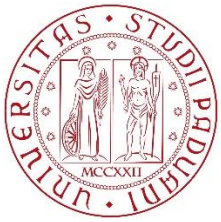


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale in Ingegneria Chimica e dei Processi Industriali

METODOLOGIE PER LA DIAGNOSI DI MODELLI A PRINCIPI PRIMI PER SISTEMI DINAMICI

Relatore: Prof. Massimiliano Barolo
Correlatrice: Ing. Natascia Meneghetti

Laureando: AMIR IBRAHIM

ANNO ACCADEMICO 2015 – 2016

Riassunto

In questa Tesi è affrontato il problema della diagnosi del possibile disallineamento (*process/model mismatch*, PMM) tra le misure di un processo e le relative predizioni da parte di un modello a principi primi. L'obiettivo della Tesi è comparare due metodologie di diagnosi di un PMM considerando un modello a principi primi sviluppato per descrivere un processo dinamico di fermentazione. Tali metodologie confrontano le strutture di correlazione di due set di dati, relativi alle misure del processo e alle predizioni da parte di un modello a principi primi. In particolare, la prima metodologia, elaborata da Meneghetti *et al.* (2014), effettua tale confronto utilizzando l'analisi delle componenti principali (PCA), mentre la seconda metodologia si basa sull'utilizzo dei coefficienti di correlazione parziale secondo una procedura elaborata da Rato e Reis (2015) nel campo del controllo statistico di processo.

Entrambe le metodologie vengono testate considerando due diversi tipi di *mismatch* parametrici, per i quali l'analisi viene effettuata sia utilizzando solo i dati in ingresso e i dati in uscita di fine batch, sia utilizzando un maggior numero di dati per tenere conto della dinamica del processo. I risultati ottenuti con la prima metodologia risultano promettenti se si considera la dinamica del processo, soprattutto per quanto riguarda la prima causa di *mismatch* analizzata, mentre rivelano i limiti dell'analisi dovuti alla forte correlazione delle variabili se si considerano solo i dati di inizio e di fine batch. Con la seconda metodologia si ottengono risultati molto promettenti utilizzando solo i dati iniziali e finali, che suggeriscono la possibilità di sfruttare questa tecnica per risolvere almeno parzialmente il problema dovuto alla presenza di variabili molto correlate in sistemi complessi. A tale scopo, sono state suggerite alcune soluzioni per l'utilizzo di tale metodologia anche considerando un set esteso di dati, in cui problemi di autocorrelazione e correlazione incrociata sono rilevanti. Sebbene nessuna di tali soluzioni possa essere utilizzata direttamente per la diagnosi di un PMM, i risultati ottenuti offrono numerose informazioni e opportunità per l'elaborazione di una diversa tecnica di diagnosi.

Indice

Nomenclatura

a = indicatore generico per il numero di componenti principali (-)
A = numero di componenti principali (-)
B = numero di simulazioni dell'n-esimo campione (-)
b_a = generico elemento del vettore dei coefficienti di regressione (-)
$\mathbf{C}$ = matrice di covarianza relativa a $\mathbf{X}$ (-)
C_{ox} = concentrazione di ossigeno (g/L)
C_p = concentrazione di penicillina (g/L)
C_s = concentrazione di substrato (g/L)
C_x = concentrazione di biomassa (g/L)
$\mathbf{D}$ = matrice di diagnosi (-)
$e_{c,k}$ = errore di ricostruzione corrispondente a $\hat{x}^g_{c,k}$ (-)
$\mathbf{e}_i$ = generico vettore della matrice dei residui $\mathbf{E}$
$\mathbf{e}_v$ = vettore colonna di $\mathbf{E}_M$ (-)
$\mathbf{E}$ = matrice degli errori statistici multivariati per la matrice $\mathbf{X}$ (-)
$\mathbf{E}_M$ = matrice degli errori statistici multivariati per la matrice $\mathbf{X}_M$ (-)
$\mathbf{E}_\Pi$ = matrice degli errori statistici multivariati per la matrice $\mathbf{X}_\Pi$ (-)
F = portata di substrato (L/h)
f_g = portata di aria (L/h)
$\mathbf{F}$ = matrice degli errori statistici multivariati per la matrice $\mathbf{Y}$ (-)
G = numero di sottogruppi per la convalida incrociata (-)
i = indicatore generico di un'osservazione o pedice generico (-)
k = indicatore generico per le variabili (-)
K = numero di variabili considerate in un set di dati (-)
K_la = coefficiente volumetrico di trasporto di massa per l'ossigeno (h^{-1})
L = numero di matrici di *lag* inserite all'interno di $\tilde{\mathbf{X}}$ (-)
$\mathbf{L}$ = matrice triangolare inferiore decomposta dalla matrice di covarianza $\mathbf{\Sigma}$ (-)
M = numero di variabili considerate nella matrice delle risposte (-)
M = indice generico dei dati di modello
$\mathbf{m}$ = vettore delle medie
$MRLR_v$ = valore del *mean residuals-to-limit ratio* per la v-esima variabile ausiliaria (-)
n = indicatore generico per i campioni (-)
N = numero dei campioni considerati in un set di dati (-)
P = potenza di agitazione (W)
$\mathbf{p}_i$ = generico vettore della matrice dei *loadings* $\mathbf{P}$ (-)

$\mathbf{P}$ = matrice dei *loadings* (-)

$\mathbf{P}_M$ = matrice dei *loadings* del modello PCA calibrato su $\mathbf{X}_M$ (-)

$PRESS_g$ = errore di predizione sulla somma dei quadrati dei residui (-)

q = grado del coefficiente di correlazione parziale (-)

$r_{x,y}$ = coefficiente di correlazione tra le generiche variabili x, y (-)

$r_{x,y.z}$ = coefficiente di correlazione parziale tra le generiche variabili x, y, z (-)

SPE_i = errore di predizione al quadrato per il generico campione i (-)

t = generico istante di tempo (-)

T = numero degli istanti di tempo del processo (-)

T^2 = statistica di Hotelling (-)

$\mathbf{t}_{CONT,i}$ = generico vettore dei contributi delle variabili alla statistica T^2 del campione (-)

$\mathbf{t}_i$ = generico vettore della matrice degli *scores* $\mathbf{T}$ (-)

$\mathbf{T}$ = matrice degli *scores* (-)

$\mathbf{T}_M$ = matrice degli *scores* per i dati di modello (-)

TSS = *total sum of squares* (-)

$\mathbf{T}_\Pi$ = matrice degli *scores* per i dati di processo (-)

$\mathbf{u}_a$ = generico vettore della matrice degli *scores* di $\mathbf{Y}$ (-)

$\mathbf{U}$ = matrice bidimensionale delle variabili decorrelate (-)

$\widetilde{\mathbf{U}}$ = matrice estesa delle variabili decorrelate (-)

v = indicatore generico per le variabili ausiliarie (-)

V = numero di variabili ausiliarie (-)

VIP_j = *variable influence on projection* per la j-esima variabile (-)

$\mathbf{w}_a$ = generico vettore della matrice dei *weights* di $\mathbf{X}$ (-)

$w_{i,j.k}$ = coefficiente di correlazione parziale normalizzato tra le variabili x, y, z (-)

x_i = generica variabile ausiliaria (-)

$x_{n,k}$ = elemento della n-esima riga, k-esima colonna di $\mathbf{X}$ (-)

$\hat{x}_{n,k}$ = elemento della n-esima riga, k-esima colonna ricostruito con il modello PCA (-)

$\hat{x}^g_{c,k}$ = elemento ricostruito della c-esima riga, k-esima colonna di $\mathbf{X}_g$ (-)

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

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

$\mathbf{X}_{dati,M}$ = matrice bidimensionale dei dati di modello (-)

$\mathbf{X}_{dati,\Pi}$ = matrice bidimensionale dei dati di processo (-)

$\mathbf{X}_M$ = matrice bidimensionale dei dati di modello (variabili ausiliarie) (-)

$\underline{\mathbf{X}}_M$ = matrice tridimensionale dei dati di modello (variabili ausiliarie) (-)

$\mathbf{X}_\Pi$ = matrice bidimensionale dei dati di processo (variabili ausiliarie) (-)

$\underline{\mathbf{X}}_\Pi$ = matrice tridimensionale dei dati di processo (variabili ausiliarie) (-)

$\hat{\mathbf{X}}$ = matrice ricostruita dal modello PCA (-)

$\mathbf{X}_g$ = matrice del sottogruppo del set di campioni per la convalida incrociata (-)

$\tilde{\mathbf{X}}$ = matrice estesa
Y_{sx} = costante di resa in biomassa rispetto al substrato (-)
$\mathbf{Y}$ = matrice bidimensionale delle variabili di risposta (-)
$z_{\alpha/2}$ = statistica *z* (-)

Apici

$^{-1}$ = inversa di una matrice

Lettere greche

λ = autovalore della matrice di covarianza **C** (-)
$\mathbf{\Lambda}$ = matrice degli autovalori (-)
ρ = media della distribuzione di coefficienti di correlazione parziale (-)
σ = deviazione standard (-)
α = percentuale di confidenza (-)
Π = indice generico dei dati di processo (-)
μ = velocità di crescita specifica perla biomassa (h^{-1})
μ_{pp} = velocità di crescita specifica per la penicillina (h^{-1})
τ = istante di campionamento (h^{-1})
Δt = intervallo di campionamento
$\mathbf{\Sigma}$ = matrice di covarianza del set di dati dinamico (-)
$\tilde{\mathbf{\Sigma}}$ = matrice di covarianza estesa (-)

Acronimi

ESS = *sum of square errors* (-)
MPCA = multi-way PCA
NOC = condizioni operative normali
PC = componenti principali
PCA = metodo dell'analisi delle componenti principali
PCC = coefficiente di correlazione parziale
PID = proporzionale integrale differenziale
PLS = *partial least squares regression*

Introduzione

Diverse attività industriali, come l'ottimizzazione e il controllo di processo, richiedono la disponibilità di modelli affidabili per la descrizione dei processi in atto. I principali modelli elaborati possono essere basati sull'utilizzo di dati (modelli empirici, *data-driven*, DD) o sulla conoscenza fisica del processo (modelli a principi primi, *first-principles*, FP). I modelli a principi primi sono costituiti da un insieme di equazioni che contengono al loro interno dei parametri: le equazioni possono essere bilanci di materia, di energia, di quantità di moto ed equazioni costitutive. Quando i risultati di un modello a principi primi (a partire dalle stesse condizioni operative e valori inziali delle variabili di processo) vengono comparati con le misure storiche di un processo, possono risultare non conformi a tali valori. Tale disallineamento tra i dati di processo e le predizioni del modello, o *process-model mismatch* (PMM), può essere dovuto a:

1. inadeguatezza delle equazioni del modello nel descrivere i fenomeni presenti nel processo a causa di un'incompleta conoscenza del processo stesso, dunque un'errata formulazione delle equazioni del modello (*mismatch* strutturale);
2. valori errati dei parametri assegnati alle equazioni di modello, che possono non riflettere il fenomeno in atto a causa di diverse condizioni operative o ambientali in cui sono stati misurati i parametri (*mismatch* parametrico; per esempio, il valore di un parametro ricavato in sede sperimentale può non essere adatto a descrivere lo stesso fenomeno riprodotto in una scala maggiore).

Rilevare il PMM può essere di grande importanza se si considera che il modello a principi primi considerato può essere sfruttato per il progetto di processi, o per ottimizzare la conduzione di un processo in apparecchiature o impianti su scala industriale.

In presenza di un PMM, è quindi necessario migliorare il modello tramite la riformulazione delle equazioni e/o una stima corretta dei parametri. Questa operazione è relativamente facile se si conosce in anticipo quale parte del modello o quali parametri rappresentino la causa principale del mismatch, grazie alla conoscenza ingegneristica del modello a principi primi e del fenomeno descritto. In generale però, questo tipo di informazioni non è disponibile, per cui è necessario effettuare nuovi esperimenti per poter ottenere un modello affidabile.

A tal scopo, recentemente è stata proposta una metodologia (Meneghetti *et al.*, 2014) per ridurre la quantità di esperimenti e il tempo necessari per migliorare un modello a principi primi affetto da errore. Tale metodologia si basa sul confronto delle strutture di correlazione di una matrice ricavata dai dati storici di processo e di una ricavata delle relative predizioni del modello a principi primi, tramite l'analisi alle componenti principali (*principal component analysis*, PCA) al fine di diagnosticare la presenza del PMM e risalire alla sua causa. I test effettuati

considerando diversi casi studio hanno rilevato alcuni limiti nell'applicazione di tale metodologia, soprattutto nel caso dell'analisi di modelli a principi primi sviluppati per la descrizione di processi dinamici.

L'obiettivo di questa Tesi è pertanto la comparazione di tale metodologia con una soluzione alternativa per la diagnosi di un PMM, in modo da rilevare i limiti e le potenziali aree di miglioramento di entrambe le metodologie. La seconda metodologia proposta si basa sugli studi condotti da Rato e Reis (2015) sull'uso dei coefficienti di correlazione parziale per individuare variabili di processo responsabili di deviazione dalle condizioni operative normali (NOC, *normal operating conditions*) in stazionari o dinamici. Le due metodologie di diagnosi vengono applicate al caso studio di un modello a principi primi, che descrive un processo dinamico di fermentazione per la produzione di penicillina (Birol *et al.*, 2002).

Tali tecniche diagnostiche vengono applicate sia utilizzando un set di dati di processo relativi agli ingressi e ai valori dello stato finale delle variabili, sia un set di dati dinamici, che comprendono le misurazioni nel tempo delle variabili. In tal modo viene verificata la capacità diagnostica delle due metodologie e i vantaggi e le problematiche emergenti dal loro utilizzo per la diagnosi in sistemi dinamici, quale è per l'appunto quello relativo al caso studio trattato.

Entrambe le metodologie sono state sviluppate per sfruttare i dati storici disponibili del processo. In generale i dati storici rappresentano le misure di un processo reale, nell'ambito della diagnosi di PMM di un impianto preesistente, sia esso in scala di laboratorio (verifica del modello) o in scala pilota o industriale (monitoraggio e controllo del processo; problemi di scale up). In questa Tesi, non sono disponibili dati storici, ma i dati utilizzati sono generati *in silico*, con un modello a principi primi in grado di simulare correttamente i valori in uscita dal processo. Questi sono confrontati con le predizioni del modello a principi primi da sottoporre alla diagnosi: esso è lo stesso modello a principi primi utilizzato per generare i dati storici, in cui sono modificate delle equazioni o dei parametri al fine di dare origine a un PMM rilevante (ovvero in grado di provocare un cambiamento della struttura di correlazione del modello a principi primi perturbato rispetto a quello originale).

La Tesi è organizzata in quattro capitoli. Nel Capitolo 1 vengono illustrate le tecniche statistiche multivariate utilizzate per l'applicazione della prima metodologia, e viene presentata la seconda metodologia utilizzata. Nel Capitolo 2 è presentato il caso studio, e il modello a principi primi utilizzato ed implementato in un simulatore, per la generazione dei dati. Nel Capitolo 3 sono presentati i risultati dell'applicazione della prima metodologia, applicata sia a dei set di dati che comprendono i dati degli ingressi al processo e i valori delle uscite nello stato finale, sia a set di dati relativi alla dinamica del processo, con misurazioni delle variabili lungo parte della durata della simulazione del processo. I risultati ottenuti vengono comparati ai risultati generati con la seconda metodologia, riportati nel Capitolo 4, dove vengono proposte delle soluzioni relative all'organizzazione dei set di dati dinamici.

CAPITOLO 1

Richiami di matematica e statistica multivariata

Nella prima parte di questo Capitolo vengono presentate le tecniche statistiche multivariate utilizzate in questa Tesi, in particolare l'analisi a componenti principali e la proiezione su strutture latenti; nella seconda, invece, sono presentate due metodologie per l'identificazione delle possibili cause del disallineamento tra modello e processo (PMM, *process/model mismatch*).

1.1 Modelli a variabili latenti

Data una matrice di dati **X** di dimensione [$N \times K$], in cui N è il numero di osservazioni o campioni disponibili per K variabili misurate, in generale affette da rumore (e.g., variabili di processo, misurazioni delle caratteristiche di un prodotto, etc.), i modelli a variabili latenti permettono di riassumere l'informazione contenuta in **X**, con un numero minore di variabili rispetto al numero di variabili originali, dette variabili latenti, (Eriksson *et al.*, 1999). Tali nuove variabili vengono definite in maniera tale da catturare la maggior parte della variabilità dei dati in analisi. Maggiore è l'informazione riassunta dal nuovo set di variabili latenti, tanto minore sarà il loro numero A rispetto al numero di variabili originali K, di solito correlate tra loro. I modelli a variabili latenti vengono utilizzati sia per descrivere le relazioni tra le variabili di uno stesso set di dati **X** che le relazione tra un set di regressori **X** [$N \times K$] e variabili risposta **Y** (e.g., specifiche di prodotto, variabili in uscita di un processo, etc.) [$N \times M$].

