio attivo) aventi lo scopo di ottenere un prodotto con le caratteristiche tecnologiche desiderate. Ciascuno di essi esplica una funzione ben definita (oppure più funzioni) nella formulazione della polvere. Quindi, seppure inerti farmacologicamente, essi non sono inerti dal punto di vista tecnologico. Eccipienti molto usati sono la cellulosa e i suoi derivati, che hanno la caratteristica di essere biocompatibili, chimicamente inerti e di avere buona comprimibilità, buone proprietà di disgregazione, buona compatibilità con molti farmaci. La cellulosa è spesso usata nella sua forma microcristallina (in parte cristallina, in parte amorfa), in cui il grado di cristallinità dipende dalla provenienza e dal metodo di preparazione. Dal grado di cristallinità dipendono le caratteristiche di comprimibilità e igroscopicità. Ne esistono numerosi tipi a seconda dello stato di aggregazione e della granulometria (Giannetti, 2012).

Secondo Johanson (1965) è possibile distinguere tre regioni differenti nella compattazione a rulli, in cui il materiale si comporta in maniera diversa: regione di scivolamento (*slip region*), regione di *nip* e regione di rilascio (*release region*) (Figura 2.1).

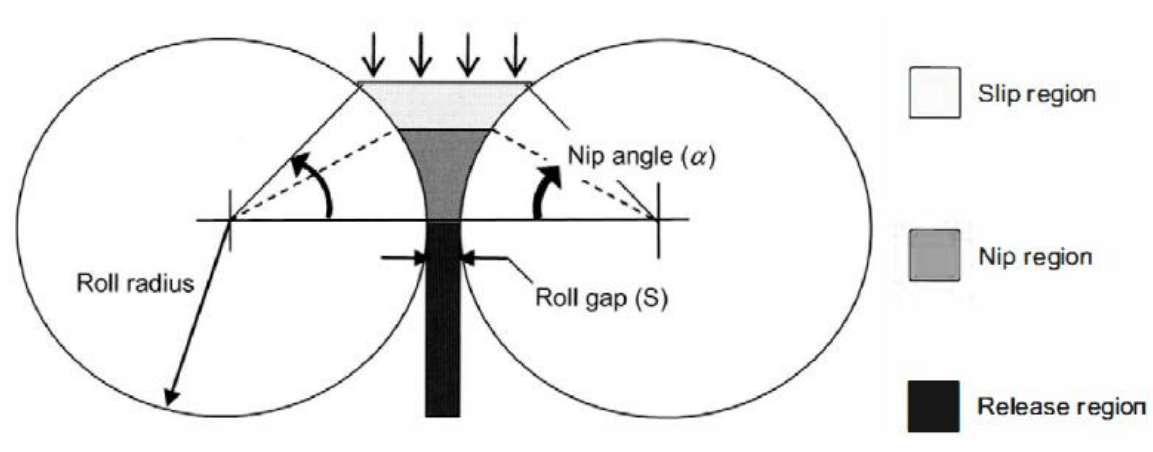


Figura 2.1: *Rappresentazione schematica del processo di compattazione a rulli e zone principali di divisione. Adattato da Guigon et al.(2007), p. 258.*

La regione di scivolamento è posizionata prima della regione di *nip* ed è caratterizzata dal fatto che in questa zona le particelle alimentate scivolano sulla superficie del rullo. La

pressione esercitata sul materiale è relativamente bassa e il comportamento della polvere dipende dall'attrito con i rulli e tra le particelle stesse. La regione di *nip* inizia all'angolo α_{NIP} (angolo di *nip*), dove la velocità di parete delle particelle coincide con quella dei rulli. L'aumento di densità è dovuto alla riduzione del *gap* tra la superficie dei cilindri. L'ultima regione, quella di rilascio, comincia quando lo spazio tra i rulli riprende ad aumentare; in questa zona lo spessore del compattato potrebbe crescere di nuovo a seconda dell'elasticità del granulato.

2.3.2 La modellazione della compattazione a rulli

Il modello poi utilizzato nel *software gSolids™* per la caratterizzazione del granulato in uscita dal compattatore a rulli è stato proposto da Johanson (1965)[3]. Il modello predice l'aumento di densità, causato dall'azione dei due rulli, del materiale in uscita dal compattatore ed è costituito da tre diverse sezioni:

- calcolo dell'angolo di *nip*;
- stima della pressione massima;
- determinazione della densità e della porosità del granulato in uscita.

2.3.2.1 Calcolo dell'angolo di *nip*

Uno dei parametri più importanti, su cui di basa il computo delle altre grandezze, è l'angolo di *nip*, che avviene secondo l'equazione:

$$\frac{4\sigma_\theta(\pi/2-\theta-\nu)\tan\delta_{EFF}}{\frac{D}{2}\left[1+\frac{S}{D}-\cos\theta\right]\left[\cot\left(\frac{\theta+\nu+\pi/2}{2}-\mu\right)-\cot\left(\frac{\theta+\nu+\pi/2}{2}+\mu\right)\right]}= \tag{2.3}$$

$$=\frac{\kappa\sigma_\theta\left[2\cos\theta-1-\frac{S}{D}\right]\tan\theta}{\frac{D}{2}\left[\frac{d}{D}+\left(1+\frac{S}{D}-\cos\theta\right)\cos\theta\right]} \quad ,$$

in cui σ_θ è lo sforzo normale medio [Pa], θ è la differenza tra l'angolo formato dal piano orizzontale e l'angolo di *nip* [rad], δ_{EFF} è l'angolo effettivo di attrito [rad], ν è l'angolo acuto tra la direzione principale di stress e la tangente alla superficie del cilindro [rad], *D* è il diametro del cilindro [m], *S* è il *gap* tra i due rulli [m], μ è l'angolo di scivolamento [rad], κ è la costante di comprimibilità del granulato solido [-], e *d* è lo spessore della bricchetta solida quando *S* = 0 [m]. Il membro a sinistra della (2.3) descrive il gradiente di pressione che si deve generare affinché ci sia scivolamento lungo la superficie del cilindro. Il termine a destra

[3]Il modello non è predittivo per quel che riguarda la PSD del granulato prodotto. Quindi, anche se nella realtà riveste un ruolo piuttosto importante, la PSD del materiale in ingresso non influisce sul compattato finale.

fornisce il gradiente di pressione quando non c'è *slip*. L'uguaglianza dei due termini, garantita dalla continuità del gradiente di pressione, permette di calcolare l'angolo di *nip*, risolvendo l'equazione rispetto θ.

2.3.2.2 Calcolo della pressione massima

La massima pressione esercitata dal compattatore a rulli è calcolata come segue:

$$P_{\max} = \frac{2F_{\text{roll}}}{WDF} \quad , \tag{2.4}$$

dove $P_{\max}$ è la pressione massima tra i due cilindri [Pa], F_{roll} è la forza applicata dal cilindro [N] e W è l'ampiezza del cilindro [m]. F è una variabile adimensionale, così definita:

$$F = \int_{\theta=0}^{\theta=\alpha_{\text{NIP}}} \left[\frac{(d+S)/D}{d/D + (1 - S/D - \cos\theta)\cos\theta} \right] \cos\theta d\theta \quad , \tag{2.5}$$

in cui α_{NIP} è l'angolo di *nip*.

2.3.2.3 Calcolo della densità e della porosità in uscita

La variabile $P_{\max}$ è utilizzata per calcolare la densità massima, $\rho_{\max}$, a cui il materiale può essere compattato. Questo valore è utilizzato come riferimento per fissare i due gradi di libertà necessari a definire la densità particellare, la densità di *bulk* e la porosità. La logica è la seguente:

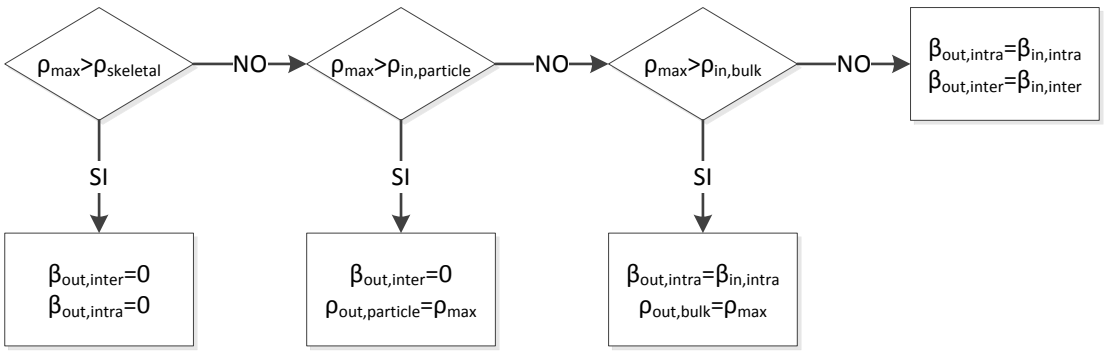


Figura 2.3*: Step logici per la determinazione di densità e porosità del granulato solido in uscita dal compattatore a rulli.*

In Figura 2.3 con $\rho_{\max}$ si è indicata la densità massima ottenuta dal *roller compactor* [kg/m^3], ρ_{skeletal} è la densità vera della polvere [kg/m^3], $\rho_{\text{in,particle}}$ è la densità della particella all'ingresso [kg/m^3], $\rho_{\text{out,particle}}$ è la densità della particella all'uscita [kg/m^3], $\rho_{\text{in,bulk}}$ e $\rho_{\text{out,bulk}}$ sono le densità di *bulk* all'ingresso e all'uscita del compattatore rispettivamente [kg/m^3], $\beta_{\text{out,inter}}$ è la porosità esterna dei solidi in uscita [m^3/m^3], $\beta_{\text{in,inter}}$ è la porosità esterna dei solidi in entrata [m^3/m^3],

$\beta_{out,intra}$ è la porosità interna dei solidi in uscita dal compattatore a rulli [m^3/m^3] e $\beta_{in,intra}$ è la porosità interna dei solidi in entrata [m^3/m^3].
Il modello di compattatore a rulli del *software* consente di utilizzare come variabili di ingresso le grandezze riportate in Tabella 2.2.

Tabella 2.2: *Variabili di ingresso del modello di compattatore a rulli implementato nel software gSolids™.*

Variabili di ingresso	Unità di misura	Nome e descrizione fisica
κ	[-]	costante di comprimibilità, capacità di un corpo di resistere ad una forza di compressione uniforme
d	[m]	spessore della bricchetta solida quando S=0
D	[m]	diametro del cilindro
W	[m]	lunghezza del cilindro
v_{roll}	[rpm]	velocità del rullo
S_{max}	[m]	*gap* massimo tra i cilindri
F_{roll}	[kN]	forza applicata dal rullo
δ_{FR}	[rad]	angolo di attrito tra solidi e cilindro
δ_{EFF}	[rad]	angolo di attrito effettivo
δ_{FEED}	[rad]	angolo di alimentazione
F_{sb}	[-]	fattore di *springback*, è il rapporto tra gli angoli finali ed iniziali di un materiale piegato e rappresenta la tendenza di quest'ultimo a tornare alla sua forma originaria
E_{add}	[J]	energia aggiunta durante la compattazione

dove v_{roll} è la velocità del cilindro [rpm], S_{max} il *gap* massimo tra i rulli [m] (diverso da quello effettivo, inserito solo per ragioni computazionali), δ_{FR} l'angolo di attrito tra solidi e cilindro [rad], δ_{FEED} l'angolo di alimentazione della polvere [rad], F_{sb} il fattore di *springback* [-] e E_{add} è l'energia aggiunta durante la compattazione [J].

2.3.3 Descrizione dei dati del caso di studio 3

I parametri di processo più importanti sono la velocità del rullo, la forza che è in grado di applicare e il *gap* tra i due cilindri (Souihi *et al.*, 2013).
Dato che il *gap* effettivo non è manipolabile nel *software* disponibile ed escluse le grandezze che non hanno influenza sull'agglomerato, si sono considerate come variabili di ingresso: κ, D, W, v_{roll}, F_{roll}, δ_{FR}, δ_{EFF} e F_{sb}. Quattro di queste sono caratteristiche del materiale (κ, δ_{FR}, δ_{EFF} e F_{sb}), mentre le rimanenti (D, W, v_{roll} e F_{roll}) dipendono dal macchinario utilizzato. Il prodotto compattato è caratterizzato in termini di $\rho_{out,particle}$, $\rho_{out,bulk}$ e $\beta_{out,intra}$ (variabili tra le

quali si andrà a scegliere la grandezza che costituisce la risposta, visto l'utilizzo di una risposta monovariata nei metodi per la stima dell'incertezza su una predizione).

I dati riguardanti le proprietà del materiale sono stati costruiti in modo tale da garantire una correlazione positiva tra angolo di attrito δ_{FR} e angolo di attrito effettivo δ_{EFF}; si è invece imposta una anti-correlazione (coefficiente di correlazione negativo) tra la costante di comprimibilità e il fattore di *springback*, κ e F_{sb} rispettivamente (Guigon *et al.*, 2007).

Nella costruzione dei dati si sono fatte le seguenti ipotesi (riassunte in Tabella 2.3):

- si hanno a disposizione due diversi macchinari, con due diverse lunghezze di cilindro *W* (costanti per macchina, ricavati da schede tecniche di case costruttrici) e diversi diametri di cilindro *D*, di cui uno in comune;
- si è in possesso di cinque diversi tipi di cellulosa microcristallina con diverse caratteristiche (le proprietà hanno differenti distribuzioni granulometriche in media e deviazione standard, Tabella 2.4). Ogni prova effettuata differisce per del rumore bianco sulla misura;
- tre materiali vengono processati da entrambi i macchinari;
- si applicano velocità e forze su diversi livelli discreti, di cui quattro in comune tra i due macchinari.

Tabella 2.3: *Riassunto delle ipotesi fatte per costruire i dati di granulazione con il software gSolids™.*

	Macchinario 1	**In comune**	**Macchinario 2**
Ampiezze del cilindro [m]	0.12	-	0.15
Diametro dei rulli [m]	0.3	0.4	0.5
Materiali processati	1	2, 3, 4	5
Velocità [RPM]	10	2, 6.5, 15.5, 20	13
Forze applicate·10^{-3} [KN]	17	4, 9, 14, 24	20

Tabella 2.4: *Proprietà medie dei materiali processati dai macchinari nelle simulazioni con gSolids™.*

	κ [-]	**δ_{FR} [°]**	**δ_{EFF} [°]**	**F_{sb} [-]**	**$\sigma_{scenario\ 1}$**	**$\sigma_{scenario\ 2}$**	**$\sigma_{scenario\ 3}$**
Materiale 1	8	20	32	0.1250	0.4	4	5
Materiale 2	9	30	48	0.1111	0.6	5	6
Materiale 3	10	25	40	0.1	0.7	6	6.5
Materiale 4	14	40	64	0.0714	0.5	3	5.5
Materiale 5	6	20	32	0.1667	0.4	4	5

Sono stati costruiti diversi *set* di misure, che differiscono solo per la variabilità dei dati contenuti in modo da ottenere tre differenti scenari:

1. i materiali hanno caratteristiche ben distinguibili tra loro;
2. si riconoscono solo due classi di materiali;
3. i materiali sono indistinguibili.

Questo permette di studiare la sensibilità alla variabilità delle misure.

Per il primo scenario si sono svolte 56 prove differenti per ciascun macchinario, 14 per materiale, raccogliendo per ogni prova le variabili di ingresso e i risultati in termini di densità particellare, densità di *bulk* e porosità interna del granulato.

Il secondo scenario contiene 26 diverse simulazioni per il primo macchinario e 24 per l'altro.

Il terzo e ultimo *set* di dati è composto da 22 campionamenti per il primo macchinario e 23 per il secondo macchinario.

Le prove effettuate per ciascun macchinario sono state riunite in un unico *set* di campionamenti, diviso poi in un *set* di calibrazione e in uno di convalida per ognuno dei diversi scenari ipotizzati.

Capitolo 3

Analisi esplorativa dei modelli predittivi della qualità

Questo Capitolo presenta un'analisi esplorativa dei casi di studio presentati nel Capitolo 2 mediante modelli a variabili latenti che stimino la qualità finale del prodotto. In particolare, si descrivono la struttura dei modelli e le correlazioni tra le diverse variabili dei dati di ingresso.

3.1 Caso di studio 1: analisi esplorativa dei dati da modello matematico

In tutti i casi di studio si è usata PCA per trovare il rango statistico della matrice dei dati di ingresso **U**, a cui corrisponde il numero corretto di variabili latenti con cui costruire il modello PLS (visto l'utilizzo di una matrice monovariata delle risposte **y**). Dopo la costruzione del modello PLS, si verifica che sia in grado di descrivere le correlazioni attese tra le variabili e si verificano le capacità predittive del modello stesso. Inoltre, in questo caso di studio diversi modelli PLS sono stati costruiti variando il numero di campioni disponibili in fase di calibrazione (numero di campioni di calibrazione: 1000, 100, 50, 20 e 10) al fine di valutare la sensitività dei modelli al numero di dati.
Nel modello preliminare PCA, il numero di variabili latenti è stato scelto mediante la regola dell'autovalore λ (§1.1.1.1). Nel caso di *autoscaling* il numero sufficiente di variabili latenti risulta essere 2, che è in grado di spiegare più del 98% della varianza degli ingressi **U**. In Tabella 3.1 si riportano i risultati di una PCA sui dati **U** autoscalati.

Tabella 3.1*: PCA per i dati* **U** *di ingresso, autovalori e varianza spiegata.*

lv	λ	Varianza spiegata (%)	Varianza cumulativa (%)
1	2.87	57.46	57.46
2	2.04	40.76	98.22
3	0.0539	1.08	99.29
4	0.0212	0.42	99.72
5	0.0141	0.28	100

La Figura 3.1 riporta il diagramma degli *score* del modello PLS per il *set* di calibrazione. Si può notare come la distribuzione dei dati sia multinormale e, come atteso, il 95% circa dei

dati sia situata all'interno dell'ellissi di fiducia al 95% definita nel §1.1.1.2, a riprova della bontà del modello costruito.

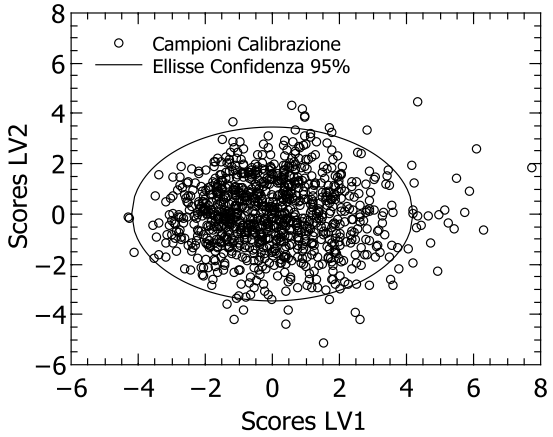


Figura 3.1: *Diagramma degli score del modello PLS con 1000 dati calibrazione e la matrice* $\mathbf{y}_0$.

Come è possibile vedere dal diagramma dei *loading* di Figura 3.2, tutte le variabili hanno un peso positivo sulla LV1 (con maggiore peso della variabile numero 5, che però presenta peso molto minore sulla LV2). Inoltre si nota la correlazione tra le variabili 1-3 e 2-4, che dominano sulla LV2 rispettivamente con peso negativo e positivo. Questo conferma che il modello a variabili latenti è in grado di descrivere correttamente la struttura di correlazione tra le variabili: la variabile 1 ($\mathbf{u}_1$) risulta molto correlata con la 3 ($\mathbf{u}_1^2$), la variabile 2 ($\mathbf{u}_2$) molto correlata con la 4 ($\mathbf{u}_2^2$). Il modello inoltre descrive la scarsa correlazione tra 1-3 e 2-4, che infatti sono indipendenti.

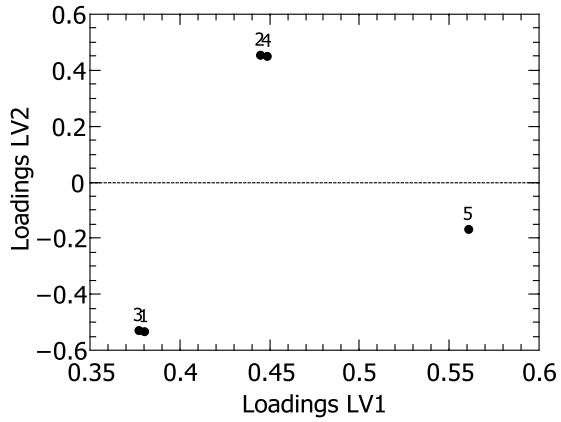
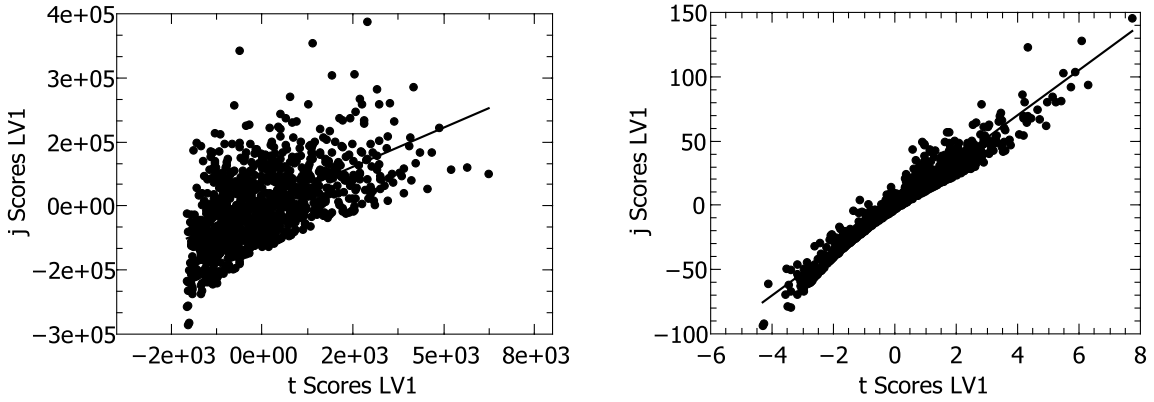


Figura 3.2: *Diagramma dei loading del modello PLS avendo utilizzato 1000 dati in fase di calibrazione e la matrice* $\mathbf{y}_0$.

Infine si noti che il *preprocessing* dei dati riveste un ruolo particolarmente importante nell'ottenimento di risultati soddisfacenti. Una verifica per valutare il *preprocessing* migliore viene fatta controllando che la relazione interna tra gli *score* di **U** e quelli di **y** (Equazione

1.15) sia il più lineare possibile (Geladi e Kowalski, 1996). È stato verificato che l'*autoscaling* dei dati, rispetto al solo *mean centering*, favorisce la linearità del modello che viene calcolato (Figura 3.3), garantendo così maggiore accuratezza ai risultati e rappresentatività al modello.



(*a*) (*b*)

Figura 3.3: *Diagramma della relazione interna t-score/j-score (1000 dati in calibrazione, assenza di rumore) a) mean centering b) autoscaling.*

In più l'*autoscaling*, rispetto il *mean centering*, garantisce la costruzione di modelli PLS con un numero minore di LV.

3.2 Caso di studio 2: analisi esplorativa dei dati di granulazione umida

Siccome i dati presentano alcuni elementi mancanti, come stadio preliminare si è utilizzata la procedura di completamento di Walczak e Massart (2001) per rimpiazzare i dati mancanti.
I dati sono stati divisi in un *set* di calibrazione di 20 campioni e uno di convalida comprensivo di 5 campioni.
Come nel caso precedente si è svolta un'analisi preliminare dei dati per comprendere quale sia il rango statistico delle matrici, cioè per capire il numero di variabili latenti necessarie a descrivere i dati. L'analisi è stata svolta sul *set* di calibrazione dei dati di ingresso **U**. In accordo con lo studio di Tomba *et al.* (2012), il numero di variabili latenti è stato scelto attraverso una PCA sui dati **U**, di cui la Tabella 3.2 indica i risultati.

Tabella 3.2: *PCA per i dati* **U** *del set di calibrazione, autovalori e varianza spiegata.*

lv	λ	Varianza spiegata (%)	Varianza cumulativa (%)
1	2.63	37.62	37.62
2	1.47	21	58.62
3	1.27	18.19	76.81
4	1	14.33	91.14
5	0.312	4.46	95.6
6	0.164	2.35	97.95
7	0.143	2.05	100

Si è dunque scelto di costruire il modello PLS con 4 variabili latenti, valore a cui corrispondono autovalori λ maggiori di 1 e in grado di spiegare mediante PCA più del 91% della varianza dei dati sulle **U**.

Si è scelto di utilizzare una matrice delle risposte **y** monovariata (e quindi descritta da un'unica LV) per poter sfruttare gli approcci di calcolo sull'incertezza di una predizione in un modello PLS. La variabile è stata selezionata tra quelle che presentano dati completi e che garantiscono la migliore costruzione del modello PLS, cioè quella che con un modello a 4 LV presenta la maggiore varianza spiegata. In particolare è stata scelta la variabile 2 *Oversize*, descritta da oltre il 93% della varianza spiegata.

Tabella 3.3: *Varianza spiegata a 4* LV *per le variabili che presentavano dati completi di* **Y**.

Variabile	Varianza spiegata a 4 LV
2 *Oversize* (%)	93.7%
4 Δ(Compattabilità) (kPa/MPa)	86%
5 *D* [3,2]	83.9%
6 *D*90/*D*10	74.3%
7 Rapporto di crescita	66.4%

Si è costruito il modello PLS e sui dati di convalida sono state applicate le procedure di calcolo dell'incertezza e di inversione.

In Figura 3.4 si riporta il diagramma degli *score* per i dati di calibrazione e per il *set* di convalida, dove è possibile vedere che i dati sono divisi in due diversi gruppi. I campioni del *set* di convalida sono stati selezionati in modo che fossero rappresentativi dei diversi gruppi, selezionando sia dati vicini alla media che campioni più esterni.

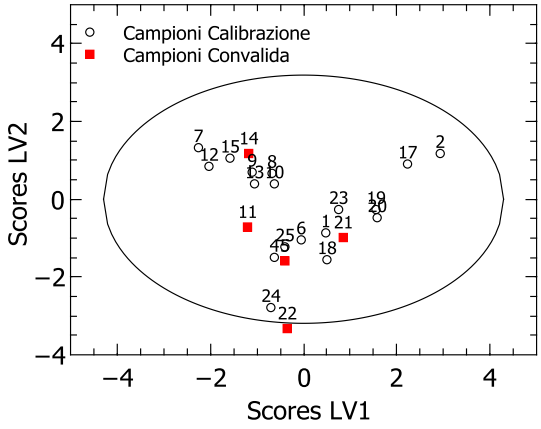


Figura 3.4: *Diagramma degli score del modello PLS del set di calibrazione e convalida per i dati di granulazione umida.*

Dal diagramma dei *loading* di Figura 3.5 (il numero delle variabili è riportato in Tabella 2.1) è possibile identificare le variabili che hanno maggiore peso nella costruzione del modello. La prima variabile latente, che spiega circa il 37% della varianza di **U**, è principalmente governata dalle variabili 1 e 4 (solubilità in acqua e *D*[3,2], rispettivamente) assieme alla variabile 6 (area superficiale), alla quale però sono anti-correlate. *D*[3,2] e *D*90/*D*10 (variabile 5) sono anti-correlate. Ciò significa che materiali con alta PSD (alto *D*[3,2]) hanno solitamente una PSD più stretta (basso *D*90/*D*10). La LV2, che rende conto del 21% circa della variabilità dei dati, è maggiormente influenzata dalle variabili 2 (angolo di contatto) e 7 (volume dei pori), che sono correlate negativamente alla variabile 3 (capacità di ritenzione dell'acqua). Questo dimostra come più alto è il volume dei pori in un materiale, minore è la sua capacità di trattenere l'acqua. Le LV3 e LV4 contribuiscono in misura minore a spiegare la varianza totale dei dati e per questo non si riporta il diagramma dei *loading*.

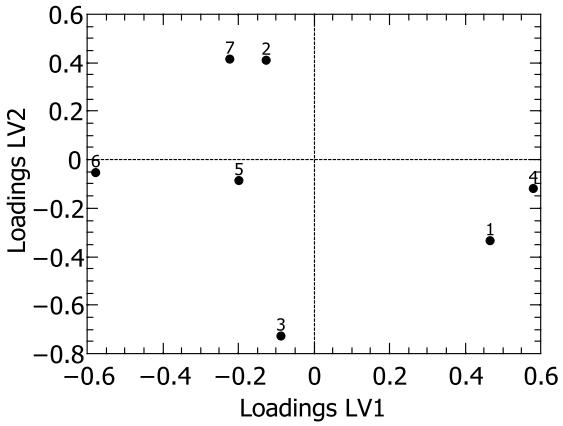


Figura 3.5: *Diagramma dei loading del modello PLS per le variabili latenti 1 e 2.*

3.3 Caso di studio 3: analisi esplorativa dei dati di granulazione mediante compattazione a rulli simulata

In questo Paragrafo vengono descritti i modelli di regressione lineare a variabili latenti ottenuti per i diversi scenari ipotizzati (che differiscono solo per la variabilità delle proprietà delle materie prime) nel caso di studio 3.

3.3.1 Modelli dello scenario 1

In questo caso sono disponibili 112 differenti misure, 56 per macchinario. Questo *set* è stato suddiviso in 100 campioni di calibrazione e 12 di convalida (6 per macchinario).
Un'analisi PCA preliminare dei dati di ingresso **U** (di calibrazione) identifica il rango statistico della matrice. La Tabella 3.4 riporta i risultati di una PCA sui dati **U**.

Tabella 3.4: *PCA per i dati* **U** *del set di calibrazione, autovalori e varianza spiegata.*

lv	λ	Varianza spiegata (%)	Varianza cumulativa (%)
1	3.68	46.03	46.03
2	1.72	21.47	67.5
3	1.11	13.93	81.43
4	0.859	10.74	92.17
5	0.324	4.05	96.21
6	0.27	3.38	99.59
7	0.0322	0.4	99.99
8	0.000623	0.01	100

Il modello PLS viene costruito con 4 LV, in grado di spiegare oltre il 92% della variabilità dei dati mantenendo autovalori prossimi all'unità.
Per quanto riguarda la variabile di risposta **y**, si sono raccolte le misure riguardanti la porosità interna del granulato $\beta_{out,intra}$, la densità di *bulk* all'uscita del compattatore $\rho_{out,bulk}$ e la densità della particella di solido in uscita $\rho_{out,particle}$. Tra queste si è scelta quella che presentava il valore minore dell'RMSECV a 4 LV, ovvero la variabile $\beta_{out,intra}$, dato che la stretta correlazione tra le grandezze implica che abbiano la stessa varianza spiegata (i risultati sono riassunti in Tabella 3.5).

Tabella 3.5: *Varianza spiegata e* RMSECV *di un modello PLS a 4* LV *per le variabili di risposta raccolte.*

Variabile	Varianza spiegata a 4 LV	RMSECV
$\beta_{out,intra}$	94.5%	25.474
$\rho_{out,bulk}$	94.5%	40.362
$\rho_{out,particle}$	94.5%	40.773

In Figura 3.6 si riporta il diagramma degli *score* del *set* di calibrazione e di quello di convalida. I materiali non sono facilmente distinguibili se si considerano anche le variabili dipendenti dal macchinario utilizzato: si possono tuttavia ancora distinguere i materiali 1, 4 e 5. È interessante notare come sono suddivise le prove per i diversi macchinari: le misure ricavate dal macchinario 1 sono tutte negative rispetto la LV2, mentre quelle del secondo sono tutte positive rispetto la stessa variabile latente.

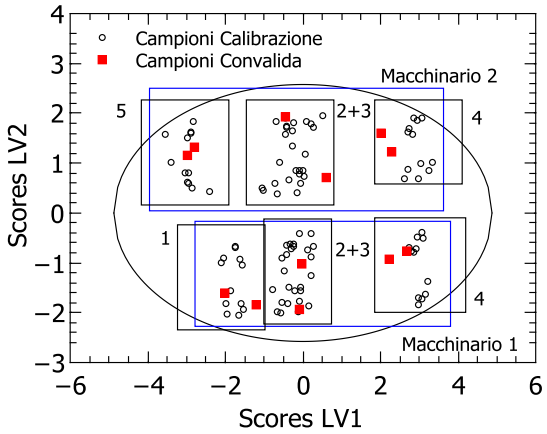


Figura 3.6: *Diagramma degli score del modello PLS del set di calibrazione e convalida per il primo scenario, con suddivisione dei diversi materiali e delle prove di ciascun macchinario.*

Il diagramma di Figura 3.7 è relativo ai *loading* del *set* di calibrazione. Guardando solo le variabili che dipendono dal materiale, è possibile vedere come vi sia una correlazione positiva tra le variabili 6 e 7 (angolo di attrito δ e angolo di attrito effettivo δ_{EFF} rispettivamente) e come la variabile 1 (costante di comprimibilità κ) e la 8 (fattore di *springback* F_{sb}) siano anticorrelate. Un'ulteriore correlazione è tra le variabili 2 (diametro del cilindro) e 3 (lunghezza del cilindro): esse sono correlate positivamente in quanto il macchinario con i diametri più grandi presenta anche la lunghezza di cilindro maggiore. Questo conferma come il modello sia in grado di descrivere in maniera corretta la struttura di correlazione delle variabili.

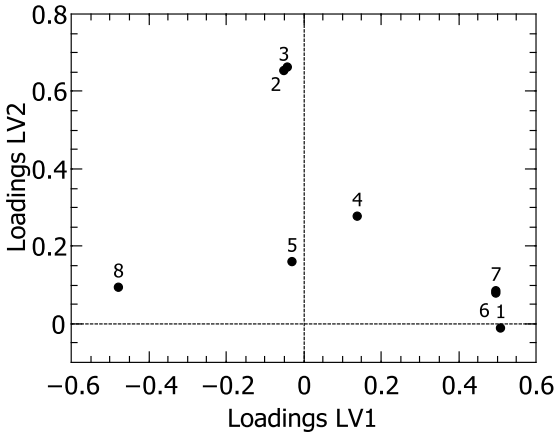


Figura 3.8: *Diagramma dei loading del modello PLS per le variabili latenti 1 e 2 per il set di calibrazione del primo scenario.*

3.3.2 Modelli dello scenario 2

In questo secondo scenario i dati di ingresso **U** sono costituiti da 26 misurazioni per il primo macchinario e 24 per il secondo. Il *set* di calibrazione conta 44 campioni e quello di convalida 6 (3 per macchinario). Anche in questo caso, un'analisi PCA sui dati **U** ha permesso di identificare il numero adatto di variabili latenti con cui costruire il modello PLS (risultati in Tabella 3.6).