1.1.1 Analisi delle componenti principali (PCA)

L'analisi a componenti principali (PCA, *principal component analysis;* Jackson, 1990), è un metodo statistico multivariato che consiste in una trasformazione lineare delle variabili originali di una matrice di dati **X**, tramite la proiezione delle stesse in un nuovo spazio vettoriale, le cui direzioni ortogonali tra loro, vengono definite in modo da individuare le direzioni di massima variabilità dei dati (Burnam *et al.*, 1996; Lopez-Negrete *et al.*, 2010). Tramite la PCA è possibile investigare in modo rapido la struttura di correlazione esistente tra le variabili originali. Matematicamente, a partire dalla matrice di covarianza **C** di **X** (Burnam *et al.*, 1996; Lopez-Negrete *et al.*, 2010):

$$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}, \tag{1.1}$$

vengono individuati gli autovettori **P** ed autovalori **Λ** ad essa associati. Ogni autovettore **p** di dimensione [K×1] della matrice di covarianza è detto *loading* ed è associato a una PC, di cui rappresenta il coseno dell'angolo formato con l'asse della k-esima variabile originale. Ad ogni autovettore **p** è associato a un autovalore λ tanto più grande quanto maggiore è la variabilità dei dati catturata dalla corrispondente componente principale. Le coordinate delle proiezioni dei dati di **X** sul nuovo spazio latente, sono rappresentate da una matrice **T** detta matrice degli scores. Ne consegue che per ogni direzione latente associata al *loading* p, viene calcolato un vettore **t** di dimensione [N×1] tale che:

$$\mathbf{t} = \mathbf{X}\mathbf{p}. \tag{1.2}$$

Pertanto, il set di dati originale può essere rappresentato dal prodotto scalare degli *scores* con i *loadings*, ovvero:

$$\mathbf{X} = \sum_{\mathbf{i=1}}^{K} \mathbf{t_i}\mathbf{p_i^T} \tag{1.3}$$

Dal momento che le variabili originali sono spesso correlate tra loro, il rango R della matrice **X** è minore di K. Di fatto il contributo di variabili correlate tra loro può essere rappresentato dalla stessa componente principale, sulla quale tali variabili presentano *loading* molto simili a indicare una direzione di variabilità comune. Questo permette di rappresentare il set di dati X con un numero A di variabili latenti < K(e < N). Dalla differenza tra la matrice **X** dei dati originali e la matrice ricostruita tramite un modello PCA, $\hat{\hat{\mathbf{X}}}$, è originata la matrice dei residui **E**, che rappresenta la parte di variabilità dei dati che non viene descritta dal modello. La matrice **E** rappresenta la variabilità non sistematica che non viene catturata dal modello, pertanto, in generale, offre un'indicazione del rumore da cui sono affetti i dati (Eriksson *et al.*, 1999): tanto più rumorose sono le misurazioni di una variabile, tanto più saranno alti, in valore assoluto, i valori degli elementi della matrice dei residui. La matrice **E** è calcolata come:

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \sum_{\mathbf{i=1}}^{\mathbf{A}} \mathbf{t_i}\mathbf{p_i^T} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{t_i}\mathbf{p_i^T} - \sum_{\mathbf{i=1}}^{\mathbf{A}} \mathbf{t_i}\mathbf{p_i^T} \tag{1.4}$$

Quindi:

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1.5}$$

In Figura 1.1, viene fornita una rappresentazione grafica del significato geometrico dell'analisi alle componenti principali, supponendo, per semplicità, di considerare 7 campioni e di prendere in esame due variabili di processo, x_1 e x_2.

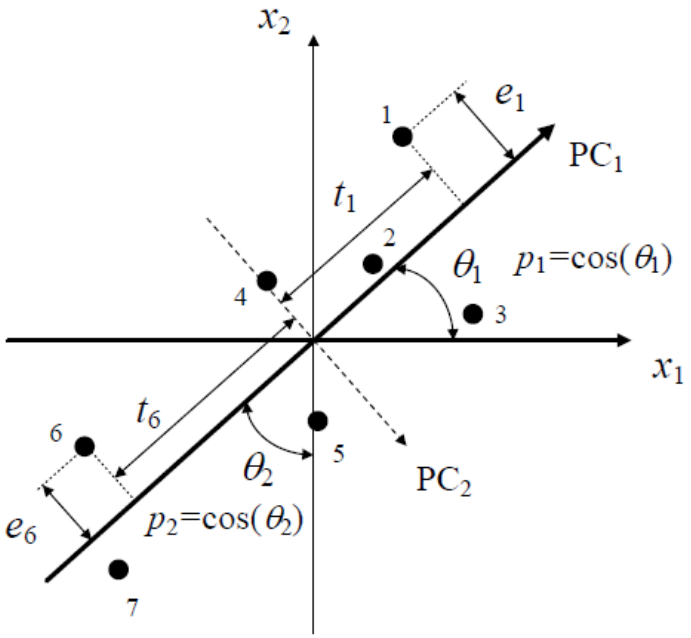


Figura 1.1. *Interpretazione geometrica di scores e loading della PCA di un set di dati con sette campioni e due variabili. Da:Emanuele Tomba, (2013).*

Nella costruzione di un modello PCA, viene identificato il vettore, in questo spazio, parallelo alla direzione di variabilità massima, ovvero PC$_1$. La mancanza di rappresentatività del modello è data dai residui *e* (ad esempio e_1 ed e_6 in Figura 1.1), che rappresentano le distanze delle osservazioni dalla direzione di PC$_1$. Se fosse considerata una seconda componente principale, ortogonale alla prima, questa catturerebbe una parte minore e poco rilevante di variabilità dei dati rispetto alla prima componente principale. Nel caso riportato, una PC è sufficiente a rappresentare gran parte della variabilità dei dati.

1.1.1.1 Pretrattamento dei dati

Si noti che solitamente, prima di costruire il modello PCA, i dati sono sottoposti a un pretrattamento (Wise *et al.*, 2006): ad ogni campione, ovvero l'*n*-esimo elemento nella *k*-esima colonna della matrice di dati **X**, è sottratto il valore medio della colonna corrispondente, tramite il *mean centering* (il bilanciamento al valor medio, ovvero la sottrazione della media); viene puoi eseguito l'*autoscaling* (riduzione della scala), cioè la divisione di ogni elemento della matrice per la deviazione standard della sua colonna. Il bilanciamento al valor medio permette di traslare i dati all'origine del sistema di riferimento (il *K*-spazio iniziale) e la riduzione della scala permette di rendere la variabilità di ciascuna variabile ugualmente importante nella

costruzione del modello PCA. Una volta sottoposta a queste operazioni, la matrice di correlazione ricavata con la (1.1) diventa la matrice dei coefficienti di correlazione.

1.1.1.2 Selezione del numero di PC

Il numero di PC per la costruzione del modello può essere selezionato in base a diversi criteri. In questa Tesi ne sono stati considerati due:

- regola dell'autovalore maggiore di 1 (Mardia *et al.*, 1979);
- convalida incrociata basata sull'indice PRESS (*prediction error sum of squares*, somma quadratica degli errori di predizione; Wold, 1978).

Il primo criterio è una semplice regola per cui tutte le PC i cui corrispondenti autovalori sono minori di 1 non vengono inclusi nel modello. L'idea alla base di questo criterio è che, essendo i dati sottoposti a riduzione di scala, l'autovalore associato a ogni PC si può assumere, approssimativamente, come il numero di variabili la cui variabilità è rappresentata dal componente principale. Dunque, se una PC non rappresenta almeno una variabile, non è necessaria alla costruzione del modello. Sebbene tale criterio sia molto semplice da implementare, è necessario considerare con cautela lo scarto di una PC il cui autovalore è molto vicino a 1, che può portare a non considerare la variabilità da esso spiegata, che potrebbe essere non trascurabile. La convalida incrociata è essenzialmente un metodo di autoverifica per convalidare il numero di PC. L'idea su cui si basa è che si debba selezionare un numero di PC tali che l'errore (rappresentato dalla somma dei quadrati della matrice dei residui) nella ricostruzione di nuovi campioni con il modello sia minimo. Un tipico algoritmo di convalida incrociata prevede di:

1. dividere il set di dati **X** in G sottogruppi $\mathbf{X}_g$ di C campioni;
2. considerare un set di dati ridotto $\mathbf{X}_1$ privato del sottogruppo $\mathbf{X}_g$;
3. costruire un modello PCA su $\mathbf{X}_1$;
4. proiettare $\mathbf{X}_g$ sul modello costruito.

L'errore di ricostruzione per il sottogruppo g è:

$$PRESS_g = \sum_{c=1}^{C}\sum_{k=1}^{K}\left(x_{c,k} - \hat{x}^g{}_{c,k}\right)^2 = \sum_{c=1}^{C}\sum_{k=1}^{K} e_{c,k}{}^2, \tag{1.6}$$

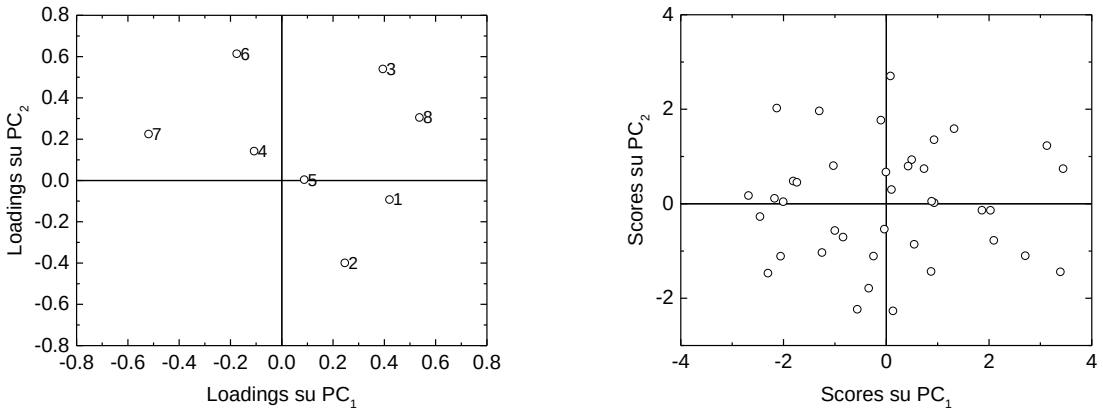
dove $\hat{x}^g{}_{c,k}$ è l'elemento ricostruito di $\mathbf{X}_g$, c-esima riga, k-esima colonna ed $e_{c,k}$ l'errore di ricostruzione corrispondente. Tale procedura sarà ripetuta per tutti i sottogruppi in cui il set di dati verrà suddiviso, fino ad avere:

$$PRESS = \sum_{g=1}^{G} PRESS_g \tag{1.7}$$

Per ogni componente principale aggiunto al modello si potrà avere il corrispondente PRESS, e si selezionerà il numero di componenti cui corrisponda l'errore minimo.

1.1.1.3 Analisi dei risultati

Da un punto di vista pratico, è utile analizzare i risultati di un modello PCA mediante la rappresentazione grafica degli *scores* e dei *loadings* risultanti, come riportato in Figura 1.2, in cui è stato analizzato un set di dati (39 campioni misurati per 8 variabili) di un processo di fermentazione. Infatti, sebbene siano possibili diverse rappresentazioni grafiche (per esempio grafici a barre lungo ogni PC), il confronto dei *loadings* e degli *scores* riportati sul piano formato da una coppia di componenti principali, è quello più utilizzato.



(a) (b)

Figura 1.2. *Diagramma dei loadings (a) e degli scores (b) sulle prime due componenti principali di un modello di fermentazione semplificato.*

Analizzando le prime due PC, l'interpretazione dei due diagrammi è la seguente:

- nel grafico dei *loadings*, (Figura 1.2*a*) ogni punto rappresenta una variabile, e le sue coordinate sono i suoi *loadings* sulle componenti principali. Due punti in uno stesso quadrante rappresentano due variabili correlate positivamente, e la loro correlazione sarà tanto più accentuata quanto più i valori dei *loadings* sono simili, come nel caso delle variabili 3 e 8. Se invece due variabili appartengono a quadranti diversi (i punti giacciono in due semipiani diversi, rispetto a uno dei due assi), le variabili sono correlate negativamente. Considerando la prima PC, le variabili 3 e 5 sono correlate positivamente, ma sulla seconda PC la correlazione, seppur positiva, è estremamente debole, poiché il *loading* della variabile 5 sulla seconda componente principale è quasi nullo;
- nel grafico degli *scores* (figura 1.*2b*) si osservano le posizioni reciproche dei campioni del sistema. La posizione di ogni punto, ovvero di ogni campione, dipende dai valori delle variabili misurate in quel campione. Campioni molto vicini nel piano presentano solitamente caratteristiche simili;

- è altresì possibile ricavare informazioni dal grafico degli *scores* confrontandolo con quello dei *loadings*: per esempio i campioni alla destra del grafico (Figura 1.2*b*), nella posizione simile a quella della variabile 8 nel grafico (Figura 1.2*a*), hanno alti valori della variabile 8 e bassi valori della variabile 3; analogamente i campioni nel quadrante in basso a destra avranno alti valori della variabile 2 e bassi valori della variabile 6, in riferimento alla seconda componente principale.

1.1.1.4 Indici diagnostici della PCA

Esistono diversi indici che permettono di valutare la capacità di un modello PCA di rappresentare il sistema in analisi, di quantificare il suo potere predittivo e di misurare l'influenza che le singole variabili in esame o i singoli campioni che compongono il set di dati hanno sul modello. Per esempio, l'indice R^2 (Eriksson *et* al.,2001) è utilizzato per stimare la variabilità dei dati spiegata dalle componenti principali, ed è calcolato come segue:

$$R^2 = 1 - \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\left(x_{n,k} - \hat{x}_{n,k}\right)^2}{\sum_{n=1}^{N}\sum_{k=1}^{K}\left(x_{n,k}\right)^2} = 1 - \frac{ESS}{TSS}, \tag{1.8}$$

dove $x_{n,k}$ rappresenta l'elemento nella *n*-esima riga, *k*-esima colonna del set di dati ricostruito con il modello PCA. Con *ESS* (*sum of square errors*) e *TSS* (*total sum of squares*) si indicano rispettivamente la somma degli errori quadratici e la somma dei quadrati dei valori originali degli elementi del set di dati. Se si calcola tale indice per un numero di PC crescente, esso aumenta fino ad arrivare a 1 quando il numero *A* di PC è uguale al numero delle *K* variabili misurate. La capacità predittiva del modello si basa sul calcolo dell'indice PRESS, già visto nel caso della convalida incrociata, e si misura tramite l'indice Q^2 (Eriksson *et* al.,2001):

$$Q^2 = 1 - \frac{PRESS}{TSS} \tag{1.9}$$

Vi sono anche degli indici usati per la calibrazione dei dati, per riuscire a scoprire eventuali *outliers* tra i campioni o le variabili che su di esso hanno maggiore influenza. Il T^2 di Hotelling (Hotelling, 1993) è un indice statistico che definisce la distanza dalla proiezione di un'osservazione del set di dati sull' *A*-spazio dei componenti principali dall'origine degli assi nello stesso spazio. Di fatto il T^2 di un'osservazione è un indice della sua *leverage* (effetto leva), cioè della deviazione dei valori delle sue variabili dai valori medi del set di dati. Se un campione presenta un alto valore di T^2 essa può rappresentare un potenziale *outlier*. Il T^2 della *i*-esima osservazione è calcolato come:

$$T_i^2 = \sum_{a=1}^{A} \frac{t^2_{a,i}}{\lambda_a} = \mathbf{t_i}^{\mathbf{T}} \mathbf{\Lambda}^{-1} \mathbf{t_i}^{\mathbf{T}}, \tag{1.10}$$

Dove $\mathbf{t_i}$ è il vettore di dimensioni [A×1] che contiene le proiezioni $t_{a,i}$ della i-esima osservazione sugli A componenti principali su cui è stato costruito il modello, mentre $\mathbf{\Lambda}$ è la matrice quadrata degli autovalori troncata alla dimensione A. L'indice SPE (*squared prediction error;* Mardia *et al.*, 1979) invece, viene utilizzato per valutare quanto un'osservazione viene rappresentata in modo adeguato dal modello:

$$SPE_i = \left(x_i - \hat{x}_i\right)^T \left(x_i - \hat{x}_i\right) = \mathbf{e_i}^{\mathbf{T}} \mathbf{e_i}, \tag{1.11}$$

dove $\mathbf{e}_\mathrm{i}$ è il vettore di dimensioni [N×1] degli scarti nella ricostruzione della i-esima osservazione $\mathbf{x_i}$. L'indice SPE misura la distanza dell'osservazione dai piani latenti definiti dalle componenti principali. Osservazioni con un alto SPE sono scarsamente rappresentate dal modello, ovvero non appartengono alla struttura di correlazione identificata dal modello. Generalmente, campioni con un basso T^2 e un alto SPE sono poco rappresentativi del modello ed eliminandoli il modello resterà pressochè invariato. Sia per il T^2 che l'SPE si possono ricavare i contributi specifici di ogni variabile al valore calcolato per un certo campione, in modo da valutare quale variabile contribuisca maggiormente alla leva delle osservazioni o alla distanza dallo spazio di modello. I contributi delle variabili al T^2 sono calcolati come:

$$\mathbf{t}_{\mathrm{CONT},i} = \left(\mathbf{t}_i^{\mathbf{T}} \mathbf{\Lambda}^{-1/2} \mathbf{P}^{\mathbf{T}}\right)^{\mathbf{T}} \tag{1.12}$$

Ogni $\mathbf{t}_{\mathrm{CONT},i}$ è un vettore di dimensione [K×1] dei contributi delle variabili al valore di T^2 di un'osservazione. La sua norma quadratica corrisponde proprio al T^2 della corrispondente osservazione. Invece il contributo di ogni variabile all'indice $\mathrm{SPE_i}$ di ogni osservazione è l'errore di ricostruzione dell'elemento nella k-esima colonna della i-esima riga, ovvero:

$$SPE_{CONT,i,k} = e_{i,k} \tag{1.13}$$

Una volta costruito un modello PCA su un set di dati che riflettono le condizioni operative normali (NOC, *normal operative conditions*) di un processo in atto, questo può anche venire utilizzato per valutare se un nuovo set di dati, $\mathbf{X}^{\mathrm{NEW}}$, rientri o meno nel campo delle NOC, proiettandolo sullo spazio del modello PCA, e successivamente ricostruendolo tramite il prodotto dei *loadings* del modello per i suoi *scores*:

$$\mathbf{t}^{\mathrm{NEW}} = \mathbf{X}^{\mathrm{NEW}} \mathbf{P} \tag{1.14}$$

$$\hat{\mathbf{X}}^{\mathrm{NEW}} = \mathbf{P}\mathbf{t}^{\mathrm{NEW}} \tag{1.15}$$

Avendo a disposizione gli *scores* del nuovo set di dati e la sua matrice ricostruita, è possibile usare le statistiche T^2 e i valori di SPE, tramite la costruzione di opportuni limiti di confidenza (Jackson, 1991; Qin, 2003) per valutare quanto i dati del nuovo set si discostino dalla media definita dalle NOC, e quanto la struttura di correlazione del nuovo set di dati si adatti al modello PCA costruito su dati in NOC.

1.1.1.5 MPCA

Nel caso in cui sia necessario analizzare le l'evoluzione nel tempo delle relazioni che intercorrono tra diverse variabili, ovvero nel caso in cui si esamini un sistema dinamico, la matrice di dati risultante $\underline{\mathbf{X}}$ è tridimensionale di dimensioni [$N \times K \times T$], dove N è il numero di campioni (spesso il numero di *batch*), K il numero di variabili e T è il numero di istanti di campionamento (o istanti di tempo del processo). In questo caso, viene utilizzata la tecnica chiamata MPCA (Nomikos and MacGregor, 1994). Tale tecnica prevede che l'analisi dei dati venga effettuata previo 'dispiegamento' (*unfolding)* della matrice tridimensionale $\underline{\mathbf{X}}$ per ottenere una matrice bidimensionale $\mathbf{X}$ su cui è poi costruito un modello PCA. L'*unfolding* può essere:

- verticale (*variable-wise unfolding*). Ogni sezione orizzontale di $\underline{\mathbf{X}}$ di dimensioni [$K \times T$] è disposta sotto a quelle precedenti in modo da creare una matrice di dimensione [$NT \times K$]. In questo caso, il numero massimo di componenti principale rimane invariato ed uguale al numero di variabili del set di dati di partenza. In tal modo è possibile osservare una struttura di correlazione mediata nel tempo, mentre si può analizzare l'andamento nel tempo degli *scores*, ovvero osservare come variano nel tempo le proprietà di un singolo campione o batch;
- orizzontale (*batch-wise unfolding*), dove si può analizzare la variazione nel tempo dei *loadings*, ovvero si può osservare la variazione, nel tempo, dell'intera struttura di correlazione del processo. Ciò corrisponde a disporre, affiancate le une alle altre, le sezioni verticali di dimensione [$N \times K$] per definire una matrice $\mathbf{X}$ di dimensioni [$N \times KT$].

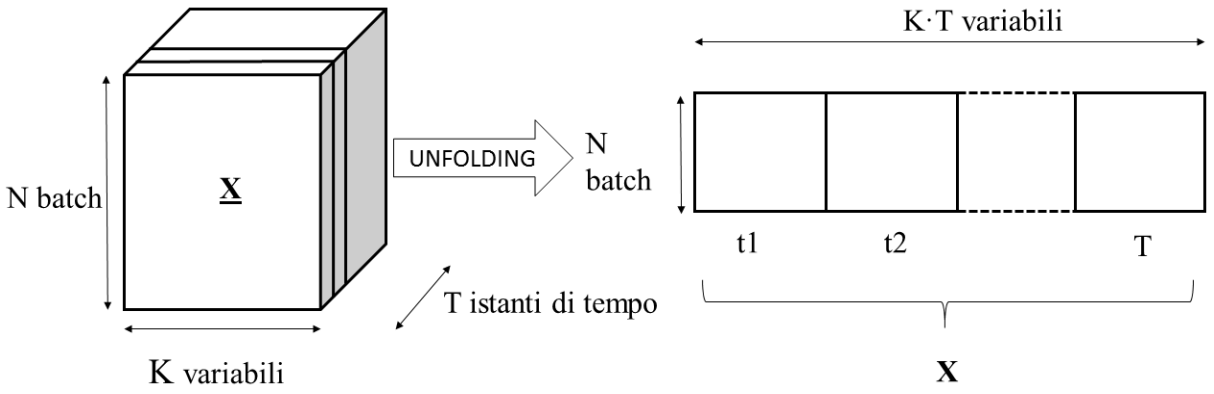


Figura 1.3. *Unfolding orizzontale della matrice tridimensionale* $\underline{\mathbf{X}}$.

In Figura 1.3 viene riportato un esempio grafico di *unfolding* orizzontale. Ognuna delle K matrici di dimensione [N×T] (che ha, per ognuna delle N righe, l'andamento della k-esima variabile nel tempo, misurata nell'N-esimo campione), costituisce una sezione orizzontale della matrice tridimensionale $\underline{\mathbf{X}}$ a sinistra; tali matrici bidimensionali vengono disposte orizzontalmente le une a fianco delle altre.

1.1.2 Proiezione su strutture latenti (PLS)

La PLS (*partial least squares regression;* Wold *et al.*, 1983; Höskuldsson, 1988), è una tecnica di regressione che correla un set di dati di regressori **X** a un set di dati di variabili di risposta **Y**. La PLS effettua una trasformazione dei dati in **X** di dimensione [$N \times K$] in modo da massimizzare la covarianza delle sue variabili latenti con le variabili del set di dati **Y** di dimensione [$N \times M$]. Il modello generato può essere utilizzato per svariati scopi, come la predizione delle variabili in uscita dal processo, oppure la valutazione dell'impatto delle variabili nel set di dati **X** sulle variabili risposta in **Y**. Dopo aver effettuato l'*autoscaling* (§ 1.1.1.1), i set di dati **X** e **Y** sono modellati come (Wold, 1976):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a{}^{\mathrm{T}} + \mathbf{E} \tag{1.16}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}_a{}^{\mathrm{T}} + \mathbf{F} \tag{1.17}$$

I vettori ortogonali degli *scores* $\mathbf{t}_a$ e $\mathbf{u}_a$, sono combinazioni lineari delle variabili in **X**, con coefficienti di combinazione lineare, i pesi (*weights*), tali che $\mathbf{t}_a = \mathbf{X}\mathbf{w}_a$, e delle variabili in **Y**, con $\mathbf{u}_a = \mathbf{Y}\mathbf{q}_a$ (per la matrice **Y** sono detti *loadings*), per ogni a-esima variabile latente. Tali *scores* devono massimizzare la covarianza tra **X** e **Y** e i pesi sono introdotti per mantenere la loro ortogonalità. La relazione che intercorre tra gli *scores* delle due matrici è la seguente:

$$\mathbf{u}_a = b_a \mathbf{t}_a, \tag{1.18}$$

dove b_a è un elemento del vettore dei coefficienti di regressione B.
Tra i vari indici diagnostici disponibili nella regressione tramite la PLS l'indice VIP (*variable influence on projection*; Chong and Jun, 2005) particolare interesse per questa Tesi, poiché viene utilizzato per la selezione delle variabili di un processo da includere in un set di dati. Tale indice quantifica, per ogni variabile risposta nel set di dati **Y**, quali sono le variabili in **X** che su di essa hanno maggior peso per la sua predizione:

$$VIP_j = \sqrt{\frac{m\sum_{a=1}^{A}\left(\mathbf{b}_a{}^2\mathbf{t}_a{}^{\mathrm{T}}\mathbf{t}_a\right)\left(\mathbf{w}_{j,a}/\left\|\mathbf{w}_{j,a}\right\|\right)^2}{\sum_{a=1}^{A}\left(\mathbf{b}_a{}^2\mathbf{t}_a{}^{\mathrm{T}}\mathbf{t}_a\right)}} \tag{1.19}$$

Dove *m* è il numero di variabili della matrice **X** e *A* è il numero di variabili latenti scelto.

1.2 Tecniche di diagnosi del PMM

In questa Tesi sono considerate due metodologie differenti per la diagnosi del PMM . La prima è stata elaborata da Meneghetti *et al.* (2014) ed utilizza un indice diagnostico basato sull'analisi dei residui di un modello PCA. Il secondo metodo di diagnosi è ricavato dagli studi di Rato e Reis (2015) sull'uso dei coefficienti di correlazione parziale per il monitoraggio di processi continui. Le due metodologie sono state testate considerando diversi tipi di disallineamento tra processo e modello (PMM, *process/model mismatch*) che possono presentarsi nel caso di trasferimento di processo tra diverse scale. Il sistema in esame è un processo di fermentazione di penicillina, descritto nel Capitolo seguente, i cui dati simulati vengono raccolti in due matrici di processo e di modello, rispettivamente $\mathbf{X}_{\text{dati},\Pi}$ e $\mathbf{X}_{\text{dati,M.}}$

1.2.1 Analisi dell'MRLR

Questa metodologia si basa sull'analisi di indici ricavati dalla matrice dei residui della matrice di processo (ovvero le misure storiche del processo), proiettata su di un modello PCA che è stato costruito sulla matrice di modello (le predizioni delle risposte del processo da parte del modello a principi primi). Ci si aspetta che i dati di processo non vengano descritti in modo ottimale dal modello PCA, calibrato sui dati di modello.

1.2.1.1 Procedura utilizzata

La metodologia proposta da Meneghetti *et al.* (2014) consiste nei seguenti passaggi, dove ci si riferisce con i pedici Π e M al processo e al modello, rispettivamente:

1. creazione delle matrici di modello e processo. Utilizzando lo stesso set di valori iniziali misurati per gli *N* campioni che costituiscono il set dei dati storici disponibile, viene generato un set di dati simulati utilizzando il modello a principi primi considerato. Ci si riferisce all'insieme degli ingressi misurati e dei risultati simulati come "misure simulate", mentre le misurazioni in ingresso e uscita dal processo vengono definite come "misure storiche". Una volta generati i dati, le variabili e i parametri del modello sono combinati per ottenere un numero di *V* variabili ausiliarie a partire dalle *K* variabili originali, i base alla struttura del modello a principi primi considerato. I valori di tali variabili ausiliarie sono

calcolati utilizzando le misure storiche del processo e le misure simulate dal modello per formare rispettivamente una matrice di modello $\mathbf{X}_M$ e di processo $\mathbf{X}_\Pi$ di dimensioni [$N \times V$]. Per rilevare un PMM (strutturale, derivante da un'errata formulazione del modello, o parametrico, causato da un valore errato dei parametri), occorre che ogni variabile ausiliaria includa almeno una variabile in ingresso o in uscita, e non solamente dei parametri (a meno che questi non siano a loro volta funzione delle variabili o delle condizioni di entrata o di processo).. A causa della presenza di un PMM, ci si aspetta che la struttura di correlazione delle due matrici sia differente. Questa differenza viene analizzata per effettuare la diagnosi del modello;

2. Creazione di un modello PCA per la matrice di modello. Entrambe le matrici sono scalate rispetto alla media e deviazione standard di $\mathbf{X}_M$. Viene costruito un modello PCA su $\mathbf{X}_M$ e viene calcolata la matrice dei residui di modello $\mathbf{E}_M$:

$$\hat{\mathbf{X}}_M = \mathbf{T}_M \mathbf{P}_M^T \tag{1.20}$$

$$\mathbf{E}_M = \mathbf{X}_M - \hat{\mathbf{X}}_M \tag{1.21}$$

3. Proiezione della matrice di processo sul modello PCA. $\mathbf{X}_\Pi$ è proiettata nello spazio latente del modello PCA costruito su $\mathbf{X}_M$, in modo da poter calcolare la matrice dei residui di processo $\mathbf{E}_\Pi$:

$$\mathbf{T}_\Pi = \mathbf{X}_\Pi \mathbf{P}_M \tag{1.22}$$

$$\hat{\mathbf{X}}_\Pi = \mathbf{T}_\Pi \mathbf{P}_M^T \tag{1.23}$$

$$\mathbf{E}_\Pi = \mathbf{X}_\Pi - \hat{\mathbf{X}}_\Pi \tag{1.24}$$

4. Analisi delle matrici dei residui e diagnosi del PMM. Le due matrici dei residui, $\mathbf{E}_\Pi$ ed $\mathbf{E}_M$, vengono analizzate per determinare quali sono le variabili ausiliarie che contribuiscono maggiormente alla diversa struttura di correlazione tra le matrici $\mathbf{X}_\Pi$ e $\mathbf{X}_M$ e, di conseguenza, quali soni i parametri o le equazioni del modello che contribuiscono a creare il PMM.

Come si vede dalla (1.4), la matrice dei residui riflette la parte della variabilità dei dati che non viene catturata dal modello. Se gli elementi $e_{i,v}$ di $\mathbf{e}_v$, dove v è la v-esima colonna di $\mathbf{E}_M$, sono distribuiti normalmente, è possibile definire dei limiti di confidenza all'interno dei quali rientra la parte di variabilità, non rilevata dal modello, di un set di dati con una struttura di correlazione simile a $\mathbf{X}_M$ (Montgomery, 2005b):

$$CL_{\alpha,\mathbf{e}_v} = z_{\alpha/2} \cdot \sigma(\mathbf{e}_v), \tag{1.25}$$

dove α è la percentuale di confidenza, generalmente compresa tra 0.01 e 0.05 (corrispondenti al 99% e 95% di confidenza), $z_{\alpha/2}$ è il corrispondente valore della statistica z e $\boldsymbol{\sigma}(\mathbf{e}_v)$ è la deviazione standard della *v*-esima colonna di $\mathbf{E}_M$. È importante notare che la matrice $\mathbf{E}_\Pi$ rappresenta la parte di variabilità del set di dati di processo che non è descritta dal modello PCA costruito su $\mathbf{X}_M$, ma il valore dei suoi elementi dipende anche dalla discrepanza tra i dati storici e quelli simulati dal modello: per tenere conto solo di quest'ultima componente, si rimuove il contributo legato alla variabilità di $\mathbf{X}_\Pi$ che non viene identificata dal modello. Si ottiene quindi, per ogni variabile, il seguente indice diagnostico, il rapporto medio tra residui e limiti di confidenza (*mean residual-to-limit ratio, MRLR*):

$$MRLR_v = \frac{\sum_{n=1}^{N}\left(\frac{\sqrt{\left(e_{\Pi n,v}\right)^2}}{CL_{\alpha,\mathbf{e}_v}}\right)}{N} \tag{1.26}$$

Per poter utilizzare tale indice ci si deve assicurare che i residui sulle colonne di $\mathbf{E}_\Pi$ siano distribuito normalmente.

1.2.2 Diagnosi del PMM tramite analisi dei coefficienti di correlazione parziale

Sebbene l'analisi dei residui di un modello PCA possa individuare un cambiamento nella struttura delle matrici analizzate, in alcuni casi potrebbe non riuscire a identificare eventuali cambiamenti localizzati della struttura di correlazione, specialmente quando le variabili in analisi sono altamente correlate tra loro. Questo perché la PCA definisce la struttura di correlazione primaria tra le variabili, cioè facendo uso di coefficienti di correlazione di grado 0, calcolati come:

$$r_{x,y} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\cdot\mathrm{var}(y)}}, \tag{1.27}$$

usati per definire lo spazio del modello PCA; una metodologia basata sulla costruzione di un modello che sfrutta le informazioni di correlazione primaria tra le variabili potrebbe non essere in grado, quando si verifica un PMM, essere in grado di individuare cambiamenti sottili e localizzati della struttura di correlazione (Rato e Reis, 2015a). Per questo motivo è stata proposta una metodologia alternativa, basata sull'utilizzo dei coefficienti di correlazione parziale (PCC, *partial correlation coefficient*) per analizzare la correlazione tra due variabili rimuovendo una terza variabile (Rato e Reis, 2014a). Considerando tre variabili *x*, *y* e *z*, il coefficiente di correlazione parziale di grado 1 tra *x* e *y* viene calcolato tenendo *z* sotto controllo, ovvero rimuovendo il suo effetto nel determinare la correlazione tra le prime due variabili (Rato

e Reis, 2014a). Rato e Reis (2014a, 2015b) hanno utilizzato il calcolo dei coefficienti di correlazione parziale in diversi casi studio di sistemi continui in problemi di monitoraggio di processo, previa applicazione di trasformazioni per l'aumento di sensibilità (*sensitivity enhancing transformations,* SET) in modo da massimizzare la variazione dei coefficienti di correlazione parziale in seguito a un cambiamento della struttura di correlazione del sistema (Rato Reis, 2014a). La logica alla base dell'uso dei coefficienti di correlazione parziale a scopo di rilevamento del PMM è che, se si verifica un cambiamento nella struttura di correlazione, i coefficienti di correlazione, in cui le variabili che causano il PMM compaiono come associate, presentino una variazione, mentre i coefficienti di correlazione in cui tali variabili sono controllate mantengano i valori che assumono in condizioni normali (Rato e Reis, 2015a).