Tabella 3.6: *PCA per i dati* **U** *del set di calibrazione, autovalori e varianza spiegata.*

lv	**λ**	**Varianza spiegata (%)**	**Varianza cumulativa (%)**
1	3.04	37.98	37.98
2	1.7	21.2	59.18
3	1.21	15.1	74.28
4	0.862	10.78	85.06
5	0.851	10.64	95.7
6	0.246	3.08	98.77
7	0.0973	1.22	99.99
8	0.000706	0.01	100

Nel secondo scenario il modello PLS viene costruito con 5 LV, che spiegano oltre il 95% della varianza dei dati.

Il modello PLS mostrato è relativo alla variabile porosità interna del granulato $\beta_{out,intra}$, selezionata in quanto presenta il valore minore dell'RMSECV a 5 LV (Tabella 3.7).

Tabella 3.7: *Varianza spiegata e* RMSECV *di un modello PLS a 5* LV *per le variabili di risposta raccolte.*

Variabile	Varianza spiegata a 5 LV	RMSECV
$\beta_{out,intra}$	91.9%	95.47
$\rho_{out,bulk}$	91.9%	151.35
$\rho_{out,particle}$	91.9%	152.88

Vista la maggiore variabilità dei dati, non è più possibile distinguere i diversi materiali nello spazio degli *score* (parte della distinzione è persa con l'inserimento delle variabili dipendenti dal macchinario e parte per come i dati sono stati costruiti). Tuttavia, come nel caso precedente, tutte le osservazioni del macchinario 2 si pongono nella parte superiore del diagramma degli *score* (positive sulla LV2) e quelle del macchinario 1 nella zona inferiore (negative sulla LV2; Figura 3.9).

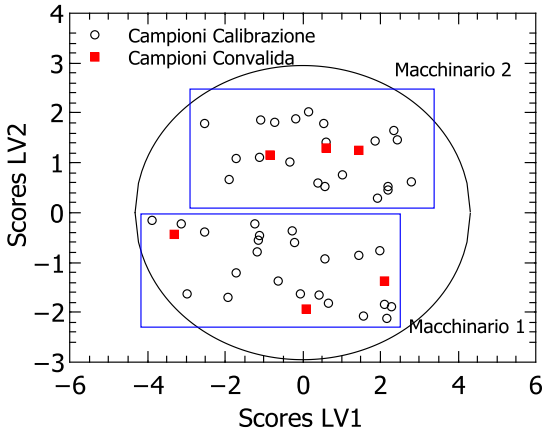


Figura 3.9: *Diagramma degli score del modello PLS del set di calibrazione e convalida per il secondo scenario, con suddivisione delle prove di ciascun macchinario.*

Il diagramma dei *loading* non si discosta da quello visto per lo scenario precedente e conduce alle stesse conclusioni. Per questo motivo non viene riportato.

3.3.3 Modelli dello scenario 3

Il terzo scenario dispone di 22 campioni per il primo macchinario e 23 per il secondo. Il *set* è stato diviso in 39 campioni di calibrazione e 6 di convalida (selezionati casualmente). Un'analisi PCA sugli ingressi **U** ha rivelato come il numero adatto di variabili latenti per la costruzione del modello sia 4, a cui corrispondono autovalori prossimi o superiori a 1 (Tabella 3.8).

Tabella 3.8: *PCA per i dati* **U** *del set di calibrazione, autovalori e varianza spiegata.*

lv	λ	Varianza spiegata (%)	Varianza cumulativa (%)
1	2.84	35.46	35.46
2	1.83	22.91	58.37
3	1.43	17.89	76.27
4	0.987	12.34	88.6
5	0.508	6.35	94.95
6	0.306	3.83	98.78
7	0.0967	1.21	99.99
8	0.000666	0.01	100

La variabile risposta, scelta tra quelle disponibili, è stata la porosità interna del granulato in uscita dal compattatore $\beta_{out,intra}$, in quanto presenta il valore minimo dell'RMSECV (Tabella 3.9).

Tabella 3.9: *Varianza spiegata e* RMSECV *di un modello PLS a 4* LV *per le variabili di risposta raccolte.*

Variabile	Varianza spiegata a 4 LV	RMSECV
$\beta_{out,intra}$	94.5%	39.526
$\rho_{out,bulk}$	94.5%	62.377
$\rho_{out,particle}$	94.5%	63.007

In questo ultimo scenario ipotizzato la variabilità dei dati è ancora maggiore: oltre a non riuscire a distinguere i diversi materiali sul diagramma degli *score*, non si individua nemmeno la distinzione delle prove tra diversi macchinari vista nei casi precedenti (Figura 3.10).

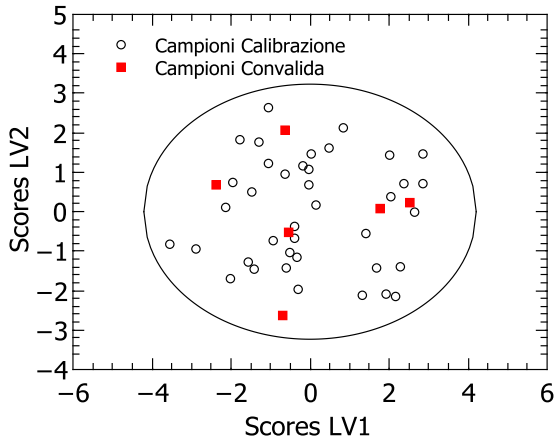


Figura 3.10: *Diagramma degli score del modello PLS del set di calibrazione e convalida per il terzo scenario.*

Anche in questo scenario il diagramma dei *loading* non è dissimile a quelli degli scenari precedenti e per brevità non è riportato.

3.4 Conclusioni sui modelli predittivi della qualità

I modelli di regressione costruiti per i diversi casi di studio riescono a descrivere in maniera corretta la struttura di correlazione attesa tra le variabili.
Il numero di variabili latenti necessario alla costruzione dei modelli, nonostante sia molto inferiore al numero di variabili originarie, spiega in tutti i casi una porzione rilevante della variabilità dei dati (non si ha quindi una perdita significativa di informazioni).
La capacità predittiva dei modelli è molto buona. Nel caso però vi siano pochi campioni di calibrazione (ad esempio, caso di studio 2) la predittività dei modelli è ridotta e può risultare compromessa.

Capitolo 4

Caratterizzazione dell'incertezza nella progettazione di prodotto

Questo Capitolo descrive i principali risultati ottenuti per la caratterizzazione dell'incertezza nei diversi casi di studio descritti nel Capitolo 2. In particolare, l'applicazione al caso di studio 1 ha permesso di definire gli approcci migliori per la stima dell'incertezza fra quelli proposti da Zhang e Garcìa-Muñoz (2009). Successivamente, si è caratterizzata l'incertezza e si sono studiate le correlazioni tra l'incertezza, l'accuratezza e la rappresentatività dei modelli.

4.1 Risultati per il caso di studio 1

Lo studio di Zhang e Garcìa-Muñoz (2009) propone diversi metodi per la stima dell'incertezza, che vengono verificati in questa prima parte di lavoro di Tesi.

4.1.1 Applicazione dei metodi di calcolo dell'incertezza

L'incertezza di predizione di un modello PLS viene calcolata in due passaggi:

1. calcolo della deviazione standard *s* dell'errore di predizione (§1.2.1);
2. stima dei gradi di libertà del modello (§1.2.2).

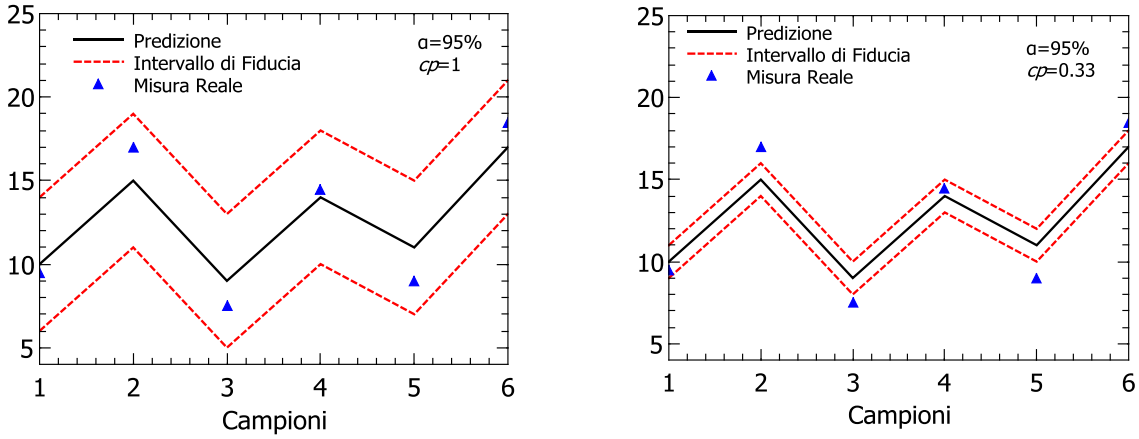
In questo Paragrafo si conduce un'analisi per identificare gli approcci più promettenti fra i diversi metodi per la stima di *s* (SF96 e UD) e per il calcolo dei gradi di libertà (Naïve, PDF e GDF), descritti nel Capitolo 1.

La bontà dei metodi è caratterizzata in termini di probabilità di copertura (*coverage probability*, *cp*), ovvero la probabilità che la risposta misurata **y** cada all'interno dell'intervallo di fiducia della predizione descritto dalla (1.17):

$$cp = \frac{n°\ di\ campioni\ di\ \mathbf{y}\ che\ cadono\ nel\ range\ definito\ dalla\ (1.17)}{n°\ di\ campioni\ di\ \mathbf{y}\ totali} \quad . \qquad (4.1)$$

Il valore di *cp* dovrà essere il più vicino possibile alla significatività scelta per la definizione della *t* di *Student* nella (1.17), ad esempio 0.95 per un intervallo di fiducia al 95%. Un valore di *cp* maggiore del limite di significatività scelta (Figura 4.1a) è indice di un intervallo di fiducia eccessivamente ampio, quindi non in grado di descrivere la variabilità dei dati, in

quanto anche misure reali lontane dalle predizioni vengono incluse all'interno dell'intervallo di fiducia. In Figura 4.1a, infatti, il 100% delle misure cade all'interno dell'intervallo di fiducia costruito con una significatività del 95%. Tuttavia, anche un valore di *cp* minore della significatività (Figura 4.1b) indica un intervallo di fiducia non rappresentativo dei campioni. In questo caso infatti il numero di campioni situati all'interno dell'intervallo è inferiore al 95% (con una significatività del 95%), ottenendo così una sottostima della *cp*.
In entrambi i casi il modello scelto per la descrizione della probabilità di copertura non è adeguato



(*a*) (*b*)

Figura 4.1: *Esempi esplicativi di: a) sovrastima della probabilità di copertura (limiti di fiducia troppo ampi); significatività α del 95% e stima di cp pari a 1; b) sottostima della probabilità di copertura (limiti di fiducia eccessivamente stretti); significatività α del 95% e stima di cp pari a 0.33.*

I risultati ottenuti nella valutazione dei metodi per la stima dell'incertezza vengono caratterizzati in termini di *cp*, numero di variabili latenti con cui è costruito il modello (*lv*), errori medi relativi di predizione (MRE) in calibrazione e convalida e gradi di libertà calcolati (*df*).

L'algoritmo UD per il calcolo della varianza di una predizione fornisce prestazioni peggiori (in tutti i casi analizzati viene sottostimato il valore di *cp*) rispetto al metodo SF96, indipendentemente da come vengono stimati i gradi di libertà (Tabella 4.1 e 4.2).
Pertanto, nel seguito si farà sempre riferimento a risultati ottenuti col metodo SF96.

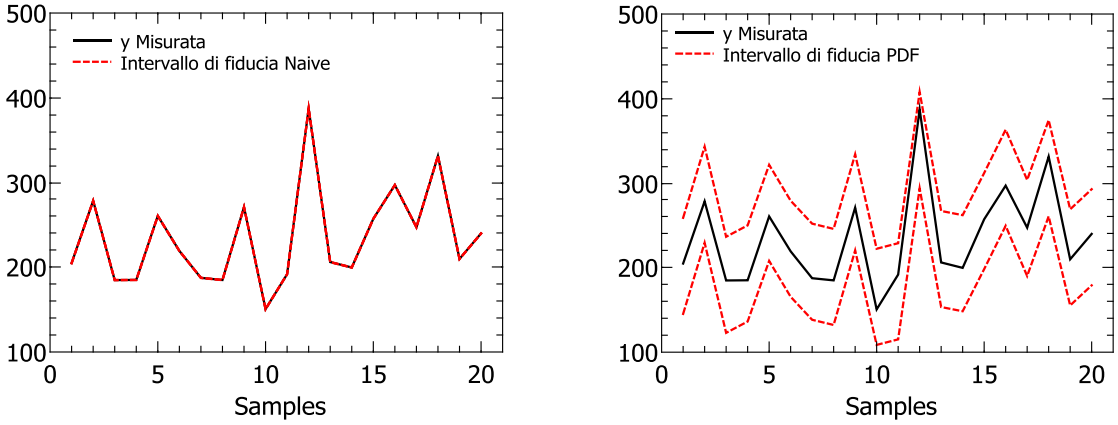
Tabella 4.1: *Confronto tra metodi Naïve e PDF: risultati ottenuti in termini di probabilità di copertura, numero di variabili latenti, errori medi relativi in fase di calibrazione e convalida, numero di gradi di libertà utilizzati. 10 dati in calibrazione, significatività al 95%.*

	Rumore							
	Assente	**10%**	**40%**	**70%**	**Assente**	**10%**	**40%**	**70%**
	Naïve				**PDF**			
cp_{SF96}	0.867	0.948	0.945	0.839	1	1	0.995	0.93
cp_{UD}	0.8	0.891	0.819	0.675	1	1	0.94	0.783
lv	2	2	2	2	2	2	2	2
MRE_{cal}	2.22	3.36	9.67	11.27	2.22	3.36	9.67	11.27
MRE_{val}	4.6	5.76	10.36	43.61	4.62	5.76	10.36	43.61
df	2	2	2	2	8.71	7.96	5.07	4.15

In Tabella 4.1 con cp_{SF96} e cp_{UD} si sono definite le probabilità di copertura ottenute dal metodo SF96 e UD rispettivamente, *df* è il numero di gradi di libertà, MRE_{cal} indica l'errore medio relativo percentuale in fase di calibrazione e MRE_{val} quello in convalida, definiti dall'equazione:

$$\mathrm{MRE} = \frac{\sum_{i=1}^{N} \frac{\left|y_i - \hat{y}_i\right|}{y_i}}{N} \cdot 100 \quad . \tag{4.2}$$

La Tabella 4.1 indica anche come l'approccio PDF peggiori le stime per l'incertezza su una predizione (sovrastima infatti i valori di *cp*). I risultati migliori si sono ottenuti per il metodo Naïve associato al SF96 (e in particolare per il 10% e 40% di rumore) rispetto le sovrastime e sottostime del metodo UD. È possibile notare anche come il metodo PDF stimi per eccesso le probabilità di copertura per entrambi i metodi (ad eccezione del 70% di rumore). L'aumento delle probabilità di copertura è dovuto all'allargamento degli intervalli d'incertezza, probabilmente a causa di una sovrastima dei gradi di libertà, che si distaccano molto dal numero di variabili latenti utilizzate, in quanto nel loro calcolo rientra il numero di dati utilizzati in fase di calibrazione (Equazione 1.30). La Figura 4.2 riporta gli intervalli di fiducia ottenuti per i metodi Naïve e PDF e indica chiaramente come i secondi siano eccessivamente ampi e quindi non rappresentativi della variabilità dei dati.



(*a*) (*b*)

Figura 4.2: *Intervalli di incertezza al 95% dal metodo SF96 a)approccio Naïve per il calcolo dei gradi di libertà b)approccio PDF per il calcolo dei gradi di libertà.*

Tabella 4.2: *Confronto tra metodi Naïve e GDF: risultati ottenuti in termini di probabilità di copertura, numero di variabili latenti, errori medi relativi in fase di calibrazione e convalida, numero di gradi di libertà utilizzati. 10 dati in calibrazione, significatività al 95%.*

	Rumore							
	Assente	**10%**	**40%**	**70%**	**Assente**	**10%**	**40%**	**70%**
	Naïve				**GDF**			
cp_{SF96}	0.867	0.948	0.945	0.839	0.877	0.958	0.955	0.946
cp_{UD}	0.8	0.891	0.819	0.675	0.833	0.935	0.834	0.797
lv	2	2	2	2	2	2	2	2
MRE_{cal}	2.22	3.36	9.67	11.27	2.22	3.36	9.67	11.27
MRE_{val}	4.62	5.76	10.36	43.61	4.62	5.76	10.36	43.61
df	2	2	2	2	2.66	3.31	2.52	4.53

L'algoritmo GDF (Tabella 4.2) migliora le prestazioni per il calcolo dell'incertezza in un modello PLS. Il valore di *cp* migliora per tutte le percentuali di rumore analizzate (si apprezzi il netto miglioramento per il 70% di rumore, nonostante gli MRE maggiori). È interessante notare come la stima dei gradi di libertà GDF si ponga sempre tra quella effettuata dal metodo Naïve e PDF, indice del fatto che probabilmente l'approccio Naïve sottostima leggermente i gradi di libertà, mentre il metodo PDF li sovrastima in modo marcato.

4.1.2 Applicazione dell'inversione diretta

La progettazione di prodotto mira a stimare quelle proprietà delle materie prime o parametri di processo (variabili di ingresso) che permettono di ottenere la qualità di prodotto desiderata (variabile risposta). In questa Tesi si studia come l'incertezza nella predizione dei modelli di regressione si propaghi dalla qualità di prodotto desiderata alle caratteristiche delle materie prime e ai parametri di processo determinati per inversione del modello.

Nei modelli di regressione multivariata a variabili latenti, diversi tipi di incertezza condizionano il risultato di predizione: incertezza nei parametri del modello di calibrazione (Martens e Martens, 2000), incertezza nei dati di calibrazione (Reis e Saraiva, 2005), incertezza di predizione (Fernandez Pierna *et al.*, 2003). In Figura 4.3 si è riportato un esempio esplicativo della rappresentazione dei diversi tipi di incertezza sul piano degli *score* delle prime due variabili latenti. Ad esempio, l'incertezza sui dati di calibrazione è un'incertezza correlata alla misura dei dati sperimentali degli ingressi e delle uscite (ellissi verde in Figura 4.3). Inoltre, vi sono anche un'incertezza legata ai parametri del modello che si ripercuote sulla collocazione dell'inversione diretta del modello (ellissi azzurra in Figura 4.3) e un'incertezza legata al calcolo dello spazio nullo (Tomba *et al.*, 2012; curve arancioni in Figura 4.3). Un'analisi completa dovrebbe considerare contemporaneamente tutti questi tipi di incertezza e l'incertezza globale, derivante dalle diverse fonti, andrebbe dunque determinata dalla probabilità congiunta di tutte le sorgenti di incertezza. L'interesse di questa Tesi è valutare unicamente l'incertezza, derivante dal modello, nella predizione della qualità di prodotto **y**, e osservare come essa si ripercuota sulla determinazione delle variabili di ingresso **U** (materie prime, parametri di processo e condizioni iniziali) nell'inversione del modello.

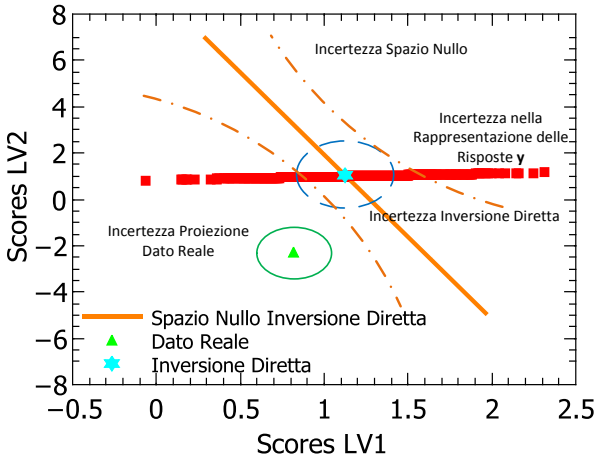


Figura 4.3: *Rappresentazione dei diversi tipi di incertezza.*

Seguendo la procedura descritta nel §1.4 e riassunta in Figura 4.4 per l'inversione diretta, è possibile ottenere un *set* di variabili di ingresso **U** (corrispondenti ad esempio a condizioni iniziali in termini di caratteristiche di materie prime o parametri di processo) in grado di fornire la risposta desiderata (y_{DES}, associabile ad esempio ad una qualità di prodotto), con una determinata incertezza calcolata con i metodi descritti nel §1.2.

L'inversione diretta del modello PLS a partire da una y_{DES} selezionata dal *set* di convalida, secondo le (1.34) e (1.35), restituisce le corrispondenti $\mathbf{u}_{NEW}$ (✶ in Figura 4.4a) nello spazio delle variabili latenti (*step* 1 e 2). Nel caso il modello sia rappresentativo, l'inversione diretta

dovrebbe essere sufficientemente "vicina" alle reali condizioni di processo che hanno permesso di ottenere y_{DES} (▲ in Figura 4.4a).

La proiezione di $\mathbf{u}_{NEW}$ sullo spazio delle **y** permette di ottenere $\hat{y}_{DES}$ (*step* 3), a cui si applica il concetto di incertezza. Nota l'incertezza nella predizione $\hat{y}_{DES}$ (dal metodo SF96, Equazione 1.26), si è costruita una distribuzione normale attorno al valore di $\hat{y}_{DES}$, che rappresenta una serie di realizzazioni di *y* a preassegnati livelli di incertezza (*step* 4). L'inversione diretta per tutte queste realizzazioni di $\hat{y}_{DES}$ preassegnate produce la corrispondente distribuzione per le variabili in ingresso $\mathbf{u}_{NEW}$ (segmento rosso in Figura 4.4c, *step* 5).

Nell'ultimo passaggio della procedura si calcola lo spazio nullo (*step* 6, Equazioni 1.39 e 1.40) associato a ciascun punto della distribuzione di $\mathbf{u}_{NEW}$ che è stato restituito dall'inversione delle realizzazioni (banda di linee blu sul piano degli *score*; Figura 4.4d). Ciascuno di questi spazi nulli rappresenta il luogo dei punti dello spazio delle **U** che garantiscono la stessa qualità di prodotto ad un preassegnato livello di incertezza rispetto all'inversione diretta[4], e quindi è uno spazio ad isoincertezza. I valori di *y* restituiti dagli spazi nulli, confrontati con il valore reale della y_{DES}, sono utili alla caratterizzazione dell'incertezza (intesa come percentile della realizzazione associata allo spazio nullo che contiene y_{DES}).

[4] Lo spazio nullo dell'inversione diretta è il luogo dei punti che garantiscono una stessa qualità con la minima incertezza.

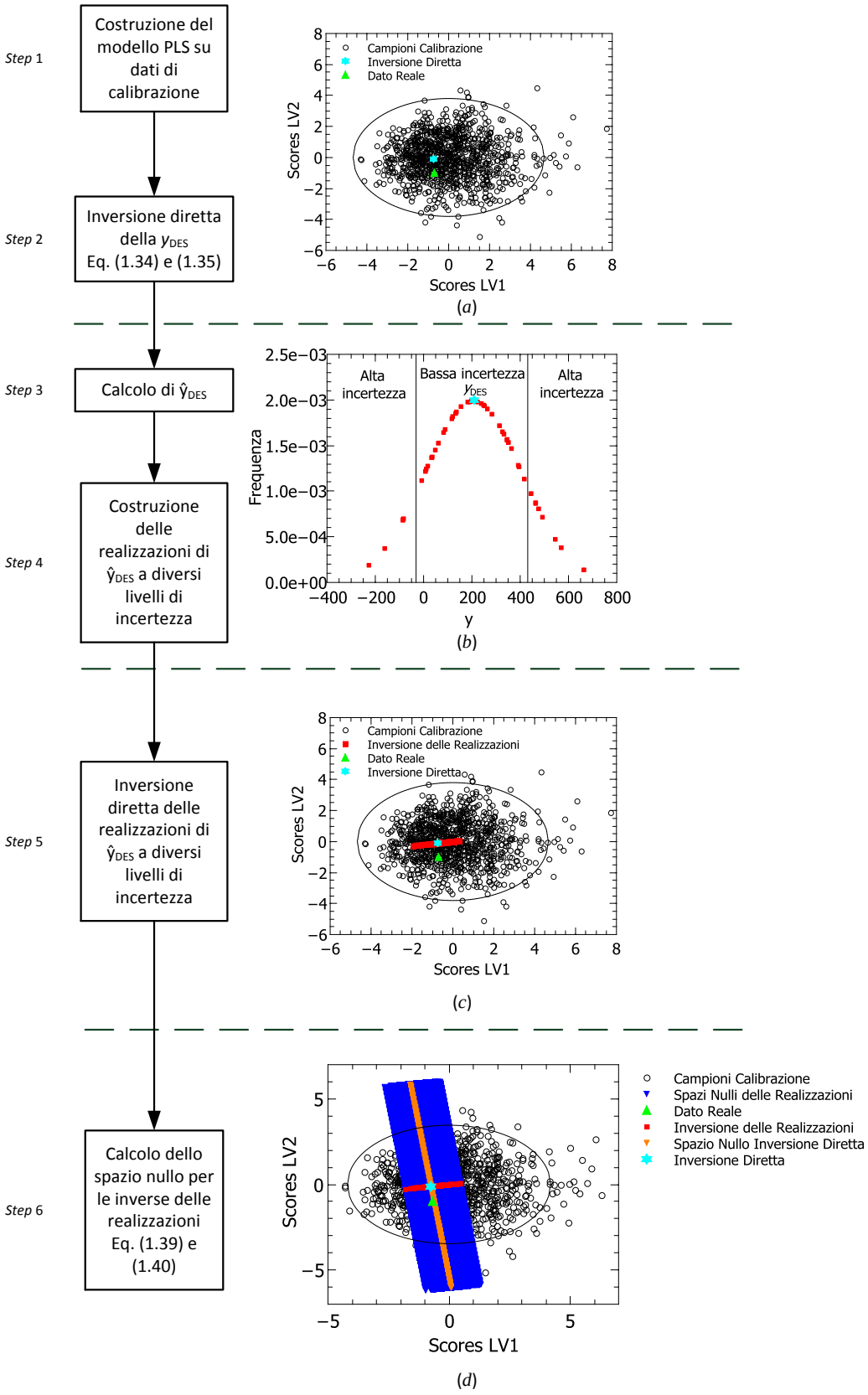


Figura 4.4: *Utilizzo dei metodi sull'incertezza con l'inversione diretta: a) proiezione sul piano degli score delle* $\mathbf{u}_{NEW}$ *ottenute dall'inversione e delle* **u** *reali; b) realizzazioni di y a differenti livelli di incertezza; c) proiezione sul piano degli score dell'incertezza nell'inversione diretta e delle* **u** *reali; d) proiezione sul piano degli score degli spazi nulli, delle realizzazioni a diversi livelli di incertezza e delle* **u** *reali.*

4.1.3 Caratterizzazione del risultato

Per descrivere l'adeguatezza del modello utilizzato per la progettazione di prodotto, si caratterizzano e confrontano l'incertezza della soluzione dell'inversione, la rappresentatività e l'accuratezza del modello.

Per caratterizzare il risultato si è scelto di utilizzare alcune metriche classiche (T^2 di Hotelling e Q, definite nel Capitolo 1) assieme altre due metriche:

- la distanza euclidea nello spazio del modello tra il punto di inversione diretta e il punto sperimentale reale; tale distanza è una misura dell'"errore" della stima dei predittori dall'inversione diretta rispetto al valore reale dei parametri di processo e delle condizioni iniziali reali utilizzati per ottenere una qualità desiderata di prodotto;
- l'incertezza con cui l'inversione diretta si discosta dal valore reale.

In dettaglio, il risultato viene caratterizzato in termini di (Figura 4.5):

- distanza euclidea tra punto reale e punto di inversione diretta (d_E);
- T^2 di Hotelling del punto di inversione diretta (T^2_{invdir});
- T^2 di Hotelling del punto reale (T^2_{real});
- Q del punto reale (Q_{real});
- incertezza della proiezione nello spazio latente del punto reale rispetto alla distribuzione delle proiezioni delle realizzazioni a diversi livelli d'incertezza (α_{inc}).

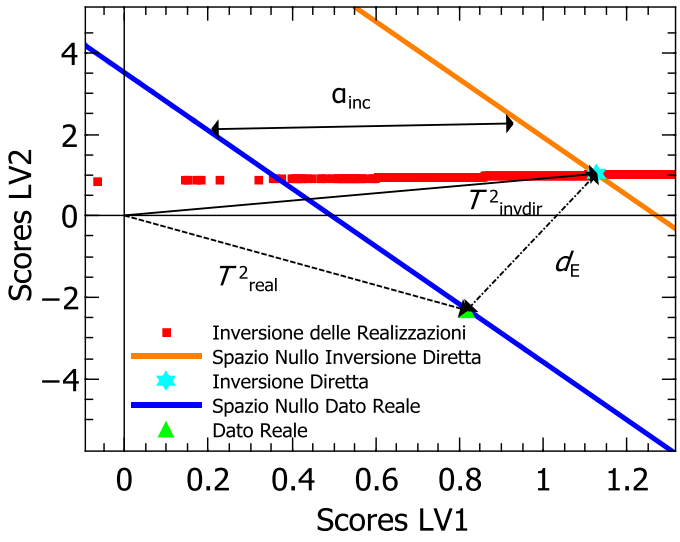


Figura 4.5: *Rappresentazione delle metriche utilizzate per caratterizzare il risultato.*

La distanza euclidea d_E fornisce una misura della bontà dell'inversione diretta: più piccola è questa grandezza migliore è il risultato, cioè la stima delle variabili di ingresso **U** dall'inversione diretta è molto vicina alla realtà fisica. Questa distanza d_E viene inoltre confrontata con i semiassi dell'ellissoide di fiducia al 95%: una distanza vicina al valore dei semiassi (o un rapporto tra d_E e i semiassi vicino a 1), unita al confronto con i valori assunti

da T^2 e Q (o dei loro rapporti con i rispettivi limiti), è indice di una bassa rappresentatività del modello, e la soluzione ottenuta potrebbe essere meno affidabile se estrapolata.
T^2 è una statistica correlata alla distanza dai valori medi (l'origine del sistema costituito dalle variabili latenti); Q fornisce un'indicazione della rappresentatività del modello e quindi dell'accuratezza con cui vengono rappresentati i dati (Wise e Gallagher, 1996).
L'inversione diretta permette di ottenere le variabili di ingresso in grado di fornire la risposta desiderata. Tuttavia nel fatto che le variabili così calcolate restituiscano *effettivamente* l'uscita voluta è insita un'incertezza dovuta al modello. Questa incertezza è legata alla deviazione standard dell'errore di predizione (s) e al numero di gradi di libertà (df). Si sono costruite realizzazioni della medesima y_{DES} a diversi livelli di incertezza, e ad ognuna di esse è stata applicata l'inversione diretta, in modo da ottenere tutte le corrispondenti realizzazioni delle variabili di ingresso **U** ai medesimi livelli di incertezza[5]. Per ogni realizzazione, si è definito lo spazio nullo corrispondente (il luogo dei punti dello spazio delle **U** che garantisce una qualità del prodotto con il predeterminato grado di incertezza). Così costruiti, gli spazi nulli delle inversioni dirette delle diverse realizzazioni costituiscono i luoghi di isoincertezza.
L'ultima metrica (α_{inc}) caratterizza questa incertezza ed è la distanza, parallela al luogo delle inversioni dirette delle realizzazioni a diversi livelli di incertezza, tra lo spazio nullo passante per il punto di inversione diretta e quello passante per il punto sperimentale reale (che non restituisce la qualità y_{DES}, ma una diversa realizzazione, y_{inc}).
La metrica viene calcolata come percentile delle realizzazioni di y_{DES} a diversi livelli di incertezza mediante le seguenti equazioni:

$$n_\sigma = \frac{|y_{inc} - y_{DES}|}{s} \tag{4.3}$$

$$\alpha_{\text{inc}} = erf\left(\frac{n_\sigma}{\sqrt{2}}\right) \quad , \tag{4.4}$$

dove n_σ è il numero di varianze, *erf* rappresenta la funzione errore e α_{inc} il livello di incertezza. In sostanza α_{inc} definisce il piano di isoincertezza dell'inversione diretta in cui il punto reale cade[6]. Più piccola è la significatività, più vicini ci si trova al luogo di isoincertezza definito dallo spazio nullo relativo al punto reale (cioè alla realtà fisica), quindi migliore è il risultato.

[5] Ciò che viene invertito è il modello; questo modello inverso sfrutta poi come ingresso la variabile risposta desiderata y_{DES}. Qui e nel proseguo della Tesi, per brevità e leggerezza di forma, si farà riferimento a questo come 'inversione di una y' o 'applicazione dell'inversione ad una y'.
[6] La scelta di applicare l'incertezza al punto di inversione diretta è dettata dal fatto che quel punto è l'unico ricavato dal modello e su cui sembra più appropriato caratterizzare un'incertezza derivante dal modello. Si potrebbe anche valutare la possibilità di applicare l'incertezza al punto reale. Però sul punto reale si dovrebbe tener conto dell'incertezza derivante dalla proiezione sul modello e contemporaneamente di altri tipi di incertezza (ad esempio, l'incertezza sulla misura).

Ciascuna metrica viene confrontata con i rispettivi valori limite: la distanza euclidea d_E viene rapportata ai semiassi dell'ellisse di fiducia in quanto è calcolata sul piano degli *score*, T^2 e Q hanno i propri limiti. Per α_{inc} non esiste un limite vero e proprio; il livello accettabile di incertezza dipende dalle esigenze del progettista/formulatore, che dovrà fissare il livello adeguato che può essere accettato e interpretare le necessità caso per caso.

4.1.4 Risultati

Questo Paragrafo presenta i risultati, sia grafici che numerici, per alcuni campioni del *set* di convalida del caso di studio 1. Sono stati scelti questi dati in quanto ritenuti rappresentativi di un ampio spettro di situazioni realistiche.