1.2.2.1 Organizzazione dei dati

La procedura di diagnosi del PMM, adattata da quella proposta da Rato e Reis (2015a), richiede di adattare il set di dati per poter calcolare i coefficienti di correlazione parziale. Per un set di dati **X**, che raccoglie *N* campioni di *K* variabili (che devono essere analizzate per stabilire quali sono maggiormente legate alla presenza di un PMM), è possibile calcolare solo un vettore di coefficienti di correlazione parziale. Nell'ambito di un metodo di controllo statistico, per ogni coefficiente parziale si deve disporre di un vettore di elementi, dunque per ognuno degli *N* campioni del set di dati vengono simulate *B* osservazioni che si differenziano per del rumore casuale. Per ognuno degli *N* campioni i valori delle variabili in uscita sono generati simulando il processo con un differente set di condizioni iniziali. Al fine di identificare il PMM, le variabili del set di dati sono variabili ausiliarie, ovvero combinazioni di una o più variabili del processo con almeno un parametro del modello che si vuole indagare.

1.2.2.2 Definizione della procedura utilizzata

La procedura usata per la diagnosi del PMM è stata adattata da quella definita da Rato e Reis (2015) e si articola come segue:

1. definizione delle matrici di processo e modello. Viene definito un set di *V* variabili ausiliarie. *N* misurazioni di ciascuna variabile, ripetute *B* volte ciascuna, sono raccolte in un set di dati di processo, la matrice $\underline{\mathbf{X}}_{\Pi}$ di dimensione [$N \times V \times B$].A partire dalle condizioni iniziali di ogni campione vengono ottenuti *N* campioni di *V* variabili ausiliarie predette, simulati ciascuno *B* volte, in modo da definire la matrice di modello $\underline{\mathbf{X}}_{M}$ di dimensioni uguali a quella di processo;
2. calcolo dei coefficienti di correlazione parziale. Per ogni campione $\mathbf{X}_{M}$ di dimensioni [$V \times B$] della matrice di modello sono calcolati i coefficienti di correlazione parziale considerando tutte le combinazioni di variabili associate e controllate. Considerando tre variabili *i, j* e *k,* il coefficiente di correlazione parziale tra *i* e *j* calcolato controllando *k* è:

$$r_{i,j.k} = \frac{r_{i,j} - r_{i,k} \cdot r_{j,k}}{\sqrt{\left(1 - r_{i,k}\right)^2 \left(1 - r_{j,k}\right)^2}}, \tag{1.28}$$

dove $r_{i,j}$, $r_{i,k}$ e $r_{j,k}$ sono calcolati secondo la (1.27).

3. normalizzazione dei coefficienti di correlazione. I coefficienti di correlazione parziale sono normalizzati rispetto al valore medio di popolazione ρ del coefficiente di correlazione (la media di ciascuna distribuzione degli N coefficienti di correlazione):

$$w_{i,j.k} = \frac{\sqrt{N - q - 1} \cdot \left(r_{i,j.k} - \rho\right)}{1 - \rho^2}, \tag{1.29}$$

in cui q è il grado del coefficiente di correlazione parziale (in questo caso 1);

4. calcolo dei coefficienti di correlazione parziale per la matrice di processo. I passaggi 2 e 3 sono ripetuti per la matrice di processo $\underline{\mathbf{X}}_{\Pi}$, ma effettuando la normalizzazione rispetto al vettore media ρ dei coefficienti di correlazione parziale di $\underline{\mathbf{X}}_{M}$. Un vettore di coefficienti di correlazione risulta quindi di dimensione $[1 \times V \cdot (V\text{-}1) \cdot (V\text{-}2)/2]$.
5. definizione dei limiti di confidenza per i coefficienti di correlazione. Sotto l'ipotesi che i coefficienti di correlazione parziale, ricavati dalla matrice di modello, siano distribuiti normalmente, vengono definiti dei limiti di confidenza, per ogni coefficiente normalizzato in base alla percentuale di confidenza α:

$$CL_{i,j.k}, = \sigma(w_{i,j.k}) \cdot z_{\alpha/2} \tag{1.30}$$

6. calcolo della matrice di diagnosi D. Per ognuno dei campioni, viene calcolata una matrice **D** di dimensioni $[V \times V]$, rappresentata in Figura 1.4. Ogni riga della matrice corrisponde a una variabile controllata; il j-esimo elemento della k-esima riga è calcolato come:

$$d_{k,j} = \sum_{k \neq i, k \neq j} f\left(w_{i,j.k}\right), \tag{1.31}$$

Dove $f(w_{i,j.k}) = 1$ se $|w_{i,j.k}| > CL_{i,j.k}$, e 0 altrimenti.

7. assegnazione dei codici di rilevanza alle variabili. Per ogni riga della matrice **D**, è calcolata la norma quadratica, definita *distanza di controllo*; la norma quadratica di ogni colonna, invece, è definita *distanza di coppia*. Una variabile con una bassa distanza di controllo è maggiormente associata al PMM in quanto i coefficienti di correlazione parziale in cui essa e controllata tendono ad avere valori simili a quelli nelle loro condizioni normali, poiché il suo contributo alla correlazione con altre variabili viene eliminato. Analogamente, una variabile con un'alta distanza di coppia è legata al PMM poiché i coefficienti di correlazione parziale in cui essa compare come variabile associata subiscono delle variazioni. A ogni variabile viene quindi assegnata una classe, rappresentabile da tre diversi indici:
   i. Rosso, se la variabile ha la minima distanza di controllo e la massima distanza di coppia;
   ii. Arancione, se ha la minima distanza di controllo ma un valore di distanza di coppia minore del massimo;

iii. Giallo, se ha massima distanza di coppia ma valore di distanza di controllo superiore al valore minimo.

In seguito, dopo aver ripetuto la procedura per ogni campione, si potrà definire la frequenza, rispetto al totale di campioni *N*, con cui ogni variabile viene etichettata con codice rosso, arancione o giallo.

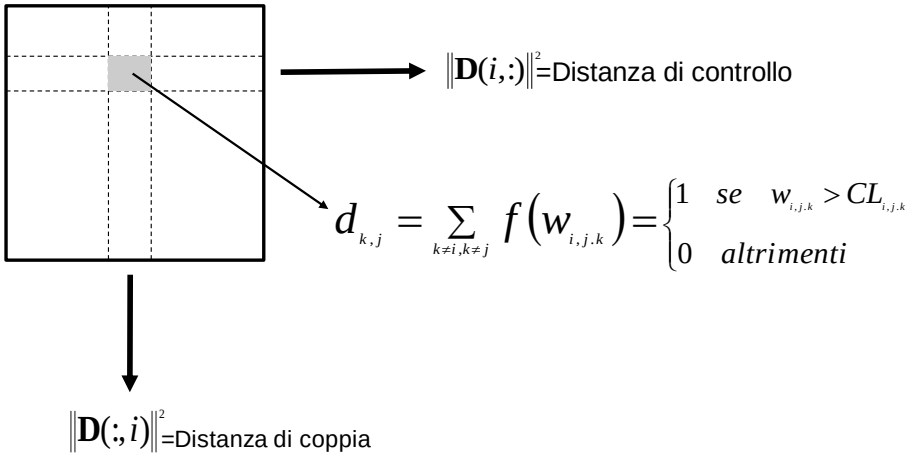


Figura 1.4. *Rappresentazione della matrice di diagnosi calcolata per ciascuno degli N campioni.*

Tabella 1.1. *Regole di assegnazione delle etichette di diagnosi (Rato e Reis, 2015).*

	$\|\mathbf{D}(i,:)\|^2 = \min\left(\|\mathbf{D}(i,:)\|^2\right)$	$\|\mathbf{D}(:,i)\|^2 = \max\left(\|\mathbf{D}(:,i)\|^2\right)$
ROSSO	si	si
ARANCIONE	si	no
GIALLO	no	si

CAPITOLO 2

Caso studio: un modello di fermentazione

In questo Capitolo viene presentato il modello a principi primi analizzato in questa Tesi, sviluppato per la descrizione di un processo di fermentazione per la produzione di penicillina. Infine viene illustrata la procedura per generare i dati su cui applicare le procedure di diagnosi del PMM (*process/model mismatch*).

2.1 Caso studio

Il caso studio in esame riguarda un processo di fermentazione per la produzione di penicillina. Le penicilline si possono considerare come prodotti secondari del metabolismo di alcuni microorganismi. La loro produzione viene effettuata in due fasi. La prima fase (batch) prevede la crescita della biomassa partire da una certa concentrazione iniziale di substrato (glucosio), in presenza di ossigeno. Quest'ultimo viene alimentato con portata costante durante tutto il processo. In tale fase la biomassa sfrutta i nutrienti fino al raggiungimento di una valore di soglia minimo di substrato, per la propria crescita e per il mantenimento e non per la produzione di penicillina. Nella seconda fase, in cui il processo viene esercitato in modalità fed-batch, viene alimentata una portata costante di substrato per il mantenimento della biomassa (la cui velocità di crescita diminuisce fortemente) e per produzione di penicillina.
Il processo dispone di un controllo della temperatura e del pH a valori ottimali tali variabili per la crescita della biomassa. Infatti, entrambe le fasi coinvolte nel processo sono complessivamente esotermiche, ovvero portano all'innalzamento della temperatura, e l'ambiente di reazione tende a diventare più acido nel tempo. Tali condizioni inibiscono la crescita della biomassa e di conseguenza causano la diminuzione della produttività.
In Figura 2.1 viene riportato il diagramma di flusso del processo. Al fermentatore, provvisto di un agitatore e incamiciato per il controllo della temperatura, viene alimentata una portata di substrato a partire dall'inizio della modalità operativa fed-batch. La temperatura e il pH sono continuamente monitorati da sensori, che trasmettono le misurazioni sotto forma di segnali ai regolatori, i quali esercitano la funzione correttiva sulle valvole che regolano le portate di acido, base, acqua di raffreddamento e riscaldamento. Nell'arco dell'intero processo una portata di aria è insufflata al reattore per ossigenare il mezzo di coltura.

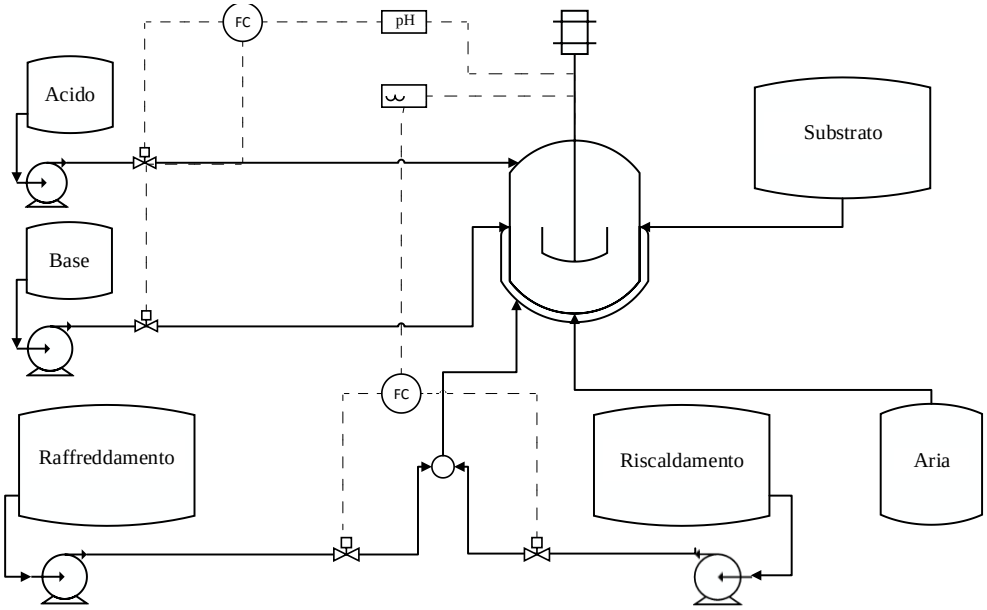


Figura 2.1. *Diagramma di flusso per il processo di produzione della penicillina. Adattato da: Process modelling, monitoring and control Research Group, Illinois Institute of Technology (2000).*

2.1.1 Equazioni e parametri del modello

In questa Tesi è stato i considerato il modello a principi primi sviluppato da Birol *et al.* (2002) per rappresentare il processo di produzione di penicillina (§2.1) descritto. Il modello consta di un set di equazioni differenziali e algebriche (DAE) che descrivono la variazione, nel tempo, della concentrazione delle specie principali coinvolte nel processo, ovvero ossigeno, biomassa, substrato (glucosio) e prodotto (penicillina), oltre alle equazioni differenziali per variazione di volume, pH, temperatura, concentrazione di CO_2, e calore di reazione. Il modello inoltre incorpora ulteriori variabili di input quali la portata di aria e la potenza di agitazione.

Il modello cinetico, utilizzato dagli autori per descrivere la formazione della biomassa e la conseguente produzione di penicillina, è stato adattato da quello proposto da Bajpai e Reuss (1980) ed è detto *unstructured* (non strutturato; Birol *et al.*, 2002) in quanto non tiene conto della fisiologia del microorganismo che costituisce la biomassa. Nei modelli non strutturati le informazioni fisiologiche del microorganismo vengono riassunte in un termine di biomassa, intesa come semplice specie chimica, semplificando notevolmente la cinetica e trascurando fenomeni di trasporto di specie chimiche tra la biomassa e il mezzo di coltura.

Alcuni parametri utilizzati nel modello di Birol *et al.* (2002) sono ricavati dallo studio di Pirt e Righelato (1967), riferiti a un sistema di 2 L di volume e ripresi da Birol *et al*.

In particolare, per descrivere la crescita della biomassa, viene utilizzata un'equazione differenziale che incorpora un termine di cinetica di Monod per evidenziare l'inibizione della crescita ad alte concentrazioni della biomassa stessa:

$$\frac{\mathrm{d}C_x}{\mathrm{d}t} = \mu \cdot C_x - \frac{C_x}{V}\frac{\mathrm{d}V}{\mathrm{d}T}, \tag{2.1}$$

dove la velocità specifica di crescita μ è espressa a sua volta in funzione dei termini di inibizione (funzioni della concentrazione di substrato, C_s, ossigeno, C_L, biomassa C_x e pH) e della temperatura T, la quale è presente in due fattori di Arrhenius di crescita e morte della biomassa:

$$\mu = \mu_x \frac{C_s}{(K_x C_x + C_s)} \frac{C_L}{(K_{OX} C_x + C_L)} \left[\frac{1}{1 + K_1/H_+ + H_+/K_2} \right] k_g \exp\left(-\frac{E_g}{RT} \right) - k_d \exp\left(-\frac{E_d}{RT} \right) \tag{2.2}$$

La variazione del pH, in termini di variazione della concentrazione di protoni, viene espressa come:

$$\frac{\mathrm{d}H_+}{\mathrm{d}t} = \gamma \left(\mu \cdot C_x - \frac{F \cdot C_x}{V} \right) + \left[\frac{-B + \sqrt{B^2 + 4 \times 10^{-14}}}{2} - H_+ \right] \frac{1}{\Delta t}, \tag{2.3}$$

dove γ è un costante di proporzionalità pari a 10^{-5} mol(H_+)/g di biomassa e il termine B è definito come:

$$B = \frac{[10^{-14}/H_+ - H_+]V - C_{a/b}(F_a + F_b)\Delta t}{V + (F_a + F_b)\Delta t} \tag{2.4}$$

F_a e F_b sono le portate di acido e base, $C_{a/b}$ è la concentrazione di entrambe le portate uguale a 3M e Δt è l'intervallo di campionamento. Il pH è mantenuto costante a 5.0. La produzione di penicillina è espressa da un'equazione differenziale in cui il primo termine rappresenta la produzione in funzione della concentrazione di biomassa, mentre il secondo termine rappresenta la degradazione del prodotto in funzione della sua stessa concentrazione, dove compare la costante di velocità di idrolisi K, già presente nell'espressione della cinetica di Bajpai e Reuss:

$$\frac{\mathrm{d}C_p}{\mathrm{d}t} = \mu_{pp} \cdot C_x - K \cdot C_p - \frac{C_p}{V}\frac{\mathrm{d}V}{\mathrm{d}t} \tag{2.5}$$

Nella (2.5) μ_{pp} è la velocità specifica di produzione di penicillina, definita come:

$$\mu_{pp} = \mu_p \frac{C_s}{\left(K_p + C_s + C_s^{\,2}/K_I\right)} \frac{C_L^{\,p}}{\left(K_{OP} \cdot C_x + C_L^{\,p}\right)} \tag{2.6}$$

La variazione, nel tempo, della concentrazione di substrato è descritta considerando, nell'ordine, i contributi relativi al consumo di substrato da parte della biomassa sia per la sua crescita che per la produzione di penicillina e per il proprio mantenimento, oltre al contributo dovuto all'alimentazione del substrato con portata di alimentazione *F*:

$$\frac{\mathrm{d}C_s}{\mathrm{d}t} = -\frac{\mu}{Y_{x/s}}C_x - \frac{\mu_{pp}}{Y_{p/s}}C_x - m_x C_x + \frac{Fs_f}{V} - \frac{C_s}{V}\frac{\mathrm{d}V}{\mathrm{d}t} \tag{2.7}$$

L'equazione della variazione nel tempo dell'ossigeno (2.8) segue la stessa logica, con la presenza di un termine di consumo dell'ossigeno per la crescita della biomassa, uno di consumo specifico per la produzione di penicillina e uno per il mantenimento della biomassa, oltre a un termine di alimentazione attraverso il trasporto nella fase liquida, che incorpora il coefficiente volumetrico di trasporto di massa dell'ossigeno, $K_l a$:

$$\frac{\mathrm{d}C_L}{\mathrm{d}t} = -\frac{\mu}{Y_{x/o}}C_x - \frac{\mu_{pp}}{Y_{p/o}}C_x - m_o C_x + K_l a\left(C_L^{\,*} - C_L\right) - \frac{C_L}{V}\frac{\mathrm{d}V}{\mathrm{d}t} \tag{2.8}$$