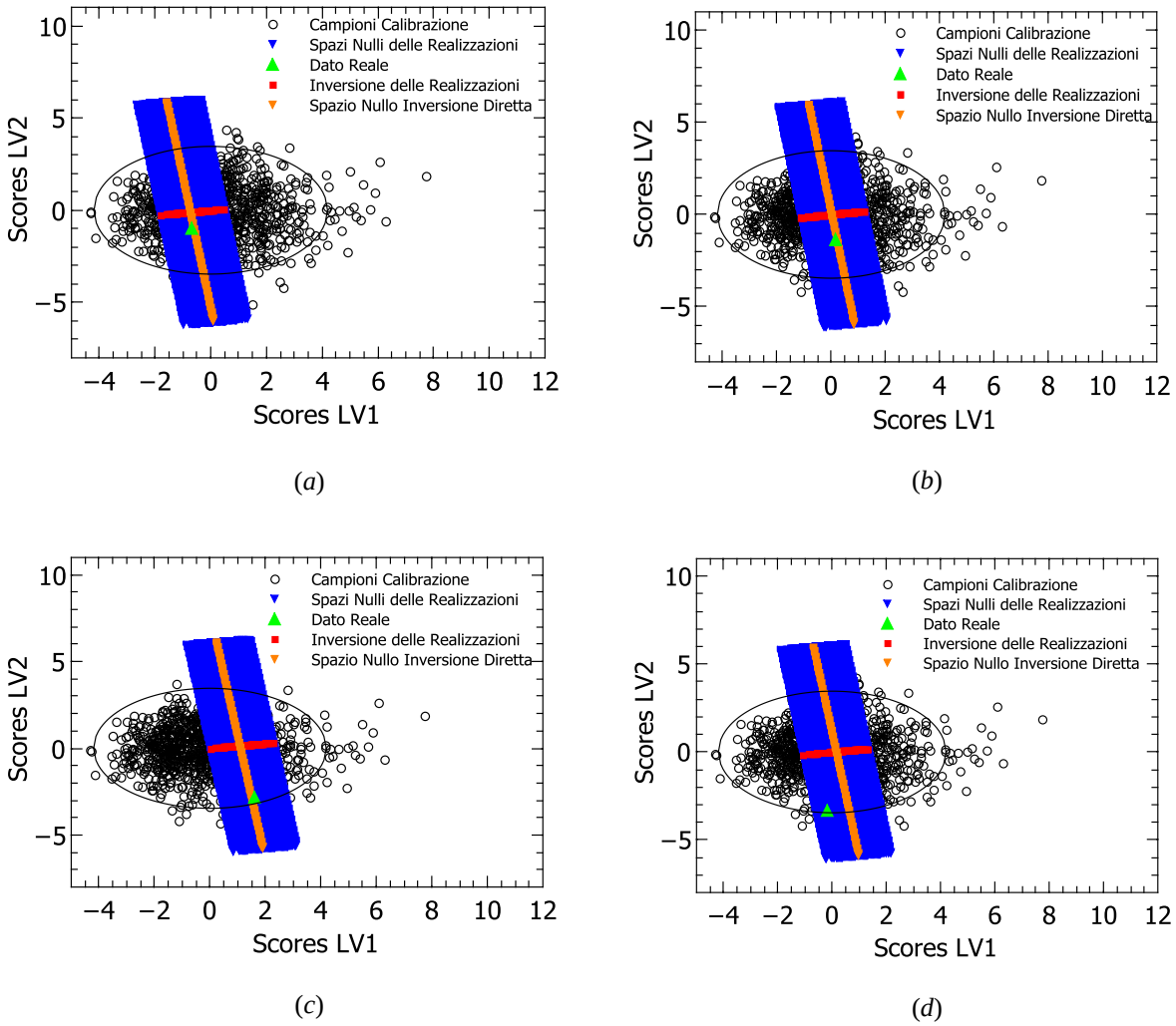


Figura 4.6: *Proiezione sul piano degli score degli spazi nulli, delle realizzazioni di* $\mathbf{u}_{NEW}$ *a diversi livelli di incertezza e delle* **u** *reali: a) campione di convalida 1, b) campione 2, c) campione 3, d) campione 4.*

Tabella 4.3: *Caratterizzazione dell'incertezza, dell'accuratezza e della rappresentatività nell'inversione diretta del modello PLS nel caso di studio 1: distanza euclidea, T^2 di Hotelling del punto di inversione diretta, T^2 di Hotelling del punto reale, Q del punto reale (rapportate al proprio limite) e incertezza.*

Campione di convalida	d_E/l_1	d_E/l_2	$T^2_{invdir}/T^2_{lim,95\%}$	$T^2_{real}/T^2_{lim,95\%}$	$Q_{real}/Q_{lim,95\%}$	α_{inc}
1	0.2105	0.2551	0.0330	0.1087	0.2485	0.1081
2	0.3247	0.3935	$1.27\cdot10^{-7}$	0.1542	0.0391	0.0616
3	0.7253	0.8789	0.0653	0.8194	0.8520	0.1520
4	0.7949	0.9634	0.0010	0.9120	1.9701	0.9712

Al fine di spiegare l'utilità e i dettagli dell'applicazione, si presenta un caso specifico di progettazione di prodotto mediante inversione diretta e della relativa caratterizzazione in termini di incertezza, accuratezza e rappresentatività. Prendiamo come esempio il primo campione di convalida (prima riga di Tabella 4.3). L'obiettivo della progettazione di prodotto/processo è suggerire quale sia il *set* di variabili di ingresso (ad esempio, materie prime, parametri di processo) che garantisca una qualità del prodotto y_{DES}=204.86. Per conoscere le variabili di ingresso necessarie a restituire y_{DES}, il modello costruito sul *set* di calibrazione viene invertito, ottenendo come soluzione delle (1.34) e (1.35) le variabili di ingresso $\mathbf{u}_{NEW}$=[38.04;9.99;$1.68\cdot10^3$;106.91;373.13]. Tuttavia, l'incertezza di predizione dovuta al modello influenza il risultato. Costruite le realizzazioni di y_{DES} a diversi livelli di incertezza (distribuite secondo una gaussiana), e applicata l'inversione diretta ad ognuna delle suddette realizzazioni, si ricavano tutte le corrispondenti realizzazioni delle variabili di ingresso **U**. Inoltre, per ogni realizzazione, si definisce lo spazio nullo corrispondente (il luogo dei punti isoincertezza a cui sono associati predeterminati livelli di incertezza). Ipotizzando di aver condotto una campagna sperimentale, i valori delle variabili di ingresso al processo che sono realmente in grado di restituire y_{DES} sono noti e sono $\mathbf{u}_{real}$=[47.13;8.93;$2.22\cdot10^3$;79.76;420.89].

A questo punto vengono utilizzate le metriche descritte nel §4.1.3 per caratterizzare l'incertezza, l'accuratezza e la rappresentatività del modello costruito. La valutazione della d_E permette di asserire quanto accurata è la soluzione stimata dal modello. In questo caso (prima riga di Tabella 4.3) si è ottenuto una valore del rapporto tra distanza euclidea e semiassi dell'ellissoide di fiducia pari a 0.2105: questo significa che la distanza euclidea è ben inferiore del semiasse dell'ellissoide e che la soluzione calcolata dall'inversione del modello è abbastanza vicina alla realtà sperimentale. Le T^2 di Hotelling sono molto basse (il rapporto è 0.033 per il punto ricavato dall'inversione diretta del modello e 0.1087 per il punto sperimentale reale): la soluzione dell'inversione diretta e il punto reale sono situate vicino ai valori medi delle variabili prese in esame, dove si ottengono risultati affidabili in quanto non si è in fase di estrapolazione. Il rapporto tra l'errore quadratico medio e il limite al 95% è

molto inferiore a 1 e pari a 0.2485. Questo dimostra come il punto sperimentale reale sia ben rappresentato dal modello. L'incertezza (α_{inc}) definisce il piano di isoincertezza dell'inversione diretta in cui il punto sperimentale cade. Quindi, più piccola è la significatività, più vicini si è al luogo di isoincertezza definito allo spazio nullo relativo al punto reale (cioè si è più prossimi alla realtà fisica). In questo caso l'incertezza è bassa (10.81%): la soluzione stimata dall'inversione del modello si trova in vicinanza dello spazio nullo a cui appartiene il punto reale. Riassumendo, nel caso descritto la soluzione ottenuta è molto accurata: l'incertezza è bassa e il modello è rappresentativo dei dati analizzati. La stima del modello giace in una zona dello spazio degli *score* ben descritta dal modello, vicina ai valori medi delle variabili prese in esame, dove si ottengono soluzioni affidabili visto che si sta interpolando. Infine, il risultato è soddisfacente visto il poco scostamento tra le variabili di ingresso ottenute dall'inversione del modello e la realtà sperimentale.

I primi due campioni presi in considerazione (osservazioni 1 e 2, Figure 4.6a e 4.6b) presentano buoni risultati per tutte le metriche: la distanza euclidea è ben al di sotto dei semiassi dell'ellisse di fiducia al 95% (l_1=4.1932 e l_2=3.4601), T^2 e Q non superano il limite al 95% ($T^2_{\mathrm{lim},95\%}$=6.0155 e $Q_{\mathrm{lim},95\%}$=0.2479 rispettivamente). L'incertezza è bassa (10.81% e 6.16%). I risultati trovano conferma visiva nelle Figure 4.6a e 4.6b: la distanza tra il punto reale (▲) il punto di inversione diretta (intersezione tra segmento rosso e arancione) è molto ridotta e il punto reale è quasi sovrapposto allo spazio nullo dell'inversione diretta (cioè l'incertezza è bassa). Più interessanti sono le soluzioni proposte per gli altri due campioni (3 e 4), che si trovano al limite dell'ellisse di fiducia al 95% sul piano degli *score*.

L'osservazione 3 presenta una d_E confrontabile con i semiassi di fiducia; in più T^2_{real} e Q_{real} sono molto vicini al limite al 95%. Tuttavia l'incertezza α_{inc} è bassa, e ciò conferma che è necessario confrontare queste metriche nella loro totalità, in quanto potrebbero fornire indicazioni contrastanti. Si osservi che per i punti periferici è più probabile un'estrapolazione della soluzione con perdita di rappresentatività del modello.

Il campione 4 presenta i risultati peggiori: la distanza euclidea quasi coincide con il semiasse minore dell'ellisse, T^2_{real} è appena al di sotto del limite al 95%, mentre Q_{real} supera il valore di 0.2479 del limite. L'incertezza inoltre è alta, pari al 97.12%.

Per entrambi gli ultimi casi T^2_{invdir} è molto minore di T^2_{real}, risultato che, unito alle altre metriche, permette di affermare come si ottengano risultati meno buoni per osservazioni lontano dall'origine.

È interessante notare come T^2_{real} sia sempre maggiore di T^2_{invdir}: questo è dovuto alla natura intrinseca dei modelli, che tendono a ottenere soluzioni il più vicino possibile alla media.

4.1.5 Correlazioni tra le metriche

L'obiettivo di questo Paragrafo è studiare se vi siano delle correlazioni tra l'incertezza α_{inc} e le metriche di scostamento dal dato reale, distanza dai valori medi e rappresentatività del modello (d_E, T^2_{invdir}, T^2_{real}, Q_{real}).

Si cerca una procedura per "certificare" il modello e caratterizzare l'incertezza in fase di calibrazione, in modo che le procedure fin qui sviluppate diventino uno strumento utile e affidabile per le industrie (manifatturiere, farmaceutiche, ecc.) nell'ambito della progettazione e sviluppo di nuovi prodotti e processi, accettate dagli enti certificatori.

Calcolate le metriche di caratterizzazione del modello (d_E, T^2_{invdir}, T^2_{real}, Q_{real} e α_{inc}) per l'intero *set* di convalida di questo caso di studio, si è costruito un modello PCA sulla matrice [1000×5] costituita dalle 5 metriche per i 1000 campioni di convalida. L'analisi del diagramma dei *loading* permette di identificare se vi siano correlazioni tra le variabili, e quanto significative esse siano (Figura 4.7).

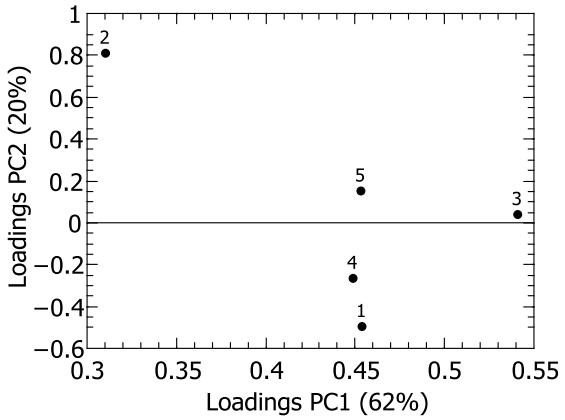


Figura 4.7: *Diagramma dei loading del modello PCA per le metriche di caratterizzazione di modello e incertezza calcolate dal set di convalida.*

Su PC1, che spiega il 62% circa della variabilità dei dati, tutte le metriche considerate sono correlate positivamente. Questa correlazione è ragionevole guardando come queste metriche di caratterizzazione dell'incertezza sono calcolate: tutte dipendono, attraverso funzioni più o meno complesse, dalle coordinate del punto reale. È importante notare la moderata correlazione lungo PC1 delle variabili 1 (d_E), 4 (Q_{real}) e 5 (α_{inc}). Questo significa che più la soluzione dell'inversione diretta si discosta dalla realtà fisica (alta d_E) o minore è la rappresentatività del modello (alto Q_{real}), più alta è l'incertezza (α_{inc}). Sempre lungo PC1, le variabili 3 (T^2_{real}) e 5 (α_{inc}) sono modestamente correlate. Ciò dimostra come più alta è la distanza del punto reale dai valori medi (l'origine degli assi sul piano degli *score*; alta T^2_{real}) maggiore è l'incertezza.

La procedura cercata per certificare il modello viene usata per calcolare dei parametri medi (in particolare l'incertezza) per il *set* di calibrazione e per studiare la sensitività dell'incertezza

in calibrazione e convalida al numero di campioni disponibili per la calibrazione con cui viene costruito il modello PLS. È stato utilizzato un approccio *jackknife* (Miller, 1964). Noti **U** e **y** come i regressori e la variabile risposta, rispettivamente, il metodo *jackknife* è una procedura iterativa che alla generica iterazione *i*, si articola nei seguenti passaggi:

1. rimozione dell'*i*-esimo campione (ad esempio, *i*-esima riga) da **U** e **y**, generando le matrici $\mathbf{U}^{(i)}$ e $\mathbf{y}^{(i)}$;
2. costruzione del modello PLS su $\mathbf{U}^{(i)}$ e $\mathbf{y}^{(i)}$, mantenendo lo stesso numero di LV utilizzato nel modello globale tra **U** e **y**;
3. calcolo dell'incertezza sull'*i*-esimo campione rimosso da **y** utilizzato in convalida;
4. iterazione della procedura per gli *N* campioni di **U**.

Con questa procedura si ottengono *N* valori di incertezza α_{inc} che, mediati sul numero di dati di calibrazione, permettono di ottenere un'incertezza "media" del modello. Successivamente, sul modello globale, si sono ricavati i valori di incertezza per l'intero *set* di campioni di convalida.

È importante chiarire che l'incertezza media ottenuta dal metodo *jackknife* si calcola a priori dai soli dati di calibrazione ed è utile a certificare il comportamento del modello. L'incertezza effettiva di convalida può essere ricavata solo a posteriori una volta che si è in possesso di un ulteriore *set* di dati (quello di convalida appunto; Figura 4.8).

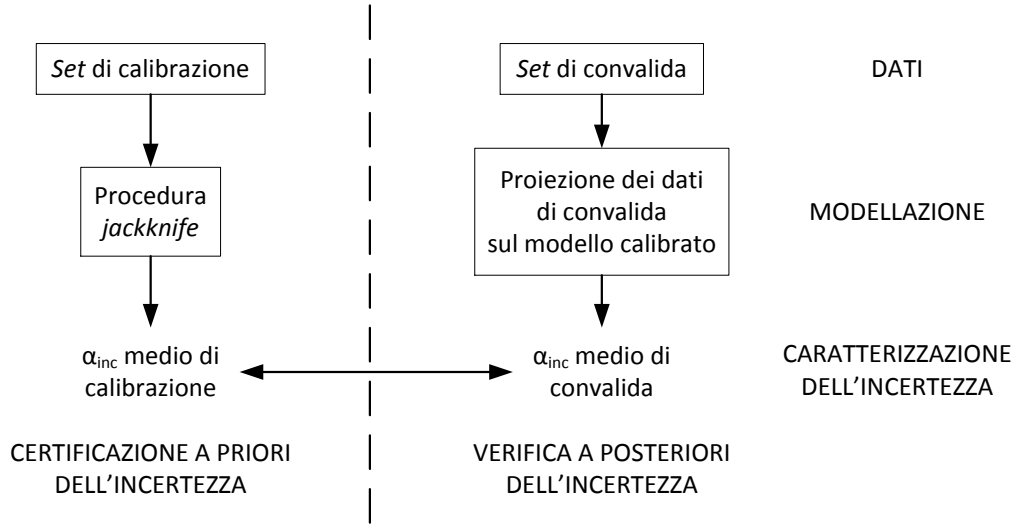


Figura 4.8: *Diagramma a blocchi delle procedure a priori e a posteriori di caratterizzazione dell'incertezza.*

Questo valore medio ottenuto dalla calibrazione è stato confrontato con il valore medio ottenuto dal *set* di convalida, studiando la sensitività dell'incertezza in calibrazione e convalida al numero di campioni disponibili per la calibrazione con cui viene costruito il modello PLS (Figura 4.9). Per chiarire la procedura *jackknife* sopra esposta, si descrive in maggior dettaglio un caso riportato in Figura 4.9. Si ipotizzi di avere 10 campioni di calibrazione; viene applicata la procedura *jackknife*:

1. rimozione del primo campione di calibrazione da **U** e **y**;
2. costruzione del modello PLS sui 9 campioni di calibrazione rimasti, mantenendo lo stesso numero di LV utilizzato sui 10 campioni originari;
3. caratterizzazione dell'incertezza sul primo campione rimosso da **U** e **y**;
4. iterazione della procedura per tutti i 9 campioni di calibrazione rimasti.

I 10 valori di incertezza calcolati vengono mediati, ottenendo un'incertezza media per la fase di calibrazione di 0.64. Successivamente si riproiettano i 1000 campioni di convalida sul modello di calibrazione costruito con 10 campioni. Le 1000 incertezze ottenute, una volta mediate, consentono di ottenere un'incertezza media in convalida di 0.86. L'incertezza in convalida è maggiore di quella di calibrazione. Ciò significa che si è in possesso di un numero eccessivamente esiguo di campioni di calibrazione per caratterizzare l'incertezza del modello. Si consiglia quindi di aumentare il numero di dati sperimentali finché l'incertezza ricavata a priori per il *set* di calibrazione e quella di convalida raggiungono un comportamento stabile ed eventualmente si avvicinano (coincidono in un caso ideale).

Nel caso non si fosse in possesso di un *set* di dati di convalida, o si volesse utilizzare solo la procedura a priori, è possibile seguire due vie di applicazione. Nella prima si dovrebbe aumentare progressivamente il numero di campioni di calibrazione (dove possibile) finché non si nota un andamento stabile dell'incertezza: quel valore dovrebbe rappresentare l'incertezza limite del modello. Nella seconda il progettista/formulatore deve fissare un valore di incertezza ritenuto accettabile (a seconda delle necessità e dell'esperienza) e verificare a quanti campioni di calibrazione tale valore corrisponde.

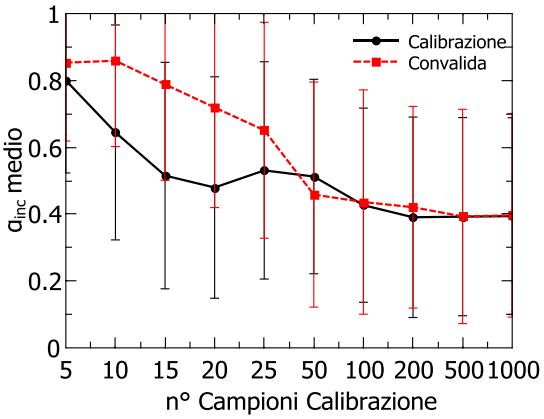


Figura 4.9: *Incertezze medie per la calibrazione e per la convalida al variare del numero di campioni di calibrazione. Le barre di errore indicano la deviazione standard dei valori di incertezza da cui si è ricavata l'incertezza media.*

Dalla Figura 4.9 si traggono due conclusioni principali:

1. all'aumentare del numero di campioni di calibrazione l'incertezza media (in genere) diminuisce;

2. l'incertezza media di calibrazione è ragionevolmente minore di quella di convalida per un numero basso di campioni di calibrazione. Tuttavia, vi è un numero di campioni di calibrazione "critico" (50/100 campioni circa in questo caso di studio), in corrispondenza del quale i valori di incertezza media in calibrazione e in convalida coincidono, e dopo il quale si assestano su un valore più o meno costante (in questo caso di studio utilizzando 100 o più campioni di calibrazione il valore di incertezza medio si fissa attorno al valore di 0.4 sia per il *set* di calibrazione che per quello di convalida). Questo significa che, quando la variabilità del *set* di calibrazione è pienamente rappresentativa della variabilità dei campioni di convalida, anche l'incertezza media che si valuta dal *set* di calibrazione è la medesima che in quello di convalida.

Successivamente, l'obiettivo è approfondire (visivamente) la relazione che intercorre tra distanza dai valori medi, rappresentatività del modello e incertezza. Per verificare quali punti presentino un'incertezza superiore alla media, è stato effettuato un controllo sul piano degli *score*, sul diagramma delle T^2 di Hotelling e sul diagramma dei residui Q.

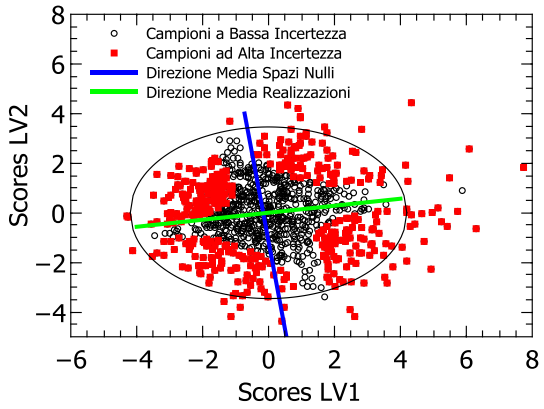


Figura 4.10: *Proiezione dei campioni di calibrazione ad elevata incertezza nel piano degli score del modello PLS (1000 dati in calibrazione).*

La Figura 4.10 indica come i campioni di calibrazione a più elevata incertezza (superiore al valore medio, ■ in Figura 4.10) siano quelli a maggiore T^2, cioè più distanti dall'origine degli assi del piano degli *score* (corrispondente ai valori medi). Si distinguono due direzioni lungo le quali non vi sono punti ad alta incertezza. Una corrisponde alla direzione media degli spazi nulli (che costituiscono infatti un luogo di isoincertezza) mentre l'altra coincide con la direzione media assunta dalle realizzazioni a diversi livelli di incertezza delle $\mathbf{u}_{NEW}$. Il punto a minor incertezza, per come è stata definita, è il punto di inversione diretta. Questo si muove lungo la direzione media delle realizzazioni, una delle direzioni dove non si riscontrano punti ad alta incertezza. La maggior parte delle soluzioni dell'inversione diretta, e gli spazi nulli ad esse associate, si concentra in vicinanza dell'origine degli assi (per la natura stessa dei modelli di regressione lineare, che tendono ad ottenere soluzioni il più possibile vicine alla

media). Questo spiega perché lungo la direzione media centrale degli spazi nulli non vi sono punti ad alta incertezza.

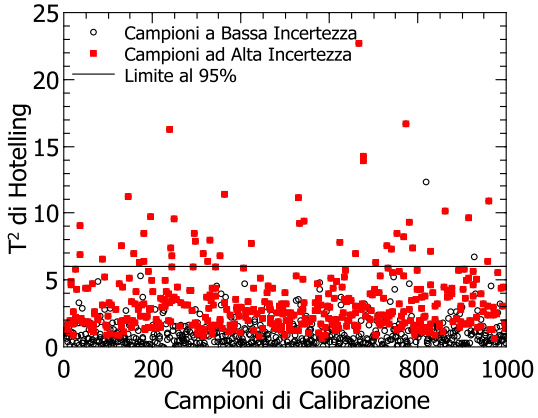


Figura 4.11: *T^2 di Hotelling per i dati di calibrazione e campioni di calibrazione ad alta incertezza. Modello PLS costruito su 1000 dati in calibrazione. La linea tratteggiata indica il limite al 95% per la T^2 di Hotelling.*

La Figura 4.11 conferma come i punti ad alta incertezza siano quelli con T^2 mediamente maggiore. Il grafico dei residui Q (che non viene riportato qui) porta alle medesime conclusioni.

L'analisi effettuata sul *set* di convalida conduce agli stessi risultati e per brevità non è qui indicata.

4.2 Risultati per il caso di studio 2

In questo Paragrafo sono applicati i metodi di inversione diretta e stima dell'incertezza nella progettazione di prodotto nel caso di studio del processo di granulazione umida (Vemavarapu *et al*., 2009), descritti nei Capitoli 2 e 3.

4.2.1 Applicazioni ai dati di granulazione umida

Innanzitutto si sono verificate le procedure per la stima della deviazione standard dell'errore di predizione s e per il calcolo dei gradi di libertà. Come nel caso precedente, gli approcci più promettenti sono SF96 per la stima di s e GDF per i gradi di libertà (i risultati per brevità non sono riportati). L'applicazione del metodo SF96, unito all'approccio GDF (df=7.65, rispetto alle 4 LV utilizzate nella costruzione del modello) per il calcolo dei gradi di libertà, ha permesso di costruire gli spazi nulli che restituiscono, teoricamente, la y_{DES} all'interno della distribuzione computata.

In questo caso di studio lo spazio nullo ha dimensione lv-V_y=4-1=3.

Tabella 4.4: *Caratterizzazione dell'incertezza, dell'accuratezza e della rappresentatività nell'inversione diretta del modello PLS nel caso di studio 2: distanza euclidea, T^2 di Hotelling del punto di inversione diretta, T^2 di Hotelling del punto reale, Q del punto reale (rapportate al proprio limite) e incertezza.*

Campione di convalida	d_E/l_1	d_E/l_2	$T^2_{invdir}/T^2_{lim,95\%}$	$T^2_{real}/T^2_{lim,95\%}$	$Q_{real}/Q_{lim,95\%}$	α_{inc}
1	0.2810	0.3780	0.1227	0.2910	1.2016	0.1816
2	0.9807	1.3192	0.0058	2.4792	0.1847	0.8395
3	0.3811	0.5126	0.0112	1.0000	0.1847	0.4153
4	0.9468	1.2736	1.5078	0.4014	0.5021	1
5	0.9894	1.3310	0.0516	2.3549	0.5937	0.9812

Il primo e il terzo campione presentano risultati piuttosto soddisfacenti: si ha un valore superiore al limite del 95% (pari a 5.5139 per il Q e a 7.5041 per il T^2) solo per il residuo della prima osservazione e per la T^2 di Hotelling del punto reale del terzo campione. La seconda osservazione mostra una d_E confrontabile con il semiasse maggiore dell'ellisse di fiducia al 95% (l_1=4.2917, l_2=3.1905, l_3= 2.5701 e l_4=2.0987), una T^2 del punto reale di molto superiore al valore limite, indice di una probabile estrapolazione della soluzione (di conseguenza non si può garantire l'affidabilità del modello); anche l'incertezza è abbastanza alta (83.95%). Il quarto campione, oltre ad avere una distanza euclidea all'incirca pari al semiasse maggiore dell'ellisse di fiducia, presenta una peculiarità: è l'unico campione per cui $T^2_{real} < T^2_{invdir}$. Queste metriche, unite all'alta incertezza, consentono di affermare la scarsa bontà del risultato. La soluzione dell'ultimo campione viene estrapolata: $d_E \approx l_1$, T^2_{real} di molto superiore al limite del 95% e incertezza all'incirca del 98%.

4.2.2 Correlazioni tra le metriche

Per questo caso di studio l'analisi correlativa tra le diverse metriche di caratterizzazione dell'incertezza ha condotto alle medesime conclusioni: vi è una moderata correlazione tra incertezza, accuratezza e rappresentatività del modello; per questo motivo, per brevità non è riportata. Tuttavia, non si è confrontata l'incertezza media di calibrazione con quella di convalida vista la minore disponibilità di campioni di calibrazione (e l'impossibilità di aumentare il *set* di dati sperimentali) e la minore predittività mostrata dal modello.

4.3 Risultati per il caso di studio 3

L'ultima verifica delle metodologie proposte in questa Tesi è stata effettuata su dati di simulazione di una granulazione a secco con compattatore a rulli, descritta nel Capitolo 2. Vengono qui di seguito riportati i risultati per i diversi scenari ipotizzati.

4.3.1 Risultati per lo scenario 1

L'applicazione del metodo SF96 per la stima della varianza di una predizione e dell'approccio GDF per il calcolo dei gradi di libertà (da cui si sono ottenuti 7.22 gradi di libertà, a differenza delle 4 LV utilizzate nella costruzione del modello) hanno permesso di ottenere dei risultati simili agli altri casi di studio: la definizione degli spazi nulli in grado di fornire, teoricamente, la qualità di prodotto desiderata (y_{DES}, in questo caso particolare la porosità del granulato $\beta_{out,intra}$) all'interno dell'incertezza definita dal metodo SF96. Anche in questo caso lo spazio nullo ha dimensione $lv\text{-}V_y$=4-1=3, per questo non rappresentabile su un piano bidimensionale.

La Tabella 4.5 riassume le metriche, definite nel §4.1.3, per l'intero *set* di convalida di questo primo scenario.

Tabella 4.5: *Caratterizzazione dell'incertezza, dell'accuratezza e della rappresentatività nell'inversione diretta del modello PLS nel primo scenario del caso di studio 3: distanza euclidea, T^2 di Hotelling del punto di inversione diretta, T^2 di Hotelling del punto reale, Q del punto reale (rapportate al proprio limite) e incertezza.*

Campione di convalida	d_E/l_1	d_E/l_2	$T^2_{invdir}/T^2_{lim,95\%}$	$T^2_{real}/T^2_{lim,95\%}$	$Q_{real}/Q_{lim,95\%}$	α_{inc}
1	0.4731	0.9118	0.0303	0.5098	1.5219	0.0392
2	0.5012	0.9659	0.0363	0.6054	0.0885	0.5895
3	0.4491	0.8655	0.0294	0.8139	0.0343	0.7769
4	0.4197	0.8088	0.1241	0.7985	0.1035	0.1810
5	0.5649	1.0886	0.4967	0.9073	0.1276	0.7632
6	0.5609	1.0809	0.2673	0.5765	0.5517	0.4685
7	0.5790	1.1159	1.6303	1.2336	0.0936	0.9188
8	0.4544	0.8758	2.5719	1.3715	0.3812	0.9927
9	0.3087	0.5949	0.1645	0.3825	0.7375	0.7122
10	0.4680	0.9018	0.0043	0.8958	0.0724	0.7985
11	0.4885	0.9413	0.1805	0.4631	0.6431	0.4421
12	0.5878	1.1327	0.4840	1.0092	0.0455	0.6211

I campioni di convalida 3 e 4 presentano risultati molto buoni: d_E inferiori rispetto i semiassi di fiducia al 95% (l_1=4.7738, l_2=2.4778, l_3= 2.0106 e l_4=2.8275), T^2 di Hotelling e Q_{real} al di sotto del limite al 95% (6.5144 e 4.6585 rispettivamente) e significatività dell'incertezza relativamente basse. Per le osservazioni 2, 6, 10 e 11 l'unico valore non buono è la distanza euclidea confrontabile con il semiasse minore dell'ellisse di fiducia, compensato però da T^2 di Hotelling e Q_{real} inferiori al limite e da un'incertezza bassa. I campioni 7 e 8 offrono una peculiarità, già riscontrata nel caso di studio 2: $T^2_{real} < T^2_{invdir}$; inoltre sono i punti che presentano l'incertezza maggiore.

4.3.2 Correlazioni tra le metriche per lo scenario 1

L'obiettivo di questo Paragrafo è studiare le correlazioni esistenti tra le metriche di caratterizzazione dell'incertezza (d_E, T^2_{invdir}, T^2_{real}, Q_{real}, α_{inc}) di questo caso di studio, confrontando i risultati con quelli ottenuti per il caso di studio 1.

Utilizzando lo stesso approccio *jackknife* utilizzato per il caso di studio 1, si sono calcolate le metriche per l'intero *set* di calibrazione. Successivamente si è costruito un modello PCA sulla matrice [100×5] costituita dalle 5 metriche per i 100 campioni di calibrazione. Analizzando il diagramma dei *loading* si identificano le correlazioni tra le variabili e la loro significatività (Figura 4.12).

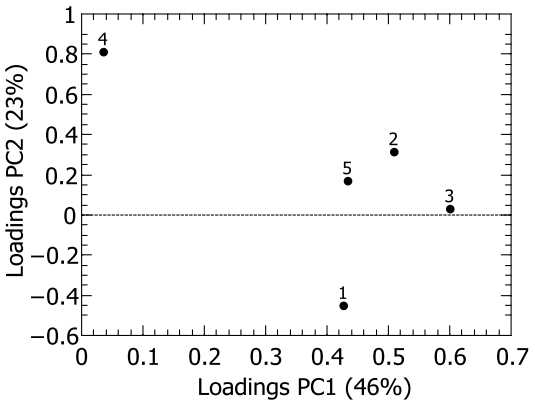


Figura 4.12: *Diagramma dei loading per le metriche di caratterizzazione dell'incertezza e del modello calcolate per il set di calibrazione.*

Lungo PC1, che spiega il 46% circa della variabilità dei dati, tutte le metriche sono, anche in questo caso, correlate positivamente. Come nel caso di studio 1, vi è una moderata correlazione sulla PC1 tra le variabili 1 (d_E) e 5 (α_{inc}). Questo a riprova di come maggiore è la distanza tra la soluzione dell'inversione diretta e la realtà fisica, più alta è l'incertezza. In questo caso di studio però la variabile 4 (Q_{real}) sembra avere un minor peso su PC1, contribuendo in misura minore all'incertezza. Sempre su PC1, le variabili 3 (T^2_{real}) e 5 (α_{inc}) sono modestamente correlate, come già visto nel caso di studio 1. In questa circostanza si aggiunge un'ulteriore correlazione lungo la PC2, che rende conto del 23% circa della variabilità dei dati, tra le variabili 2 (T^2_{invdir}) e 5 (α_{inc}). Questo significa che più la soluzione dell'inversione diretta si discosta dall'origine degli assi sul piano degli *score* (cioè dai valori medi, alta T^2_{invdir}), maggiore è l'incertezza.