La formulazione di $K_l a$ è stata proposta da Bailey e Ollis (1986) e vede il parametro di trasporto dell'ossigeno dipendente dalla portata di aria, dalla potenza di agitazione e dal volume:

$$K_l a = \alpha\sqrt{f_g}\left(\frac{P_W}{V}\right)^{\beta} \tag{2.9}$$

Nella (2.9) i valori dei parametri α e β sono stati assegnati in modo che la dipendenza della concentrazione di penicillina da $K_l a$ fosse conforme ai risultati predetti con il modello di Bajpai e Reuss. La variazione di volume, che tiene conto delle portate in entrata di substrato (*F*) e di acido/base per il controllo del pH ($F_{a/b}$), e della portata in uscita per evaporazione (F_{loss}), viene espressa come:

$$\frac{dV}{dt} = F + F_{a/b} - F_{loss}, \tag{2.10}$$

La variazione di volume dovuta all'evaporazione, viene espressa a sua volta in funzione della temperatura dell'alimentazione di substrato e di quella all'interno del reattore:

$$F_{loss} = V\lambda\left(\exp\left(5(T - T_0)/T_v - T_0\right) - 1\right), \tag{2.11}$$

dove T_0 e T_v sono le temperature di ebollizione e di fusione del mezzo di coltura e λ è stimato come portata di evaporazione di 2.5×10^{-4} L/h. La variazione nel tempo del calore di reazione è

ascrivibile a due termini, uno di crescita della biomassa e uno di mantenimento anche in assenza di crescita:

$$\frac{\mathrm{d}Q_{rxn}}{\mathrm{d}t} = r_{q1}\frac{\mathrm{d}C_x}{\mathrm{d}t}V + r_{q2}\cdot C_x \cdot V, \tag{2.12}$$

dove i parametri r_q sono assunti come costanti di resa (Nielsen e Villadsen, 1994). Infine, dal bilancio di energia basato su uno scambiatore di calore a serpentina, disponibile in scala di laboratorio (Nielsen, 1997), si ricava l'equazione differenziale per definire la variazione di temperatura nel tempo:

$$\frac{dT}{dt} = \frac{F}{s_f}\left(T_f - T\right) + \frac{1}{V\rho c_p}\left[Q_{rxn} - \frac{aF_c^{\,b+1}}{F_c + \left(aF_c^{\,b}/2\rho_c c_{pc}\right)}\right] \tag{2.13}$$

2.1.2 Simulazione del processo e risultati

Per generare i dati analizzati in questa Tesi, è stato utilizzato il simulatore Pensim (di cui è disponibile una versione compatibile con Matlab®) che fornisce una soluzione numerica delle equazioni differenziali e algebriche che compongono il modello a principi primi di Birol *et al.* (2002).

2.1.2.1 Caratteristiche del simulatore

Il simulatore permette la scelta di un set di condizioni iniziali, ovvero concentrazioni di substrato, C_s, di biomassa, C_x, di ossigeno disciolto, C_L, di penicillina, C_p e il volume al tempo iniziale, che costituiscono le condizioni al contorno per risolvere le equazioni e rappresentano gli ingressi del processo.

Inoltre, possono essere modificati anche i valori di alcune variabili operative, quali la portata di aria, la potenza di agitazione, la portata di alimentazione di substrato e la sua temperatura, i valori di *set point* della temperatura e del pH per i sistemi di regolazione.

Infatti, nel simulatore sono implementati dei sistemi di regolazione *feedback* di temperatura e pH, in modo da mantenere i valori di set point desiderati tramite la regolazione di due correnti (acida e basica) per il controllo del pH e due (acqua di riscaldamento e raffreddamento) per quello della temperatura. Per tutte le correnti la regolazione è di tipo PID. Nella Tabella 3.1 sono riportati i valori delle condizioni iniziali, i parametri del modello (Equazioni 2.1-2.13) e quelli dei sistemi di regolazione.

Tabella2.1 *Valori delle condizioni iniziali delle concentrazioni di substrato, ossigeno, biomassa, prodotto; volume iniziale; parametri cinetici e costanti di resa; parametri dei controllori PID.*

Variabile/parametro	Valore
Condizioni iniziali	
Concentrazione di substrato S (g/L)	15
Concentrazione di ossigeno disciolto C_L (=C_L^* alla saturazione) (g/L)	1.16
Concentrazione di biomassa X (g/L)	0.1
Concentrazione di penicillina P (g/L)	0
Concentrazione di ioni idrogeno H_+ (mol/L)	$10^{-5.1}$
Temperatura T (K)	298
Parametri cinetici e variabili	
Concentrazione di substrato nell'alimentazione s_f (g/L)	600
Costante di resa $Y_{x/s}$ (g biomassa/g substrato)	0.45
Costante di resa $Y_{x/o}$ (g biomassa/g ossigeno)	0.04
Costante di resa $Y_{p/s}$ (g penicillina/g substrato)	0.90
Costante di resa $Y_{p/o}$ (g penicillina/g ossigeno)	0.20
Costanti K_1, K_2 (mol/L)	10^{-10}, 7×10^{-5}
Coefficiente di mantenimento (substrato) m_x (h^{-1})	0.014
Coefficiente di mantenimento (ossigneo) m_o (h^{-1})	0.467
Massima velocità specifica di crescita della biomassa μ_x (h^{-1})	0.092
Costante di saturazione di Contois K_x (g/L)	0.15
Costanti di limitazione dell'ossigeno K_{ox}, K_{op} (in condizioni di saturazione)	2×10^{-2}, 5×10^{-4}
Velocità specifica di produzione di penicillina μ_p (h^{-1})	0.005
Costante di inibizione K_P (g/L)	0.0002
Costante p	3
Costante di idrolisi della penicillina K (h^{-1})	0.04
Costante di Arrhenius per la crescita di biomassa k_g	7×10^3
Energia di attivazione per la crescita E_g (cal/mol)	5100
Costante di Arrhenius per la morte della biomassa k_d	10^{33}
Energia di attivazione per la morte E_d (cal/mol)	50000
Densità×calore specifico del mezzo di coltura ρC_p ($°C^{-1}$)	1/1500
Densità×calore specifico del liquido di raffreddamento ρC_{pc} ($°C^{-1}$)	1/1200
Resa di generazione del calore r_{q1} (cal/g biomassa)	60
Costante di generazione del calore r_{q2} (cal/g biomassa h)	1.6783×10^{-4}
Coefficiente di trasporto del calore del liquido di servizio a (cal/h°C)	1000
Costante b	0.60
Costanti nel parametro $K_l a$ α,β	70, 0.4
Costante in F_{loss} λ (h^{-1})	2.5×10^{-4}
Costante di inibizione per la formazione di prodotto K_I (g/L)	0.10
Costante di proporzionalità γ (mol H_+/g biomassa)	10^{-5}
Parametri di controllo (PID)	
pH (base) K_c, τ_I (h), τ_D (h)	8×10^{-4}, 4.2, 0.2625
pH (acido) K_c, τ_I (h), τ_D (h)	1×10^{-4}, 8.4, 0.125
Temperatura (raffreddamento) K_c, τ_I (h), τ_D (h)	70, 0.5, 1.6
Temperatura (riscaldamento) K_c, τ_I (h), τ_D (h)	5, 0.8, 0.005

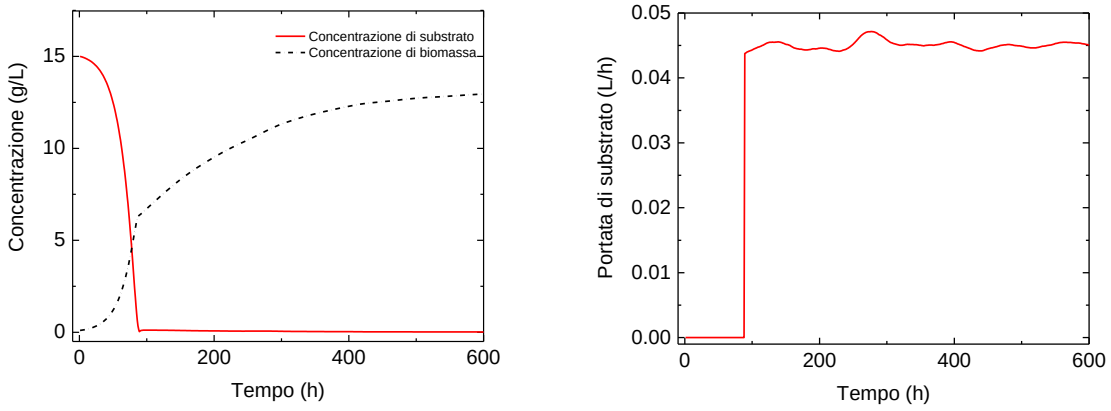
Nel simulatore, allo scopo di riprodurre delle condizioni di processo verosimili, vengono introdotte delle deviazioni dai valori nominali delle variabili operative, definite come PRBS (*pseudo random binary signals*), che producono piccole fluttuazioni nei profili delle concentrazioni; del rumore bianco viene aggiunto anche al profilo di concentrazione dell'ossigeno disciolto, alla portata di alimentazione di substrato e alla potenza di agitazione. Gli intervalli suggeriti dagli sviluppatori del simulatore per alcune variabili in ingresso sono riportati nella Tabella 2.2.

Tabella 2.2 *Intervalli di validità delle variabili in ingresso.*

Variabile	**Valore iniziale**
Volume, V	100-200 L
Portata di aria, f_g	3-10 L/h
Potenza di agitazione, P_w	20-50 W
Portata di substrato, F	0.035-0.045 L/h
Temperatura di alimentazione del substrato, T_i	296-298 K

2.1.2.2 Risultati ottenuti nella simulazione del processo

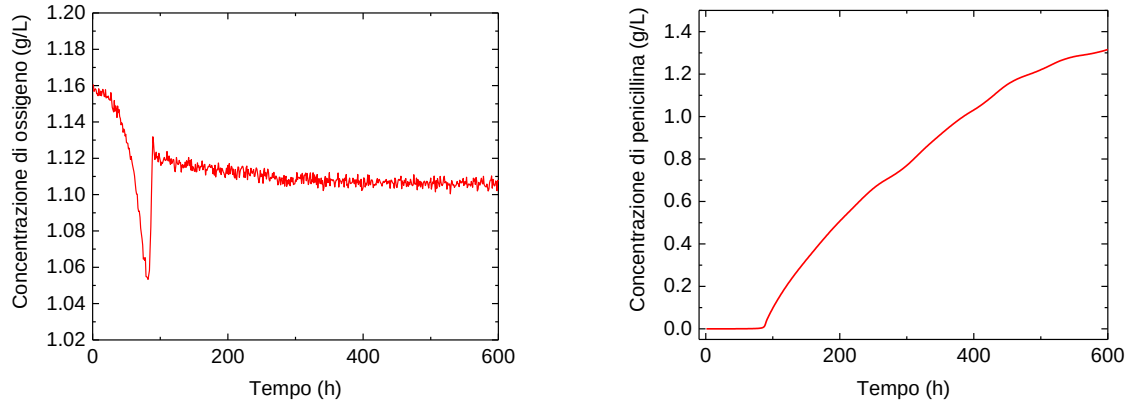
In Figura 2.1*a* sono riportati gli andamenti nel tempo della concentrazione di biomassa e quella di substrato. Si nota una forte dipendenza tra il consumo del substrato e la crescita della biomassa. Si riconoscono due fasi del processo, quella propriamente batch in cui non viene alimentato il substrato, e quella fed-batch dove, una volta raggiunto il valore minimo della sua concentrazione, entra una portata costante per il mantenimento e crescita della biomassa (Figura 2.*2a* , linea tratteggiata).



(a) (b)

Figura 2.2. *Andamenti nel tempo (a) di concentrazione di substrato e di biomassa e (b) di portata di substrato.*

La variabile di alimentazione di substrato *F* è una funzione a tratti, e assume valore costante per concentrazioni substrato inferiori a una certa soglia, mentre è nulla per valori superiori (Figura 2.2*b*).



(a) (b)

Figura 2.3. *Andamenti nel tempo della concentrazione di ossigeno disciolto (a) e della concentrazione di penicillina (b).*

Il profilo della concentrazione di ossigeno (Figura 2.3*a*) riflette i fenomeni coinvolti nel processo: nella fase fed batch, in presenza di un'alta concentrazione di substrato (Figura 2.2*a*, linea rossa), si assiste ad una rapida crescita di biomassa, che richiede un ingente consumo di ossigeno (per cui la sua concentrazione cala, Figura 2.3*a*); nella fase fed-batch il substrato disponibile è quello alimentato, e, sebbene la crescita della biomassa continui, il suo incremento nel tempo è molto meno marcato; e la concentrazione di ossigeno ritorna a valori prossimi a quelli di saturazione. Il profilo di concentrazione di penicillina nel tempo è simile a quello della biomassa (Figura 2.3*b*).

2.2 Generazione dei dati per la procedura di diagnosi del PMM

Il simulatore del processo di fermentazione viene utilizzato per generare due set di dati. Il primo, secondo le definizioni fornite in § 1.2.2.1, corrisponde alle misure storiche del processo, ovvero ad un insieme di misure delle variabili in ingresso e in uscita dal processo. Poiché non sono disponibili dei dati di linea di un processo reale, tale set di dati è generato con il simulatore, utilizzando le equazioni e i parametri originali del modello. Il secondo set di dati è costituito dagli stessi valori degli ingressi del primo set di dati e dalle predizioni delle variabili di uscita, ottenute modificando il modello, ovvero introducendo un certo errore, in modo da forzare una discrepanza tra i valori delle misure storiche e quelli delle misure simulate.

I due set di dati sono rappresentati tramite due matrici, la matrice di modello, $\mathbf{X}_{dati,M}$ e quella di processo, $\mathbf{X}_{dati,\Pi}$. Ogni colonna delle matrici corrisponde a una variabile del processo, le righe

invece rappresentano i campioni, ognuno dei quali viene generato con una differente combinazione di variabili in entrata.

2.2.1 Scelta delle variabili incluse nel set di dati

Le variabili di input scelte per generare i due set di dati sono solo alcune delle variabili in ingresso disponibili, ovvero quelle la cui variazione ha un effetto maggiore sulla resa finale di penicillina. Tale selezione è stata effettuata utilizzando un modello PLS (*partial least square regression*, regressione parziale ai minimi quadrati; §1.1.2).

Vengono simulati 100 batch, usando valori di ingresso selezionati casualmente all'interno degli intervalli di validità del simulatore. I dati generati sono utilizzati per costruire un modello PLS per rappresentare le relazioni che intercorrono tra una matrice degli ingressi (predittori) **X** e una matrice delle uscite del processo (risposte) **Y**. Le due matrici includono rispettivamente i valori di:

- **X**: concentrazione iniziale di substrato e biomassa, portata di aria, potenza di agitazione, temperatura dell'alimentazione, portata di substrato;
- **Y**: concentrazione finale di substrato, di biomassa, di penicillina, di ossigeno disciolto, e il volume.

Grazie ai parametri del modello PLS vengono calcolati gli indici VIP (*variable influence on projection*, influenza delle variabili sulle proiezioni), i quali danno un'indicazione su quali siano le variabili che hanno maggiore influenza sulle risposte. Una regola proposta da Eriksson *et al.* (1999) prevede di scartare i predittori la cui VIP è minore di uno, anche se una soglia di 0.7-0.8 può essere accettabile.

In Figura 2.4 vengono riportati i risultati dell'analisi: si nota che la temperatura dell'alimentazione (a causa della presenza del sistema di regolazione) e la concentrazione iniziale di biomassa sono poco rilevanti nel determinare le risposte del processo, poiché presentano valori poco significativi dell'indice VIP. La potenza di agitazione presenta valori di VIP generalmente bassi, tranne rispetto alla concentrazione di ossigeno disciolto e di substrato.

Nella generazione dei set di dati con il simulatore per applicare la procedura di diagnosi del PMM , i valori di concentrazione iniziale di biomassa e la temperatura di alimentazione vengono quindi posti ai valori nominali (riportati nel caso base del simulatore), mentre è inclusa una variabilità dell'input della potenza di agitazione, importante nel determinare il valore del coefficiente di trasporto dell'ossigeno, implicato nella determinazione del PMM (vedi §2.2.2), oltre a valori variabili della portata di aria, della portata di substrato e della concentrazione iniziale di tale specie.

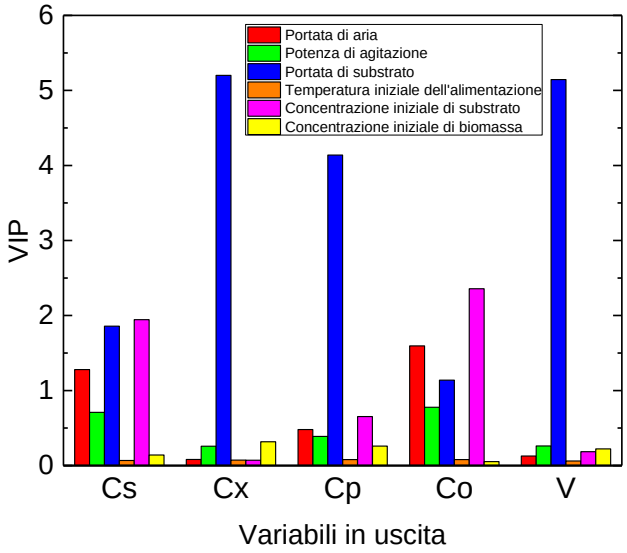


Figura 2.4. *Indici VIP delle variabili in ingresso. Le variabili in uscita dal processo sono (da sinistra): Concentrazione finale di substrato Cs, Concentrazione finale di biomassa Cx, concentrazione finale di penicillina Cp, concentrazione finale di ossigeno disciolto Co, volume finale V.*

Le variabili considerate nella generazione di $\mathbf{X}_{dati,M}$ e $\mathbf{X}_{dati,\Pi}$ sono riportate in Tabella 2.3.