Con la stessa procedura *jackknife* utilizzata nel caso di studio 1 si è caratterizzata l'incertezza media per il *set* di calibrazione, da usare per "certificare" il modello in fase di convalida. Successivamente, sul modello globale, si sono calcolati i valori di incertezza per l'intero *set* di convalida che, mediati, hanno fornito l'incertezza media per la fase di convalida. Questi valori (di calibrazione e di convalida) sono stati confrontati al variare del numero di campioni di calibrazione con cui viene costruito il modello PLS (Figura 4.13).

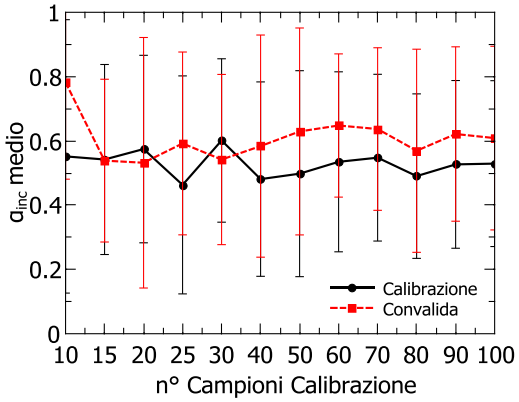


Figura 4.13: *Incertezze medie per la calibrazione e per la convalida al variare dei campioni di calibrazione. Le barre di errore indicano la deviazione standard dei valori di incertezza da cui si è ricavata l'incertezza media.*

Dalla Figura 4.13 si vede come l'incertezza in fase di calibrazione è generalmente minore di quella di convalida. All'aumentare del numero di campioni di calibrazione l'incertezza media non diminuisce sensibilmente ma, mentre con un numero ridotto di campioni in calibrazione il valore dell'incertezza mostra maggiore variabilità, l'incertezza si assesta stabilmente attorno al valore di 0.5 per la calibrazione e 0.6 per la convalida quando il numero di campioni di calibrazione è sufficientemente alto. A differenza del caso di studio 1, non si raggiunge un numero di campioni di calibrazione per cui i due valori coincidono, molto probabilmente a causa del numero ridotto di campioni a disposizione e dell'alta variabilità delle variabili.

Anche in questo caso si è verificata la collocazione dei punti con incertezza superiore al valore medio nel piano degli *score*, nel diagramma delle T^2 di Hotelling e nel diagramma dei residui Q.

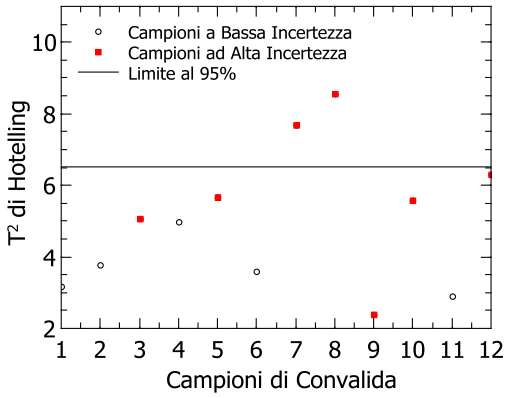


Figura 4.14: *T^2 di Hotelling per i dati di convalida e campioni di convalida ad alta incertezza. Modello PLS costruito su 100 dati di calibrazione. La linea tratteggiata indica il limite al 95% per la T^2 di Hotelling.*

Il diagramma di Figura 4.14, relativo alle T^2_{real} dei campioni di convalida, conferma la relazione che sussiste tra incertezza e distanza dai valori medi: i punti a maggiore incertezza sono quelli che con T^2_{real} mediamente più alta (ad eccezione del campione 9).

4.3.3 Risultati per lo scenario 2

Il secondo scenario differisce dal primo solo per la variabilità dei dati, che in questo caso viene accentuata, permettendo di distinguere con molta minor chiarezza le differenti classi di materiali. Sono stati calcolati, utilizzando il metodo GDF, 8.27 gradi di libertà, valore quasi doppio rispetto le variabili latenti sfruttate nella costruzione del modello PLS. In questo caso il modello è costruito con 5 LV; lo spazio nullo ha quindi dimensione lv-V_y=5-1=4. Sono state applicate le procedure già impiegate nei casi precedenti e i risultati sono riportati in Tabella 4.6.

Tabella 4.6: *Caratterizzazione dell'incertezza, dell'accuratezza e della rappresentatività nell'inversione diretta del modello PLS nel secondo scenario del caso di studio 3: distanza euclidea, T^2 di Hotelling del punto di inversione diretta, T^2 di Hotelling del punto reale, Q del punto reale (rapportate al proprio limite) e incertezza.*

Campione di convalida	d_E/l_1	d_E/l_2	$T^2_{invdir}/T^2_{lim,95\%}$	$T^2_{real}/T^2_{lim,95\%}$	$Q_{real}/Q_{lim,95\%}$	α_{inc}
1	0.6508	0.9522	0.5506	0.7216	0.4058	0.9932
2	0.5041	0.7376	0.0220	0.7001	0.2093	0.5241
3	0.4067	0.5950	0.0831	0.3108	0.1081	0.5653
4	0.3411	0.4991	0.0565	0.3401	0.0988	0.6425
5	0.5310	0.7768	0.0198	1.0100	0.1736	0.8371
6	0.3825	0.5597	0.0320	0.2826	0.2098	0.9687

I risultati ottenuti sono nel complesso soddisfacenti. Il secondo, terzo e quarto campione presentano ottimi valori: distanze euclidee inferiori ai semiassi dell'ellissoide di fiducia (che corrispondono a l_1=4.3430, l_2=2.9684, l_3= 2.4990, l_4=2.4297 e l_5=2.3905), sia T^2 che Q sono molto al di sotto del limite al 95% (6.5932 e 3.6112) e le incertezze sono abbastanza contenute. La prima osservazione presenta buoni valori per le T^2 di Hotelling e per il Q del punto reale, che si pongono al di sotto del limite. La distanza euclidea però è confrontabile, se non superiore, ai semiassi minori dell'ellisse di fiducia e l'incertezza è elevata (99% circa). I rimanenti due campioni analizzati hanno risultati buoni, d_E minori dei semiassi e Q al di sotto del limite (T^2 supera il valore al 95% solo per il quinto campione). Da notare però gli intervalli di fiducia all'interno dei quali si trova la proiezione del punto sperimentale reale abbastanza elevati, pari all'83.71% e al 96.87%.

Anche in questo scenario tutti le osservazioni analizzate hanno $T^2_{real} > T^2_{invdir}$.

4.3.4 Risultati per lo scenario 3

Nel costruire l'ultimo scenario si è considerata una variabilità dei dati molto maggiore rispetto al caso precedente che, come descritto nel Capitolo precedente, non permette di distinguere le diverse classi di materiali utilizzate. Il modello PLS è stato costruito con 4 LV e visto l'utilizzo di una matrice delle risposte **y** monovariata, lo spazio nullo ha dimensione $lv\text{-}V_y$=4-1=3. L'approccio GDF ha permesso di calcolare il numero più adatto di gradi di libertà (7.38) da utilizzare nella stima dell'intervallo di fiducia di una predizione.
La Tabella 4.7 riassume i risultati ottenuti per le procedure sull'utilizzo contemporaneo dei metodi sull'incertezza e sull'inversione.

Tabella 4.7: *Caratterizzazione dell'incertezza, dell'accuratezza e della rappresentatività nell'inversione diretta del modello PLS nel terzo scenario del caso di studio 3: distanza euclidea, T^2 di Hotelling del punto di inversione diretta, T^2 di Hotelling del punto reale, Q del punto reale (rapportate al proprio limite) e incertezza.*

Campione di convalida	d_E/l_1	d_E/l_2	$T^2_{\text{invdir}}/T^2_{\text{lim,95\%}}$	$T^2_{\text{real}}/T^2_{\text{lim,95\%}}$	$Q_{\text{real}}/Q_{\text{lim,95\%}}$	α_{inc}
1	0.6832	0.8889	0.0524	0.9151	0.8042	0.1112
2	0.3942	0.5130	0.0012	0.3728	0.0618	0.6367
3	0.4143	0.5390	0.1240	0.5538	0.2757	0.0873
4	0.3787	0.4928	0.0189	0.3099	0.0345	0.9241
5	0.6459	0.8404	0.0385	0.7965	0.1566	0.1555
6	0.6792	0.8837	0.1316	0.8957	0.6887	0.0292

Per questo scenario si sono calcolati i semiassi dell'ellisse di fiducia al 95% (l_1=4.2061, l_2=3.2325, l_3= 2.4007 e l_4=3.1638) e i limiti del 95% per le statistiche T^2 di Hotelling (6.6796) e per il residuo Q (4.4798). Il primo e il sesto campione di convalida presentano grandezze molto simili: una d_E confrontabile con i semiassi dell'ellisse di fiducia e una T^2 del punto reale molto vicina al limite, indice di una possibile cattiva predizione, nonostante il Q inferiore al limite e un'incertezza molto bassa (11.12% e 2.92% rispettivamente). La seconda e terza osservazione presentano i risultati migliori di tutto il *set*: le d_E sono inferiori ai semiassi dell'ellisse di fiducia, nessuna metrica supera il proprio limite e l'intervallo d'incertezza è piuttosto basso (soprattutto per il campione 3). L'osservazione 4 offre ottimi risultati per gran parte delle metriche prese in esame (possiede infatti una distanza euclidea minore dei semiassi di fiducia e le metriche statistiche classiche non superano i rispettivi limiti); tuttavia l'incertezza è abbastanza alta (92.41%). Il quinto campione nonostante un alto valore della d_E (rapportabile ai semiassi dell'ellisse di fiducia) e della T^2 del punto reale (vicina al limite al 95%) mostra buoni risultati per le altre grandezze analizzate: Q al di sotto del limite e un intervallo di incertezza contenente la proiezione del dato sperimentale reale basso (15.55%). Anche in questo scenario $T^2_{\text{invdir}} < T^2_{\text{real}}$, a causa della natura stessa dei modelli di regressione.

4.3.5 Correlazioni tra le metriche per gli scenari 2 e 3

L'analisi correlativa tra le metriche di caratterizzazione dell'incertezza e il confronto tra le incertezze medie di calibrazione e convalida porta alle medesime conclusioni dedotte dallo scenario 1. Per brevità non sono dunque riportati i risultati.

4.4 Conclusioni

In questo Capitolo sono stati riportati i risultati relativi alla verifica dei metodi per la stima dell'incertezza di predizione e alla caratterizzazione dell'incertezza.

Per quanto riguarda la verifica dei metodi più affidabili per la stima dell'incertezza, per il calcolo della deviazione standard dell'errore di predizione si suggerisce di utilizzare il metodo SF96, che ha avuto prestazioni superiori (ha permesso di ottenere probabilità di copertura più vicine alla significatività desiderata) e si è dimostrato più robusto, nel senso che è risultato essere meno sensibile alle variazioni di rumore e del numero di dati di calibrazione con cui il modello PLS viene costruito.

È preferibile stimare i gradi di libertà sfruttando l'approccio GDF date le sue migliori prestazioni, nonostante l'allungamento dei tempi di calcolo (dell'ordine dei minuti rispetto le risposte immediate fornite dagli altri metodi) e il ridotto scostamento dei risultati rispetto al metodo Naïve.

I risultati hanno trovato conferma in letteratura, suffragando il lavoro di Zhang e Garcìa-Muñoz (2009), che sono giunti alle medesime conclusioni. Questi risultati sono stati confermati anche dai casi di studio 2 e 3, che per brevità non sono stati riportati.

Una parte del lavoro è stata dedicata alla caratterizzazione dell'adeguatezza, della rappresentatività e dell'incertezza del modello nell'applicazione dell'inversione diretta, in modo da comprendere quanto la soluzione ottenuta si discosta dalla realtà fisica.

Si sono cercate le correlazioni tra l'incertezza e le metriche utilizzate per caratterizzare il risultato, in modo da collegare l'incertezza allo scostamento della soluzione dalla realtà fisica (d_E), alla distanza dai valori medi (T^2_{real}) e alla rappresentatività del modello (Q_{real}).

L'analisi correlativa tra le metriche di caratterizzazione del modello e dell'incertezza ha condotto alle seguenti conclusioni:

- vi è una moderata correlazione tra scostamento dalla realtà fisica (d_E), distanza dai valori medi (T^2_{real}) e incertezza (α_{inc}). Questo significa che più la soluzione ottenuta dall'inversione diretta si discosta dal punto sperimentale reale o più il punto reale è distante dall'origine degli assi del piano degli *score* (corrispondente ai valori medi), maggiore è l'incertezza. La variabile d_E contribuisce in misura maggiore alla determinazione dell'incertezza, in quanto la correlazione è stata riscontrata lungo la PC1, che spiega la percentuale maggiore di variabilità dei dati;

- nel primo caso di studio si è osservata una modesta correlazione tra rappresentatività del modello (Q_{real}) e incertezza (α_{inc}). Ciò significa che minore è la rappresentatività del modello (alto Q_{real}), più alta è l'incertezza.

L'analisi correlativa trova conferma nel controllo effettuato sul piano degli *score*, sul diagramma delle T^2 di Hotelling e sul diagramma dei residui Q. I punti a più alta incertezza sono quelli che presentano la T^2 e il Q maggiore.

È stata poi proposta una procedura per "certificare" il modello in fase di calibrazione. La procedura, un approccio *jackknife* permette di calcolare l'incertezza media per il *set* di calibrazione. Questo valore è stato confrontato con il valore medio ottenuto dal *set* di convalida, studiando la sensitività dell'incertezza in calibrazione e convalida al numero di campioni di calibrazione con cui viene costruito il modello PLS. Questo confronto ha portato alle seguenti conclusioni:

- all'aumentare del numero di campioni di calibrazione, l'incertezza media diminuisce sia in fase di calibrazione che di convalida;
- l'incertezza in fase di convalida è generalmente maggiore dell'incertezza in fase di calibrazione;
- in generale viene raggiunto un numero di campioni di calibrazione per cui l'incertezza di calibrazione e quella di convalida si avvicinano (o al limite coincidono) e mantengono un valore all'incirca costante. Questo significa che quando la variabilità dei dati nel *set* di calibrazione è pienamente rappresentativa della variabilità dei dati del *set* di convalida, le due incertezze coincidono. Per un numero di campioni di calibrazione ridotto, l'incertezza presenta valori mediamente più alti e il suo comportamento è instabile.

Conclusioni

In questa Tesi è stata proposta una procedura per caratterizzare l'incertezza nella modellazione a variabili latenti, con lo scopo di valutare e aumentare l'affidabilità nell'utilizzo di tali tecniche nella progettazione di nuovi prodotti e processi. In particolare, la Tesi mira a caratterizzare la ricaduta dell'incertezza di predizione dalla qualità di prodotto desiderata sulle qualità delle materie prime e sui parametri di processo stimati dal modello a variabili latenti.

I contributi della Tesi sono stati principalmente due: la verifica dei metodi più promettenti tra quelli proposti da Zhang e Garcìa-Muñoz (2009) per stimare l'incertezza di predizione, e la caratterizzazione dell'incertezza in fase di calibrazione e convalida dei modelli a variabili latenti, studiandone le correlazioni con l'accuratezza e la rappresentatività del modello.

Per quanto riguarda il primo contributo, è stato verificato che i metodi suggeriti da Zhang e Garcìa-Muñoz (2009) sono i più idonei a caratterizzare l'incertezza di predizione.

Nella seconda parte della Tesi si sono analizzate le correlazioni tra l'incertezza, l'accuratezza e la rappresentatività del modello, in modo da osservare se vi sia un legame tra l'incertezza e la differenza tra il valore stimato dal modello e il valore reale delle caratteristiche delle materie prime e dei parametri di processo utilizzati nella progettazione al fine di ottenere un prodotto di una qualità predeterminata. Dallo studio si è rilevata l'esistenza di correlazione tra incertezza, accuratezza e rappresentatività dei modelli costruiti. I risultati mostrano come maggiore è la rappresentatività e l'accuratezza del modello, tanto minore è l'incertezza.

Infine, è stata proposta una procedura per caratterizzare l'incertezza media del modello in calibrazione e convalida, e si è studiata la sensitività dell'incertezza al numero di campioni di calibrazione con cui viene costruito il modello a variabili latenti. In particolare è stata proposta una procedura *jackknife* (Miller, 1964) per "certificare" il modello in fase di calibrazione in termini di incertezza media per il *set* di calibrazione. Questo valore è stato poi confrontato con l'incertezza media ottenuta in fase di convalida. In generale si è osservato che l'incertezza diminuisce all'aumentare del numero di campioni di calibrazione, sia in fase di calibrazione che di convalida, e che l'incertezza media in convalida è maggiore di quella in calibrazione. Inoltre, è stato rilevato che nel caso si sia in possesso di un numero di campioni di calibrazione sufficientemente alto (in funzione dalle caratteristiche di variabilità e correlazione dei dati disponibili), esiste un numero di campioni "ideale", per cui l'incertezza di calibrazione e convalida si assestano attorno ad un valore all'incirca stabile e tendenzialmente si avvicinano (al limite coincidono). Il valore comune d'incertezza rappresenta il limite inferiore di incertezza per quel caso di studio preso in esame.