Tabella 2.3. *Variabili che compaiono nel set di dati utilizzato per le procedure di diagnosi del PMM e loro ordine di apparizione nelle matrici di dati.*

Numero variabile	Tipo variabile	Simbolo
1	Portata di aria	f_g
2	Potenza di agitazione	P_w
3	Portata di alimentazione di substrato	F
5	Concentrazione di substrato	C_s
6	Concentrazione di biomassa	C_x
7	Concentrazione di penicillina	C_p
8	Concentrazione di ossigeno disciolto	C_{ox}
9	Volume	V

2.3 Introduzione del PMM

In questa Tesi vengono analizzati due diversi mismatch parametrici, causati da un'errata stima del valore di uno dei parametri presenti nel modello a principi primi considerato.

La presenza di discrepanze tra i dati simulati e i dati reali del processo è uno dei tipici problemi riscontrati nell'utilizzo di modelli a principi primi a scopi industriali, soprattutto quando un modello, sviluppato per un certo sistema, viene utilizzato per rappresentare un sistema simile, per esempio nel caso del trasferimento di processo tra due apparecchiature di diversa scala. Il passaggio della produzione da un impianto di laboratorio ad un impianto pilota è un'operazione

che viene eseguita nelle prime fasi di progettazione degli impianti industriali per verificare, prima del dimensionamento su scala industriale, che il processo, adattato su nuove dimensioni, abbia le caratteristiche di produttività (come resa e conversione) richieste, e che esso presenti, rispetto alle misurazioni delle variabili in uscita, risultati conformi al modello che deve descrivere i fenomeni nel processo.
In un processo governato da svariati fenomeni, come quello del caso studio in esame, l'applicazione di un criterio di *scale-up* che non tenga conto di una variazione della modalità con la quale avvengono tali fenomeni (per esempio la fluidodinamica nel reattore, la crescita della biomassa, fenomeni di trasporto di specie chimica e dell'energia, e del regime di aerazione) può facilmente portare a un'errata stima dei parametri, tale da causare il PMM. Nel caso studio in esame, sono state considerate due diverse possibili cause di PMM nel caso di *scale up* del reattore (Appendice A), che vengono descritte in seguito.

2.3.1 Primo PMM parametrico: modifica del valore di $K_l a$

Durante uno *scale-up* del reattore di fermentazione da una scala di laboratorio a una scala di impianto pilota (Appendice A), uno dei parametri critici da valutare è il coefficiente volumetrico di trasporto di massa dell'ossigeno. Nei processi in reattori agitati, il trasporto di ossigeno alla biomassa è infatti un fenomeno di cruciale importanza per il dimensionamento dell'apparecchiatura e la conduzione del processo, ed esso dipende da diverse variabili, tra cui le più importanti sono la velocità di agitazione (o la potenza fornita alla girante), il numero e la tipologia degli agitatori, la portata di gas e la geometria dell'ambiente di reazione. Il trasporto di ossigeno si può anche considerare come lo stadio cineticamente determinante del processo di fermentazione, a causa della sua solubilità generalmente bassa nel mezzo di coltura.

2.3.1.1 Determinazione sperimentale di $K_l a$

Il fenomeno di trasporto dell'ossigeno all'interno della fase liquida viene generalmente modellato facendo ricorso alla teoria dei 2 film (Whitman, 1923) in cui il flusso J (portata per unità di superficie è:

$$J = K_l\left(C_i - C\right) \tag{2.14}$$

Dove $K_l a$ è il coefficiente di trasporto di massa locale (m/s); siccome è difficile determinare la concentrazione di ossigeno all'interfaccia gas-liquido, si considera la concentrazione C^* di saturazione del liquido (*bulk*) in equilibrio con la concentrazione nel *bulk* gassoso:

$$J = K_l a\left(C^* - C\right), \tag{2.15}$$

dove *a* è l'area interfacciale specifica. Generalmente si determina sperimentalmente l'intero termine $K_l a$. Usando metodi sperimentali si usa un'equazione di bilancio di massa dell'ossigeno, che include l'OTR (*oxygen transfer rate*, dalla fase gas al liquido) e l'OUR (*oxygen uptake rate*, ossigeno trattenuto dal microorganismo):

$$\frac{dC}{dt} = OTR - OUR \tag{2.16}$$

I principali metodi sperimentali per la determinazione del coefficiente di trasporto sono i seguenti.

- metodi chimici: i primi a essere utilizzati, tendono a sovrastimare il coefficiente di trasporto in caso di reazioni molto veloci;
- metodi fisici: i più usati, stimano la concentrazione di ossigeno disciolto in un processo del suo assorbimento o desorbimento;
- misure dirette dell'OTR: un analizzatore misura la concentrazione di ossigeno nelle portate in entrata e uscita dal reattore, in presenza del microorganismo, e viene determinata l'OUR.

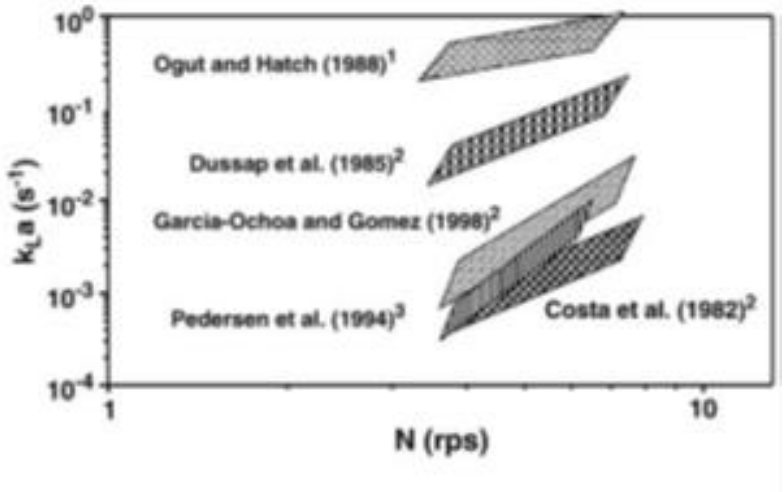


Figura 2.5 *Confronto dei valori di $K_l a$ ottenuti con differenti metodi, in funzione della velocità dell'agitatore in soluzioni a comportamento non newtoniano. Legenda:1:metodo chimico;2: metodo fisico;3:metodo Kr (misurazione di radioattività nella portata in uscita a seguito dell'iniezione di un tracciante nel mezzo di coltura). Da: Garcia-Ochoa e Gomez, 1988.*

Comparando risultati sperimentali per la stima di $K_L a$ ottenuti con metodi diversi, si osserva che il suo valore oscilla da 10^{-1} h a 10^3 h. In Figura 2.5, dove si ha una panoramica del suo *range* dei valori di determinazione sperimentale.

2.3.1.2 Correlazioni empiriche per $K_L a$

In letteratura sono presenti numerose correlazioni per la predizione del coefficiente di trasporto dell'ossigeno, sia in forma di equazioni dimensionali che adimensionali. Tuttavia, esistono

frequenti discrepanze tra il valore dei dati sperimentali e quelli predetti dalle formule. Nei bioreattori con agitazione le variabili associate al regime di moto e alla geometria del reattore giocano un ruolo importante nella determinazione di $K_l a$. La formulazione utilizzata è stata adottata da Bajpai e Reuss (1980) e utilizzata da Birol *et al.* nel modello implementato nel simulatore, e si basa sulla relazione proposta da Van't Riet (1979), dove compare il rapporto tra potenza di agitazione P e volume V:

$$K_l a = C \cdot V_s^{\ a} \cdot (P/V)^b \cdot \mu_a^{\ c} \tag{2.17}$$

La costante C dipende dalle caratteristiche geometriche del reattore e dal tipo di agitatore impiegato, μ_a è la viscosità del liquido e V_s è la velocità superficiale del gas.

2.3.1.3 Perturbazione di $K_l a$

Come mostrato in 2.3.1.1, un'errata predizione del valore di $K_l a$ rappresenta un caso molto verosimile, in quanto vi sono molte correlazioni per tale parametro, con un vasto *range* di valori predetti e dipendenze rispetto a molte variabili operative, caratteristiche geometriche e proprietà reologiche e termodinamiche del fluido. Anche le numerose regole di *scale-up* portano a risultati differenti, alcuni tali da produrre deviazioni notevoli dei valori di variabili di processo misurate rispetto a quelli calcolati usando il modello, una volta derivato il valore di $K_L a$. Appare dunque opportuno, utilizzare questo parametro per generare il PMM al fine di condurre uno studio sulle tecniche di diagnostica. Il valore del coefficiente di trasporto dell'ossigeno è stato opportunamente perturbato in modo da causare un PMM che infici sul funzionamento del processo, soprattutto sulla sua resa, in base ai risultati ottenuti da un'analisi di sensitività (Figura 2.6). In Figura 2.6 si nota che, per variazioni del -90% di $K_L a$, la concentrazione del prodotto alla fine della simulazione diminuisce del 10%.

Considerando che, a seconda dei metodi utilizzati per stimare $K_L a$ sperimentalmente, si ottengono valori nei range di 10^{-1} h e 10^3 h, una variazione negativa del 90% è considerata accettabile, dal momento che, per il valore iniziale di 100 L del volume il valore del coefficiente di trasporto dell'ossigeno è 123 h^{-1}. Nell'analisi di sensitività non sono state studiate variazioni della resa per variazioni positive del parametro: un incremento positivo del coefficiente volumetrico di trasporto di massa produce variazioni estremamente ridotte sulla concentrazione di prodotto.

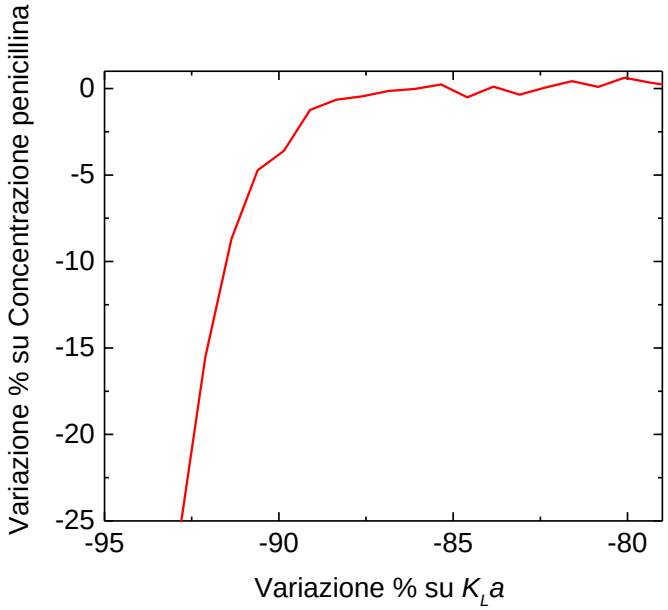


Figura 2.6. *Variazione percentuale sulla concentrazione finale di penicillina in funzione della variazione percentuale sul coefficiente volumetrico di trasporto di massa dell'ossigeno.*

Osservando l'equazione di variazione della concentrazione di ossigeno nel tempo della simulazione, si nota che, per un valore più alto di K_La, la forza motrice (C_L^*-C_L) diminuisce come effetto dell'aumentato trasporto di ossigeno nella fase liquida, creando così una compensazione che impedisce un impatto apprezzabile sulla resa del processo. Pertanto, in una diagnosi del PMM parametrico, il set di dati storico viene generato utilizzando il valore originale del parametro α che compare nella (2.9); il set di dati di modello riceve gli stessi valori delle variabili in ingresso delle misure storiche, ma α è diminuito del 92%, in modo da garantire una variazione nella resa di prodotto.

2.3.2 Secondo PMM parametrico: modifica del valore di Y_{sx}

Un altro parametro viene scelto per determinare un PMM, in modo da poter testare le metodologie di diagnosi in più casi. Questo è la costante di resa di biomassa rispetto al substrato $Y_{sx.}$ Su tale parametro è stata condotta un'analisi di sensitività per stabilire un'opportuna variazione di tale parametro, tale da causare una variazione della resa in penicillina rilevante.
Come si nota in Figura 2.7, per variazioni tra il +50% e il -50% di tale parametro si riscontrano forti variazioni nella resa in penicillina. Dal punto di vista di uno studio per la diagnosi del PMM, anche questo parametro può essere utilizzato per dare luogo a due set di dati in cui vi sia una differenza nella struttura di correlazione, analogamente al procedimento presentato in §2.3.2.3. In tal caso, il parametro Y_{sx} è stato aumentato del 50% per creare il set di dati di modello.

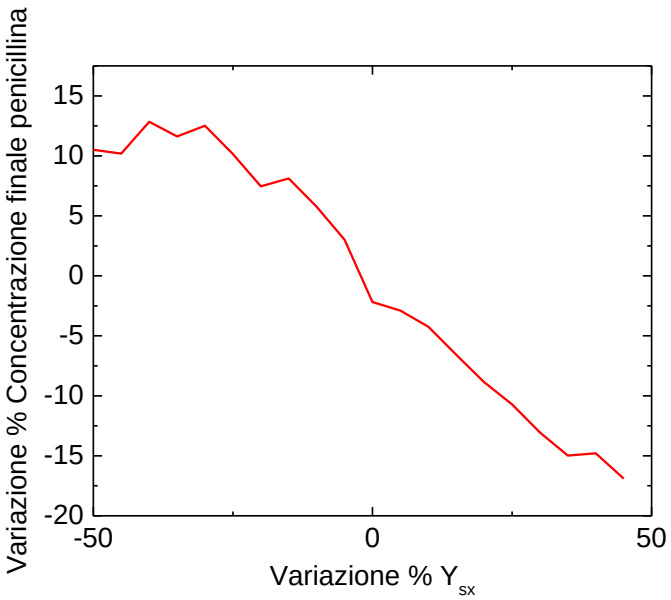


Figura 2.7. *Variazione percentuale della concentrazione finale di penicillina in funzione della variazione percentuale della costante di resa in biomassa rispetto al substrato,* Y_{sx}.

Capitolo 3

Diagnosi della mancata corrispondenza tra modello e processo: metodo 1

In questo Capitolo la metodologia di diagnosi di un PMM (*process/model mismatch*) elaborata da Meneghetti *et al.* (2014) è applicata ad un nuovo caso studio, in cui viene considerato un modello a principi primi di un processo di fermentazione della penicillina (Birol *et al.*, 2002). Vengono analizzate due possibili cause di PMM parametrico. Per entrambi i casi trattati, sono comparati i risultati ottenuti testando la metodologia utilizzata su due diversi tipi di set di dati simulati: il primo, in cui sono considerati solo i valori in ingresso e in uscita dal processo, e il secondo, in cui vengono considerati tutti i dati disponibili dall'inizio alla fine del processo. I risultati ottenuti hanno contribuito al lavoro riportato in Meneghetti (2016).

3.1 Generazione dei dati

Viene creato un set di dati di ingresso di L=100 combinazioni delle variabili riportate in Tabella 3.1, i cui valori rispettano gli intervalli di validità del modello riportati in Tabella 2.2. Ognuna di queste combinazioni rappresenta le condizioni iniziali utilizzate per simulare (utilizzando il simulatore Pensim, come illustrato in §2.1.2) un batch di produzione di penicillina. A partire da questo set di ingresso, vengono quindi generati i valori di concentrazione della biomassa, del substrato, dell'ossigeno disciolto, della penicillina e del volume di reazione per tutta la durata del processo (T = 600 h). Tra le 100 combinazioni di variabili di ingresso e le relative uscite generate, ne vengono scelte N = 28. Per garantire una distribuzione dei valori di concentrazione finale di penicillina sufficientemente uniforme tra il valore minimo e massimo disponibili, la selezione avviene rimuovendo quelle combinazioni che garantiscono una distribuzione uniforme dei valori di concentrazione del prodotto tra il valore minimo e quello massimo, una volta scelto il numero di combinazioni N. [1]

Utilizzando il modello originale (il set di Equazioni 2.1-2.13, senza aver perturbato alcun parametro) viene generato un set di dati di misure storiche, di dimensione [N×K×T], che

[1] La selezione dei campioni avviene usando la funzione Matlab® '*reducennsamples.m*', che opera una riduzione del set di campioni selezionandone un sottoinsieme in base alla tecnica della *nearest neighbor distance*. Questa consiste in una selezione dei campioni, rimuovendo i campioni simili tra loro sulla base della loro distanza nello spazio multivariato (formato dai campioni).

raccoglie le misure delle variabili del processo considerate per ogni istante della simulazione, per tutti i campioni (matrice $\mathbf{X}_{\text{dati},\Pi}$).

Tabella 3.1. *Valori iniziali delle variabili in entrata nei campioni delle matrici di dati.*

Variabile	Valori iniziali
Portata di aria, f_g [L/h]	3; 4.4; 5.8; 7.2; 8.6; 10
Potenza di agitazione, P [W]	20; 32; 38; 44; 50
Portata di substrato, F [L/h]	0.035; 0.037; 0.039;0.0431;0.0451
Concentrazione iniziale di substrato, C_s**(0)** [g/L]	5;8;11;17;20

Lo stesso modello a principi primi, in cui uno dei due parametri descritti in §2.3 è stato modificato, viene utilizzato per generare un secondo set di dati, partendo dalle stesse condizioni iniziali, denominato $\mathbf{X}_{\text{dati,M}}$. In base alla metodologia descritta in §1.2.1, le K variabili originarie sono combinate tra loro e con i parametri di modello per dare origine a V variabili ausiliarie. La scelta delle variabili ausiliarie viene effettuata in base ai termini che compaiono nelle equazioni di modello. In particolare, si considerano gli addendi delle varie equazioni differenziali, che sono già delle combinazioni non lineari di variabili del processo e alcuni parametri. In tal modo le variabili ausiliarie riflettono un preciso significato nell'ottica dell'analisi ingegneristica del processo, poiché sono termini presenti in bilanci di materia, energia e fattori cinetici.
Secondo questa logica, in base alle equazioni algebriche e differenziali definite in §2.1.1, vengono identificate le seguenti variabili ausiliarie:

$$\begin{array}{lll} x_1(n,t) = K_l a \cdot \left(C_l^{\,*} - C_l(n,t)\right) & x_4(n,t) = K \cdot C_p(n,t) & x_7(n,t) = C_l(n,t) \\ x_2(n,t) = C_x(n,t) \cdot \mu & x_5(n,t) = \dfrac{x_2(n,t)}{Y_{x/s}} + \dfrac{x_3(n,t)}{Y_{p/s}} + C_x(n,t) \cdot m_x & x_8(n,t) = C_x(n,t) \\ x_3(n,t) = \mu_{pp} \cdot C_x(n,t) & x_6(n,t) = \dfrac{x_2(n,t)}{Y_{x/o}} + \dfrac{x_3(n,t)}{Y_{p/o}} + C_x(n,t) \cdot m_o & x_9(n,t) = C_{s,}(0) - C_s(n,t) \end{array} \tag{3.1}$$

Ciascun elemento $x_v(n,t)$ della matrice $\mathbf{X}_v$ di dimensione $[N \times T]$ rappresenta la v-esima variabile ausiliaria calcolata per il t-esimo istante di tempo dell'n-esimo campione.
In questa Tesi, la metodologia viene applicata considerando due diversi set di dati per ognuno dei *mismatch* parametrici analizzati:

- caso 1: vengono considerati solo i valori delle variabili in ingresso ei valori finali delle variabili in uscita dal processo. In tal caso le matrici di dati di processo e di modello, $\mathbf{X}_\Pi$ e $\mathbf{X}_M$, risultano matrici bidimensionali $[N \times V]$ per cui la diagnosi del PMM viene effettuata basandosi sulla metodologia sviluppata per sistemi stazionari. Questa scelta offre l'opportunità di verificare le prestazioni della metodologia effettuando l'analisi di un sistema più semplice di quello dinamico. L'analisi di un sistema dinamico, infatti, è

solitamente caratterizzata da una serie di problemi dovuti per esempio a dati mancanti o al fatto che le variabili di processo vengono misurate con intervalli di campionamento diversi, in base alla strumentazione disponibile. Inoltre, poiché per ottenere una data resa di prodotto in diverse condizioni di processo la durata di diversi batch è solitamente diversa (ma anche, se come nel caso in esame la durata è la stessa, l'evoluzione di batch diversi è diversa), una sincronizzazione preliminare dei dati è spesso suggerita, se non necessaria;

- caso 2: si tiene conto dell'evoluzione delle variabili nel tempo. In questo caso vengono considerate tutte le misurazioni disponibili per i T campionamenti nel tempo, generando due matrici tridimensionali $\mathbf{X}_\Pi$ e $\mathbf{X}_M$ di dimensioni $[N \times V \times T]$. L'uso di set di dati dinamici permette di ricavare maggiori informazioni sul PMM rispetto ai dati in stato stazionario, potendo analizzare la sua evoluzione nel tempo e l'effetto sulle diverse variabili in differenti istanti del processo.

Si noti che le variabili ausiliarie x_7, x_8 e x_9, che rappresentano misure di variabili del processo, sono aggiunte al set di variabili ausiliarie per evidenziare il cambiamento della struttura di correlazione tra le due matrici $\underline{\mathbf{X}}_\Pi$ e $\underline{\mathbf{X}}_M$.

3.2 Applicazione del metodo di diagnosi: caso 1

Nel primo caso vengono considerate due matrici bidimensionali di processo e di modello, $\mathbf{X}_\Pi$ e $\mathbf{X}_M$, di dimensioni $[N \times V]$, in cui le variabili ausiliarie vengono calcolate utilizzando solo i valori delle condizioni iniziali e i valori finali (calcolati per l'ultimo istante $t = T$ di simulazione del processo) delle variabili in uscita.

3.2.1 Esempio 1a: modifica di ($K_l \cdot a$)

Nel primo esempio trattato, un errore viene introdotto nel modello a principi primi utilizzato per descrivere il processo di fermentazione della penicillina, modificando il coefficiente di trasporto dell'ossigeno come riportato in §2.3.1.3. In particolare, considerando la formulazione adottata da Birol *et al.* (2002, Eq. 2.9), viene ridotto del 92% il parametro α, portando a una sottostima della resa in penicillina del 10% in media.

I valori degli elementi della matrice di processo $\mathbf{X}_\Pi$ vengono calcolati in base alle variabili ausiliarie calcolate considerando i valori delle variabili misurate del processo (ovvero i valori simulati senza alcuna modificati dei parametri del modello). Si sottolinea che, dato che in generale non vi è alcuna conoscenza a priori della causa del mismatch, il valore del parametro ($K_l \cdot a$) utilizzato sia per il calcolo della matrice di processo che di modello, è quello errato. Seguendo la procedura descritta in § 1.2.1, viene costruito un modello PCA sulla matrice di modello $\mathbf{X}_M$, considerando solo le prime due componenti principali, scelte in base alla regola

dell'autovalore maggiore di uno (criterio definito in §1.1.1.2). Queste PC descrivono il 91% della variabilità dei dati (Tabella 3.2).

Tabella 3.2 *Esempio 1a: variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su* $\boldsymbol{X}_M$.

Numero PC	Autovalore della matrice di covarianza	R^2	R^2 cumulato
1	5.48	60.01	60.01
2	2.85	31.66	91.67

La matrice $\mathbf{X}_\Pi$ viene poi scalata sulla media e la deviazione standard di $\mathbf{X}_M$, e viene proiettata nello spazio del modello calibrato su $\mathbf{X}_M$.

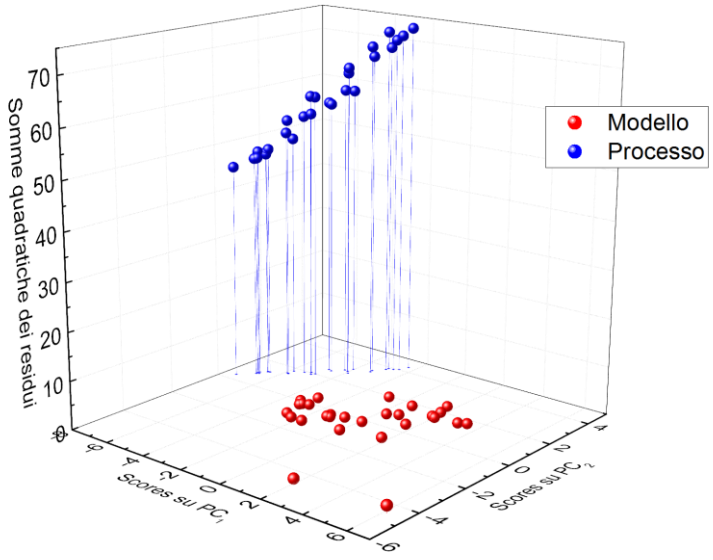


Figura 3.1. *Esempio 1a: residui dei campioni delle matrici di processo e modello,* $\boldsymbol{X}_\Pi$ *e* $\boldsymbol{X}_M$*, lungo le prime due componenti principali del modello PCA.*

In Figura 3.1 sono riportati gli *scores* dei campioni dei due set di dati sul piano formato dalle due componenti principali (PC_1 e PC_2), mentre lungo l'asse perpendicolare al piano sono riportati i residui quadratici del modello PCA relativi ad ogni campione. Si nota che, mentre i campioni di $\mathbf{X}_M$ (evidenziati in rosso) si trovano molto vicino piano formato dalle prime due componenti principali, per i campioni di $\mathbf{X}_\Pi$ (in blu) la distanza dal piano è rilevante. Ciò dimostra che le direzioni delle prime due PC, che sono in grado di rappresentare in modo soddisfacente la variabilità dei dati di $\mathbf{X}_M$, a causa della presenza di un PMM non sono in grado di rappresentare anche la struttura di correlazione di $\mathbf{X}_\Pi$.

In Figura 3.2, vengono riportati i *loadings* delle variabili ausiliarie rispetto alle due componenti principali del modello PCA. Si osserva che le variabili ausiliarie x_1 x_5, x_6 e x_8 presentano valori dei *loadings* molto simili, in particolare alti *loadings* sulla prima PC, che cattura il 60% circa

della variabilità dei dati, e appaiono anticorrelati, lungo PC1, con x_2, x_3, x_4 e x_7. Le variabili ausiliarie x_3, x_4, x_5, x_7 (e, marginalmente, x_9) presentano alti valori di *loadings* su PC2, risultando correlate positivamente lungo tale PC.

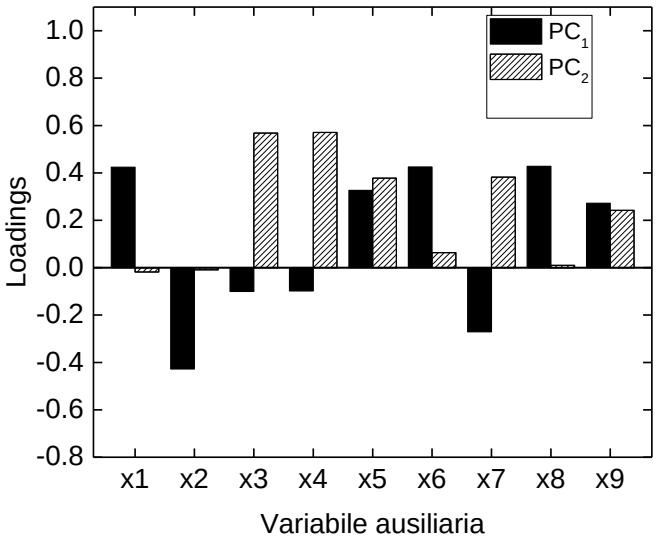


Figura 3.2. *Esempio 1a: loadings per il modello PCA costruito su* $\boldsymbol{X}_M$ ..

Per estrarre indicazioni utili per comprendere la reale causa del PMM analizzato, le matrici dei residui $\mathbf{E}_M$ ed $\mathbf{E}_{\Pi,}$ risultanti dal modello PCA, vengono utilizzate per calcolare gli indici MRLR (*mean residuals-to-limit ratio*) per ogni variabile ausiliaria secondo le (1.25) e (1.26).
I risultati della diagnosi sono mostrati in Figura 3.3, in cui vengono riportati i valori dell'indice diagnostico MRLR calcolato per ciascuna variabile ausiliaria. Si osserva che la variabile ausiliaria x_1 presenta un valore di MRLR più alto rispetto alle altre variabili ausiliarie; questo risultato permette di identificare la causa del PMM nel termine del modello in cui compare il coefficiente di trasporto dell'ossigeno.

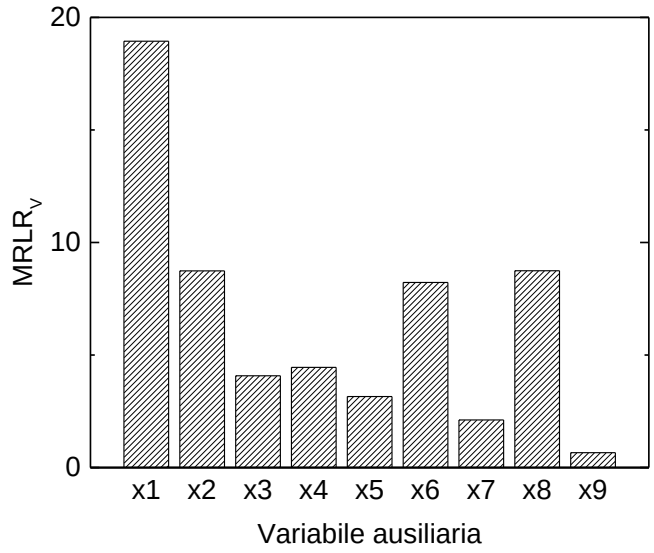


Figura 3.3. *Esempio 1a: indici $MRLR_v$ calcolati a partire dalle proiezioni di $\boldsymbol{X}_{\Pi}$ sul modello PCA costruito su $\boldsymbol{X}_M$.*

3.2.2 Esempio 1b: modifica di Y_{sx}

La stessa metodologia di diagnosi di un PMM viene applicata nel caso in cui nel modello a principi primi venga modificato il valore della costante di resa, Y_{sx} (come riportato in §2.3.2), a cui viene assegnato un valore pari a 0.2 (invece di 0.45), causando una sottostima del valore finale di concentrazione di penicillina del 32% in media.
Anche in questo caso, il modello PCA viene costruito sulla matrice di modello $\mathbf{X}_M$, utilizzando 3 PC, che catturano il 93% della variabilità dei dati (Tabella 3.3).

Tabella 3.3 *Esempio 1b: variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su $\boldsymbol{X}_M$.*

Numero PC	Autovalore della matrice di covarianza	R^2	R^2 cumulato
1	7.55	83.84	83.84
2	0.86	9.51	93.35

Il confronto delle matrici dei residui $\mathbf{E}_M$ e $\mathbf{E}_\Pi$, utilizzando l'indice MRLR, è riportato in Figura 3.4, dove la variabile ausiliaria x_4 presenta un valore particolarmente alto di questo indice, insieme a x_3. In questo caso i risultati suggeriscono che il PMM sia dovuto a questi due termini del modello, che in realtà, anche se vengono influenzati dall'errore introdotto, non rappresentano l'effettiva causa del mismatch, la quale dovrebbe invece essere riconosciuta in x_5. Infatti, poiché la variazione della resa in penicillina dei dati del modello rispetto a quelli di processo causata dal PMM per la modifica della costante di resa è elevata (il 30% circa), è

possibile che la variabilità di x_4, nella costruzione del modello PCA, nasconda il contributo al PMM delle altre variabili ausiliarie nell'analisi dei residui.

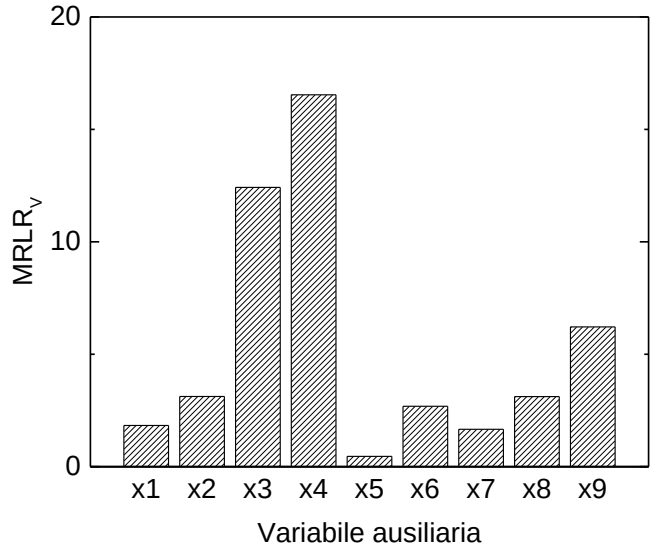


Figura 3.4. *Esempio 1b: indici $MRLR_v$ calcolati a partire dalle proiezioni di $\boldsymbol{X}_{\Pi}$ sul modello PCA costruito su $\boldsymbol{X}_M$.*

3.3 Applicazione del metodo di diagnosi: caso 2

In questa seconda applicazione, la diagnosi è effettuata considerando gli andamenti nel tempo delle variabili di ciascun campione per il processo e il modello, ovvero $\underline{\mathbf{X}}_{\Pi}$ e $\underline{\mathbf{X}}_{M}$ risultano matrici di dimensioni [N×V×T]. Poiché il processo presenta una fase batch e una fed-batch, risulta conveniente analizzare separatamente le due fasi, previa sincronizzazione delle traiettorie dei campioni (a causa del fatto che il passaggio da fase fed-batch a fase batch avviene in istanti di tempo diversi a seconda delle condizioni iniziali del campione). Per semplicità è stata analizzata solo la fase fed-batch, dopo aver effettuato un troncamento della fase batch di tutti i campioni. Il troncamento è stato eseguito dall'istante di tempo relativo al cambiamento di modalità del campione che presenta la massima durata della fase batch.

3.3.1 Esempio 2a: modifica di $K_l a$

La stessa procedura applicata in §3.2 è adattata alle matrici tridimensionali $\underline{\mathbf{X}}_{M}$ e $\underline{\mathbf{X}}_{\Pi}$, che pertanto sono sottoposte ad uno srotolamento orizzontale, seguendo la procedura descritta in §1.1.1.5 (e utilizzata da Meneghetti *et al.* (2015) nell'applicazione del metodo di diagnosi di un modello sviluppato per un processo di essiccazione), ottenendo due matrici di dimensioni [N×KT], $\mathbf{X}_{M}$ e $\mathbf{X}_{\Pi}$. Successivamente è costruito un modello MPCA su $\mathbf{X}_{M}$. Sono considerate 2 componenti principali (Tabella 3.4) basandosi sul fatto che in questo caso, avendo un sistema

dinamico in cui si considerano T istanti di tempo, si mantiene una componente principale se il relativo autovalore è tale che $\lambda \geq T$.