Le metodologie proposte sono state applicate a tre diversi casi di studio: la determinazione degli ingressi ad un modello matematico, la progettazione di prodotto di una granulazione umida (Vemavarapu *et al.*, 2009) e la progettazione di prodotto e processo di una granulazione a secco di cellulosa microcristallina simulata. Queste applicazioni sono rappresentative di situazioni realmente riscontrabili da progettisti di processo e formulatori dell'industria manifatturiera e farmaceutica. Il primo caso di studio è stato costruito per valutare i metodi di stima dell'incertezza. Il secondo esempio applicativo tratta un caso reale di interesse farmaceutico, con uno studio che coinvolge un numero ridotto di campioni, rispecchiando quindi una sperimentazione reale. Il terzo caso di studio riguarda un processo di granulazione a secco mediante un compattatore a rulli, un'operazione largamente utilizzata nell'industria farmaceutica, chimica ed alimentare e quindi di notevole interesse.

In base ai promettenti risultati ottenuti, si auspicano ulteriori sviluppi futuri delle metodologie proposte, per esempio considerando non un'unica variabile risposta, ma diverse caratteristiche di prodotto contemporaneamente (cioè estendere la stima dell'incertezza ad una risposta multivariata). Ulteriori sviluppi futuri potrebbero riguardare la valutazione congiunta dei diversi tipi di incertezza che influenzano la modellazione a variabili latenti e la possibilità di applicare il concetto di incertezza all'inversione mediante la risoluzione di un problema di ottimizzazione (Tomba *et al.*, 2012), in grado di tener conto delle esigenze economiche e qualitative del prodotto.

Nomenclatura

a	=	componente principale generica (-)
A	=	numero di componenti principali (-)
$\mathbf{b}$	=	vettore dei coefficienti della relazione interna (-)
b_a	=	elemento del vettore dei coefficienti della relazione interna b (-)
CI	=	intervallo di fiducia (-)
cp_{SF96}	=	probabilità di copertura ottenuta dal metodo Simple-Faber96 (-)
cp_{UD}	=	probabilità di copertura calcolata dal metodo U-deviation (-)
D	=	diametro del cilindro (m)
d	=	spessore della bricchetta solida quando S = 0 (m)
d_E	=	distanza euclidea tra punto sperimentale reale e punto di inversione diretta (-)
df	=	numero di gradi di libertà (-)
E_{add}	=	energia aggiunta durante la compattazione (J)
$\mathbf{e}_{N+1}$	=	elemento della matrice dei residui relativo all'osservazione N+1 (-)
$\mathbf{E}_{\text{U}}$	=	matrice dei residui di **U** (-)
$\mathbf{E}_{\text{Y}}$	=	matrice dei residui di **Y** multivariata (-)
$\mathbf{E}_{\text{y}}$	=	matrice dei residui di **y** monovariata (-)
F	=	valore della distribuzione di Fisher-Snedecor (-)/variabile adimensionale utilizzata nel calcolo della pressione massima esercitata dal compattatore (-)
F_{roll}	=	forza applicata dal cilindro (kN)
F_{sb}	=	fattore di *springback* (m)
$\mathbf{G}$	=	matrice della decomposizione ai valori singolari (-)
gdf	=	numero di gradi di libertà dati dall'approccio GDF (-)
h_{N+1}	=	leveraggio per l'osservazione N+1 (-)
$\mathbf{J}$	=	matrice degli *score* di **Y** (-)
k	=	indice delle iterazioni nell'approccio GDF (-)
K	=	numero di iterazioni nell'approccio GDF (-)
$\mathbf{k}$	=	coefficienti del modello di processo di **y** (-)
lv	=	numero di variabili latenti (-)
$\mathbf{M}_a$	=	matrici di rango unitario di decomposizione di **U** nella PCA (-)
MRE_{cal}	=	errore medio relativo percentuale in fase di calibrazione (-)
MRE_{val}	=	errore medio relativo percentuale in fase di convalida (-)
N	=	numero di campioni/dati di calibrazione (-)
n_σ	=	numero di varianze corrispondenti alla proiezione del punto sperimentale reale sulla distribuzione di incertezza (-)

$\mathbf{p}$	=	vettore degli *loading* (-)
$\mathbf{P}$	=	matrice dei *loading* di **U** (-)
pdf	=	numero di gradi di libertà dati dall'approccio PDF (-)
$P_{\max}$	=	pressione massima tra i due cilindri (Pa)
$\mathbf{Q}$	=	matrice dei *loading* di **Y** (-)
$\mathbf{q}$	=	vettore dei *loading* di **Y** (-)
Q	=	somma dei quadrati di ogni riga della matrice dei residui **E** (-)
Q_{real}	=	errore quadratico medio del punto sperimentale reale (-)
r	=	indicatore generico per il rango di una matrice (-)
R_{U}	=	rango della matrice **U** (-)
R_{Y}	=	rango della matrice **Y** (-)
s	=	deviazione standard dell'errore di predizione (-)
S	=	*gap* tra i due cilindri (m)
l_a	=	semiasse dell'ellisse di fiducia (-)
$S_{\max}$	=	*gap* massimo tra i due cilindri (m)
$s_{\text{U,pr}}$	=	varianza del residuo del predittore per l'osservazione N+1 (-)
$s_{\text{Utol,val}}$	=	varianza media del residuo del predittore nel *set* di convalida (-)
$s_{\text{y\_val}}$	=	varianza del residuo di **y** nel set di convalida (-)
$\mathbf{t}$	=	vettore degli *score* (-)
$\mathbf{T}$	=	matrice degli *score* di **U** (-)
t	=	valore della *t* di *Student* (-)
$T^2_{A,N,\alpha}$	=	limite di fiducia per il diagramma degli *score* (-)
T^2_{invdir}	=	T^2 di Hotelling per il punto di inversione diretta (-)
T^2_{real}	=	T^2 di Hotelling per il punto sperimentale reale (-)
$\mathbf{t}_{\text{DES}}$	=	*score* corrispondenti a y_{DES} (-)
$\mathbf{t}_{\text{NULL}}$	=	*score* corrispondenti a $\mathbf{u}_{\text{NULL}}$ (-)
$\mathbf{U}$	=	matrice dei dati di ingresso del processo (-)
u	=	generica variabile manipolabile (-)
$\mathbf{U}^{(i)}$	=	matrice dei dati di ingresso a cui è stata rimossa la riga *i*-esima (-)
$\mathbf{u}_{\text{NEW}}$	=	nuove condizioni di ingresso (-)
$\mathbf{u}_{\text{NULL}}$	=	condizioni di ingresso appartenenti allo spazio nullo (-)
$\mathbf{u}_{N+1}$	=	vettore predittore per l'osservazione N+1 (-)
$\mathbf{u}_{\text{PRED}}$	=	condizioni di ingresso a cui vengono aggiunte quelle dello spazio nullo (-)
$\mathbf{V}$	=	matrice della decomposizione ai valori singolari (-)
v_{roll}	=	velocità del cilindro del compattatore a rulli (rpm)
V_U	=	numero di variabili misurate in **U** (-)
V_y	=	numero di variabili misurate in **Y** (-)

$\mathbf{W}$	=	matrice dei pesi (-)
$\mathbf{w}$	=	vettore dei pesi (-)
W	=	ampiezza del cilindro (m)
x	=	generica variabile misurabile (-)
$\mathbf{Y}$	=	matrice delle risposte multivariata (-)
$\mathbf{y}$	=	vettore delle risposte monovariate (-)
$\bar{y}$	=	risposta media misurata (-)
$\mathbf{y}^{(i)}$	=	vettore delle risposte a cui è stata rimossa la riga *i*-esima (-)
$y^*_{i,k}$	=	y *i*-esima a cui è stato aggiunto il rumore $\sigma_{i,k}$ (-)
$\mathbf{y}_0$	=	vettore delle risposte a cui non è stato aggiunto rumore (-)
$\mathbf{y}_{10}$	=	vettore delle risposte a cui è stato aggiunto del rumore con media 0 e varianza il 10% della varianza di $\mathbf{y}_0$ (-)
$\mathbf{y}_{40}$	=	vettore delle risposte a cui è stato aggiunto del rumore con media 0 e varianza il 40% della varianza di $\mathbf{y}_0$ (-)
$\mathbf{y}_{70}$	=	vettore delle risposte a cui è stato aggiunto del rumore con media 0 e varianza il 70% della varianza di $\mathbf{y}_0$ (-)
$y_{cal,i}$	=	valore assunto dalla risposta *i*-esima nel *set* di calibrazione (-)
$\hat{y}_{cal,i}$	=	valore predetto dal modello per la risposta *i*-esima nel *set* di calibrazione (-)
$\mathbf{y}_{DES}$	=	*set* di variabili di risposta desiderato (-)
y_{inc}	=	y corrispondente alla proiezione del punto reale sulla distribuzione di incertezza (-)
$\hat{y}_{N+1}$	=	valore assunto dalla risposta per una nuova osservazione N+1 (-)

Lettere greche

α	=	significatività per l'intervallo di fiducia (-)
α_{inc}	=	percentile di incertezza (-)
α_{NIP}	=	angolo di *nip* (rad)
$\hat{\boldsymbol{\beta}}$	=	coefficienti del modello di regressione lineare (-)
$\beta_{in,inter}$	=	porosità esterna dei solidi in entrata al compattatore a rulli (m^3/m^3)
$\beta_{in,intra}$	=	porosità interna dei solidi in entrata al compattatore a rulli (m^3/m^3)
$\hat{\boldsymbol{\beta}}_{OLS}$	=	coefficienti del modello di regressione lineare ai minimi quadrati ordinari (-)
$\beta_{out,inter}$	=	porosità esterna dei solidi in uscita dal compattatore a rulli (m^3/m^3)
$\beta_{out,intra}$	=	porosità interna dei solidi in uscita dal compattatore a rulli (m^3/m^3)
δ	=	costante arbitraria (-)
δ_{EFF}	=	angolo effettivo di attrito (rad)
δ_{FEED}	=	angolo di alimentazione della polvere nel compattatore a rulli (rad)
δ_{FR}	=	angolo di attrito tra solidi e cilindri del compattatore a rulli (rad)
θ	=	differenza tra l'angolo formato dal piano orizzontale e l'angolo di *nip* (rad)
κ	=	costante di comprimibilità del granulato solido (-)

λ	=	autovalori (-)
μ	=	angolo di scivolamento (rad)
ν	=	angolo acuto tra la direzione principale di sforzo e la tangente alla superficie del cilindro (rad)
$\rho_{in,bulk}$	=	densità di *bulk* all'ingresso del compattatore a rulli (kg/m^3)
$\rho_{in,particle}$	=	densità della particella all'ingresso del compattatore a rulli (kg/m^3)
ρ_{max}	=	densità massima ottenuta dal compattatore a rulli (kg/m^3)
$\rho_{out,bulk}$	=	densità di *bulk* all'uscita del compattatore a rulli (kg/m^3)
$\rho_{out,particle}$	=	densità della particella all'uscita del compattatore a rulli (kg/m^3)
$\rho_{skeletal}$	=	densità vera della polvere (kg/m^3)
$\hat{\sigma}$	=	estimatore della deviazione standard del rumore (-)
Σ	=	matrice della decomposizione ai valori singolari (-)
σ^2	=	varianza (-)
$\sigma_{i,k}$	=	rumore per l'oggetto *i*-esimo all'iterazione *k* (-)
σ_θ	=	sforzo normale medio (Pa)
ϕ	=	intercetta per il fitting lineare tra le coppie $y^*_{i,k}$ e $\sigma_{i,k}$ (-)

Acronimi

CI	=	intervallo di fiducia (*confidence interval*)
GDF	=	gradi di libertà generalizzati (*generalized degrees of freedom*)
LV	=	variabile latente (*latent variable*)
LVM	=	modellazione a variabili latenti (*latent variable modeling*)
LVRM	=	regressione lineare alle variabili latenti (*latent variable regression model*)
MRE	=	errore medio relativo (*mean relative error*)
OLS	=	metodo ai minimi quadrati ordinari (*ordinary least squares*)
PC	=	componenti principali
PCA	=	analisi delle componenti principali (*principal component analysis*)
PDF	=	pseudo gradi di libertà (*pseudo degrees of freedom*)
PLS	=	proiezione su strutture latenti (*projection on latent structures*)
PRESS	=	somma dei quadrati dell'errore di predizione
PSD	=	distribuzione di dimensione delle particelle (*particle size distribution*)
RMSECV	=	errore medio quadratico della convalida incrociata (*root mean squared error of cross-validation*)
SEC	=	errore standard di calibrazione (*standard error of calibration*)
SF96	=	metodo *Simple-Faber96*
TSM	=	tecniche statistiche multivariate
UD	=	metodo *U-deviation*

Riferimenti bibliografici

Baumann, K. e N. Stiefl (2004). Validation tools for variable subset regression. *J. Comput. Aided Mol. Des.*, **18**, 549-562.

De Vries, S. e C.J.F. Ter Braak (1995). Prediction error in partial least squares regression: A critique on the deviation used in the Unscrambler. *Chemom. Intell. Lab. Syst.*, **30**, 239-245.

Faber, K. e B.R. Kowalski (1996). Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler. *Chemom. Intell. Lab. Syst.*, **34**, 283-292.

Facco, P., F. Doplicher, F. Bezzo e M. Barolo (2009). Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process. *J. Process Control*, **19**, 520-529.

FDA (2004). Pharmaceutical CGMPs for the 21st century - A risk based approach. Report finale. *U.S. Department of Healt and Human Services. U.S. Food and Drug Administration.*

Fernandez Pierna, J.A., L. Jina, F. Wahl, N.M. Faber e D.L. Massart (2003). Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. *Chemom. Intell. Lab. Syst.*, **65**, 281-291.

Garcìa-Muñoz, S., T. Kourti e J.F. MacGregor (2003). Troubleshooting of an industrial batch process using multivariate methods. *Ind. Eng. Chem. Res.*, **42**, 3592-3601.

Garcìa-Muñoz, S., T. Kourti e J.F. MacGregor, F. Apruzzese e M. Champagne (2006).Optimization of batch operating policies. Part I. Handling multiple solutions. *Ind. Eng. Chem. Res.*, **45**, 7856-7866.

Geladi, P. e B. Kowalski (1986). Partial least squares: a tutorial. *Anal. Chim. Acta*, **185**, 1-17.

Giannetti, P. (2012). Studio sperimentale di metodologie alternative di aggiunta del legante nella granulazione low shear. *Tesi di Laurea Specialistica in Ingegneria Chimica per lo Sviluppo Sostenibile*, Dipartimento di Ingegneria Industriale, Università di Padova.

Guigon, P., O. Simon, K. Saleh, G. Bindhumadhavan, M.J. Adams e J.P.K. Seville (2007). Roll pressing. In*: Handbook of powder technology-Granulation (*Elsevier Ltd., Amsterdam, Netherlands).

Höskuldsson, A. (1988). PLS regression methods. *J. Chemom.*, **2**, 211-228.

Jackson, J.E. (1991). *A user's guide to principal components*. John Wiley & Sons Inc., New York (U.S.A.).

Jaeckle, C.M. e J.F. MacGregor (1998). Product design through multivariate statistical analysis of process data. *AiChE J.*, **44**, 1105-1118.

Jaeckle, C.M. e J.F. MacGregor (2000). Industrial applications of product design through the inversion of latent variable models. *Chemom. Intell. Lab. Syst.*, **50**, 199-210.

Johanson, J.R. (1965). A rolling theory for granular solids. *J. Appl. Mech.*, **14**, 842-848.

Martens, H. e M. Martens (2000). Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference*, **11**, 5-16.

McCormick, D. (2005). Evolutions in direct compression. *Pharm. Technol.*, **17**, 52-62.

Miller, R.G. (1964). A trustworthy jackknife. *Ann. Math. Statist.*, **35**, 1594-1605.

Nomikos, P. e J.F. MacGregor (1995). Multi-way partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.*, **30**, 97-108.

Reis, R.S. e P.M. Saraiva (2005). Integration of data uncertainty in linear regression and process optimization. *AIChE J.*, **51**, 3007-3019.

Souihi, N., M. Josefson, P. Tajarobi, B. Gururajan e J. Trygg (2013). Design space estimation of the roller compaction process. *Ind. Eng. Chem. Res.*, **52**, 12408-12419.

Tomba, E., M. De Martin, P. Facco, J. Robertson, S. Zomer, F. Bezzo e M. Barolo (2013). General procedure to aid the development of continuos pharmaceutical processes using multivariate statistical modeling - An industrial case study. *Int. J. Pharm.*, **444**, 25-39.

Tomba, E., M. Barolo e S. Garcìa-Muñoz (2012). General framework for latent variable model inversion for the design and manufacturing of new products. *Ind. Eng. Chem. Res.*, **51**, 12886-12900.

van der Voet, H. (1999). Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *J. Chemom.*, **13**, 195-208.

Vemavarapu, C., M. Surapaneni, M. Hussain e S. Badawy (2009). Role of drug substance material properties in the processability and performance of a wet granulated product. *Int. J. Pharm.*, **374** , 96-105.

Walczak, B. e D.L. Massart (2001). Dealing with missing data: Part I. *Chemom. Intell. Lab. Syst.*, **58**, 15-27.

Wennerstrum, S. (2000). Ten things you need to consider when choosing and installing a roller press system. *Powder Bulk Eng.*, **14**, 37-50.

Wise, B. e N.B. Gallagher (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Control*, **6**, 329-348.

Wold, S., H. Martens e H. Wold (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Lectures in Math.*, **973**, 286-293.

Zhang, L. e S. Garcìa-Muñoz (2009). A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective. *Chemom. Intell. Lab. Syst.*, **97**, 152-158.