Tabella 3.4 *Esempio 2a: variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su* $\boldsymbol{X}_M$.

Numero PC	Autovalore della matrice di covarianza	R^2	R^2 cumulato
1	2736.1	63.33	63.33
2	1156.4	26.77	90.1

La matrice $\mathbf{X}_\Pi$ è poi proiettata nello spazio del modello e i residui della matrice di modello e di processo sono confrontati calcolando gli indici MRLR che, in questo caso, presentano anch'essi una specifica evoluzione nel tempo, come riportato in Figura 3.5.

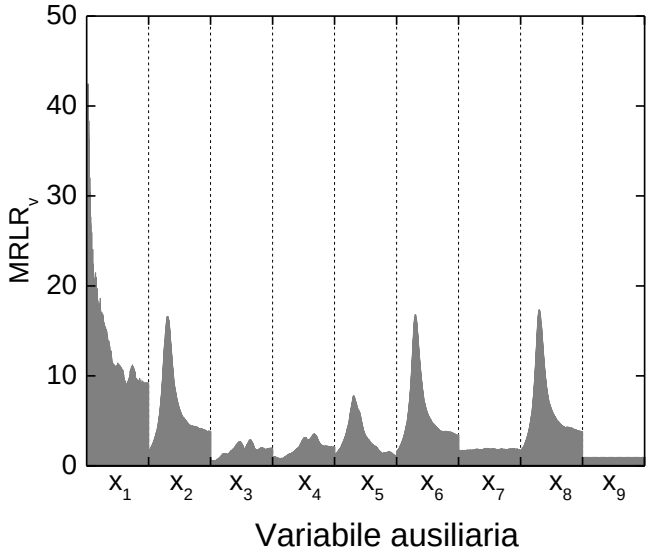


Figura 3.5. *Esempio 2a: indici* $MRLR_v$ *calcolati a partire dalle proiezioni di* $\boldsymbol{X}_\Pi$ *sul modello PCA costruito su* $\boldsymbol{X}_M$.

L'analisi dei residui evidenzia come la variabile x_1 presenti, in ogni istante di tempo, un valore di MRLR maggiore (o comparabile) rispetto a quelli delle altre variabili ausiliarie confermando i risultati ottenuti in §3.2.1. Anche in questo caso, le variabili ausiliarie x_2, x_6 e x_8 presentano valori simili e rilevanti di MRLR, ma in generale inferiori a quello della prima variabile ausiliaria. Si nota anche che l'andamento nel tempo degli indici MRLR delle variabili x_2, x_6 e x_8 appare influenzato da x_1: infatti la prima variabile ausiliaria presenta fin dal primo istante valori alti di MRLR, per poi decrescere, mentre il picco dell'indice per le altre variabili compare dopo un inziale ritardo, per poi assumere un andamento analogo a quello dei residui di x_1. Dalla dinamica del processo è quindi possibile dedurre come il PMM si manifesti lungo la durata del processo, in particolare nella parte iniziale della fase fed-batch del processo.

3.3.2 Esempio 2b: modifica di Y_{sx}

La stessa procedura utilizzata in §3.3.1 viene ripetuta considerando il caso in cui il modello a principi primi è perturbato modificando la costante di resa. Pertanto, un modello MPCA viene costruito sulla matrice $\underline{\mathbf{X}}_M$, e considerando 3 PC in grado di catturare il 95% della variabilità dei dati (Tabella 3.5), viene proietta su di esso la matrice srotolata $\mathbf{X}_\Pi$. L'analisi dei residui delle due matrici, effettuata tramite l'indice MRLR, viene riportata in Figura 3.6. Come nel caso precedente, si considera solo l'intervallo di tempo in cui si svolge la fase fed-batch del processo.

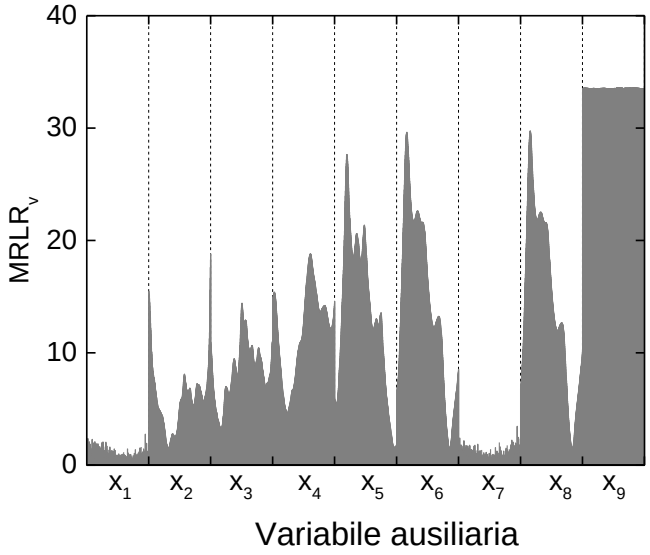


Figura 3.6. *Esempio 2b: indici $MRLR_v$ calcolati a partire dalle proiezioni di $\boldsymbol{X}_\Pi$ sul modello PCA costruito su $\boldsymbol{X}_M$.*

L'analisi dei residui fornisce indicazioni più chiare sul disallineamento tra processo e modello, rispetto all'analisi dei residui effettuata per un set di dati dove sono considerati solo i valori delle variabili in entrata al processo e quelli finali delle variabili in uscita. In Figura 3.6, le variabili ausiliarie x_6 e x_5 presentano valori più elevati dell'indice MRLR rispetto alle altre variabili, dimostrando che in questo caso è possibile ottenere dei risultati più consistenti rispetto all causa del PMM introdotto, essendo x_5 la variabile ausiliaria contenente il parametro modificato. Le variabili originali di concentrazione di biomassa e substrato presentano inoltre valori molto alti dell'indice diagnostico: ciò è coerente con la causa del PMM, dato che ad essere modificato è il parametro della costante di resa in biomassa rispetto alla concentrazione di substrato. Dato che la quinta e sesta variabile ausiliaria hanno valori analoghi di MRLR, non è comunque chiaro quale tra le due possa essere maggiormente associata al PMM; inoltre anche altre variabili ausiliarie, come x_2, x_3 e x_4 presentano valori rilevanti dell'indice MRLR. In tal caso, può essere conveniente reiterare la procedura di diagnosi, escludendo però le variabili ausiliarie x_1 e x_7, che presentano valori poco significativi dell'indice diagnostico.

Tabella 3.5 *Esempio 2b: variabilità dei dati catturata da ciascuna componente principale del modello PCA costruito su* $\boldsymbol{X}_M$.

Numero PC	Autovalore della matrice di covarianza	R^2	R^2 cumulato
1	3196.48	73.99	73.99
2	586.01	13.56	87.55
3	346.05	8.01	95.56

3.4 Conclusioni

L'analisi dei residui di un modello PCA facendo dell'uso dell'indice MRLR su un set di variabili ausiliarie è un utile strumento di identificazione del PMM. Considerando un processo dinamico, anche facendo uso di un set di dati semplificato (limitato ai valori di ingresso e ai valori di stato finale del processo) è possibile estrarre informazioni sulle variabili maggiormente responsabili del PMM.

Tuttavia, la presenza di variabili fortemente correlate rappresenta una limitazione all'utilizzo di tale procedura di diagnosi: una variabile che, in seguito alla proiezione dei dati di processo sul modello PCA, risulta associata al PMM, tende infatti a condividere simili valori dell'indice MRLR con variabili ad essa correlate.

Sfruttando la dinamica del processo, e quindi utilizzando un set di dati che raccoglie le misurazione *nel tempo* delle variabili, si possono ottenere maggiori informazioni sull'effetto del PMM nel corso del processo, soprattutto nel caso in cui questo si manifesti solo durante il processo e non nelle fasi iniziale e finale. Come visto nei risultati nel caso 2a, l'analisi sul set di dati dinamici può mostrare come il PMM si manifesta nel tempo, con tempistiche differenti sulle variabili incluse nell'analisi.

L'analisi in regime dinamico è comunque più complessa rispetto a quella condotta utilizzando solo i valori in entrata e i valori di stato finale del processo, nel caso i set di dati presentino valori mancanti delle variabili nel tempo o necessitino di un allineamento delle traiettorie delle variabili. Nel caso in cui si voglia indagare il PMM presente in un processo dinamico, ma senza utilizzare una procedura eccessivamente complessa, una possibile soluzione è quella di effettuare la diagnosi usando differenti set di dati, costruiti considerando i valori delle variabili in entrata al processo e quelli delle variabili calcolate in differenti istanti di tempo per ogni set di dati.

Infine, un aspetto importante per il miglioramento di questa tecnica diagnostica è la scelta delle variabili ausiliarie, al fine di massimizzare l'informazione deducibile dall'analisi dei residui. Limitando la scelta ad addendi presenti nelle equazioni di modello, l'analisi offre un riscontro in base a termini legati a fenomeni fisici presenti nel processo. Altre possibilità sono tuttavia in esame per potenziare la capacità diagnostica del metodo, mantenendone la robustezza.

I risultati della diagnosi possono altresì servire da base di comparazione con la metodologia di diagnosi presentata in §1.2.2 e discussa in §4.1.

Capitolo 4

Diagnosi della discrepanza tra modello e processo: metodo 2

In questo Capitolo viene proposto un metodo alternativo per la diagnosi della causa di un PMM (*process/model mismatch*) sviluppato adattando la metodologia proposta da Rato e Reis (2015), utilizzata nell'ambito del monitoraggio di processo. Lo scopo è fornire una possibile soluzione ai problemi riscontrati applicando la metodologia di diagnosi basata sull'analisi dei residui di un modello PCA (Capitolo 3), in particolare a causa della presenza di variabili ausiliarie fortemente correlate tra loro. La metodologia è applicata a due differenti set di dati: il primo (caso 1), in cui si considerano solo i valori delle variabili in entrata al processo e i valori finali delle variabili in uscita, il secondo (caso 2), in cui si considera la dinamica del processo, sfruttando i dati disponibili per la fase fed-batch del processo. Alcuni di questi risultati hanno contribuito al lavoro di Meneghetti (2016).

4.1 Caso 1

Vengono generati due set di dati, uno relativo alle misure storiche del processo (simulate con il modello in cui i parametri non sono stati modificati) e uno relativo alle misure simulate con il modello affetto da errore, in cui si considerano solo i valori delle variabili in entrata e i valori delle variabili nello stato finale del processo. Questi set di dati sono utilizzati per generare una matrice di processo e una di modello da analizzare utilizzando la procedura descritta in §1.2.2.2. In questo modo si considerano le stesse condizioni di analisi di un processo stazionario.

4.1.1 Generazione dei dati

In base alla procedura definita in §1.2.2, il secondo metodo proposto per la diagnosi delle cause di un *mismatch* tra processo e modello si basa sull'ipotesi di distribuzione normale dei coefficienti di correlazione parziale. A questo scopo, viene selezionato un set di valori delle variabili in ingresso (Tabella 4.1), utilizzato per generare diversi batch distribuiti normalmente intorno a dei valori di riferimento. In particolare, i valori medi selezionati per generare queste distribuzioni delle variabili in ingresso sono le combinazioni di concentrazione di substrato, potenza di agitazione, portata di alimentazione di substrato e portata di aria che garantiscono la massima variazione nella resa in prodotto in seguito alla modifica di uno dei due parametri del

modello considerati per produrre il PMM (il coefficiente di trasporto dell'ossigeno, $K_l a$, e la costante di resa in biomassa rispetto al substrato, Y_{sx}). Tale combinazione è stata selezionata dall'insieme di combinazioni utilizzate per generare il set di dati in §3.1.

Vengono prodotte quindi N = 100 combinazioni di variabili di ingresso, distribuite normalmente, che permettono di ricavare le matrici di dati di processo e modello, $\mathbf{X}_{\mathbf{dati},\Pi}$ e $\mathbf{X}_{\mathbf{dati},\mathrm{M}}$, di dimensioni [$N$×$K$], in cui K è il numero delle variabili di processo originali definite in §2.2.1. Dalle combinazioni non lineari tra le K variabili originali e i parametri del modello (inclusi quelli modificati per generare il PMM) si originano V variabili ausiliarie. In particolare, vengono scelte le combinazioni non lineari di variabili e parametri definiti in Eq. (3.1):

$$\begin{aligned} x_1(n,t) &= K_l a \cdot \left(C_l^{\,*} - C_l(n,t)\right) & x_4(n,t) &= K \cdot C_p(n,t) \\ x_2(n,t) &= C_x(n,t) \cdot \mu & x_5(n,t) &= \frac{x_2(n,t)}{Y_{x/s}} + \frac{x_3(n,t)}{Y_{p/s}} + C_x(n,t) \cdot m_x \\ x_3(n,t) &= \mu_{pp} \cdot C_x(n,t) & x_6(n,t) &= \frac{x_2(n,t)}{Y_{x/o}} + \frac{x_3(n,t)}{Y_{p/o}} + C_x(n,t) \cdot m_o \end{aligned} \quad (4.1)$$

Tale scelta è utile per fornire un termine di paragone dell'utilizzo di questa tecnica di diagnosi con i risultati forniti nel Capitolo 3. In base alla definizione riportata in §1.2.2.2, per ogni terna di variabili ausiliarie i, j e k si può calcolare un vettore di coefficienti di correlazione parziale di dimensioni [1×V·(V-1)·(V-2)/2], ovvero un totale di 60 coefficienti (con riferimento all'Eq. 1.29):

$$r_{i,j.k} = \frac{r_{i,j} - r_{i,k} \cdot r_{j,k}}{\sqrt{\left(1 - r_{i,k}\right)^2 \left(1 - r_{j,k}\right)^2}} \quad . \quad (4.2)$$

Per generare una distribuzione di ciascun coefficiente di correlazione parziale, calcolato per i campioni della matrice di modello, utile a definire appropriati intervalli di confidenza rispetto a cui comparare i coefficienti di correlazione calcolati per i campioni della matrice di processo, ognuno degli N campioni delle matrici $\mathbf{X}_\Pi$ e $\mathbf{X}_\mathrm{M}$ viene replicato per B = 300 volte. I due set di dati risultano quindi due matrici tridimensionali $\underline{\mathbf{X}}_\Pi$ e $\underline{\mathbf{X}}_\mathrm{M}$ di dimensioni [N×V×B]. Ogni sezione di dimensione [V×B] viene utilizzata per calcolare un vettore di coefficienti di correlazione parziale. All'interno di tale sezione è garantita una certa variabilità dei dati grazie al rumore sulle variabili introdotto dal simulatore Pensim. Dunque, per ognuno dei 60 coefficienti di correlazione parziale è possibile calcolare un vettore di N elementi, uno per ognuno degli N campioni del set di dati. Per ogni vettore di coefficienti di correlazione calcolato rispetto alla matrice di modello ne è calcolata una media ρ, che viene poi utilizzata per normalizzare gli elementi del vettore secondo la (1.30).

Infine, per ogni distribuzione di coefficienti di correlazione parziale normalizzati generati dalla matrice di modello, vengono calcolati i limiti di confidenza (secondo l'Eq.1.31) necessari per

applicare il metodo di diagnosi descritto in §1.2.2.2. Ogni variabile viene quindi classificata come 'GIALLO', 'ARANCIONE' o 'ROSSO' in base al suo grado di coinvolgimento nel determinare il PMM.

4.1.2 Esempio 1a: modifica di (K_l·a)

Nel primo caso, la presenza di un PMM viene forzata modificando il coefficiente di trasporto dell'ossigeno (K_l·a), modificando il parametro α nell'Eq. (2.9). Sebbene venga applicata una procedura di normalizzazione dei coefficienti di correlazione come descritto nel paragrafo precedente, la normalità delle distribuzioni di tali coefficienti non è verificata per tutti i coefficienti di correlazione parziale generati dalla matrice $\mathbf{X}_M$. Infatti, effettuando il test Anderson-Darling (1952) per verificare la validità dell'ipotesi di normalità, su 60 coefficienti di correlazione parziale generati, solo 33 risultano normalmente distribuiti. Per quanto riguarda i coefficienti per cui viene respinta l'ipotesi di normalità, circa il 50% di essi presenta comunque un andamento tendente alla normalità. In queste condizioni, si suggerisce pertanto di considerare i risultati ottenuti con cautela.

Per ogni coefficiente di correlazione parziale di $\underline{\mathbf{X}}_M$, vengono quindi definiti i limiti di confidenza calcolati con la (4.3). Infine, utilizzando le medie ρ delle distribuzioni di tali coefficienti, viene effettuata la normalizzazione dei coefficienti calcolati a partire dalle misure storiche ($\underline{\mathbf{X}}_\Pi$). Infine si conduce la procedura di diagnosi basata sull'assegnazione delle classi definite in §1.2.2.2. Per ogni campione, si verifica quali variabili abbiano la maggior distanza di coppia e la minore distanza di controllo. In Figura 4.1 sono riportati i risultati della diagnosi, in cui si osserva che la prima variabile ausiliaria presenta il massimo valore di campioni classificati come 'ROSSO' e una significativa percentuale di campioni classificati come 'ARANCIONE' (~20%). Questo risultato suggerisce che questa variabile ausiliaria può essere considerata come maggiore responsabile del PMM. Si osserva inoltre che, come conseguenza della presenza del PMM, altre variabili ausiliarie presentano una percentuale rilevante di campioni (60% dei batch) in cui sono classificate come 'GIALLO', come per esempio la seconda e la sesta variabile.

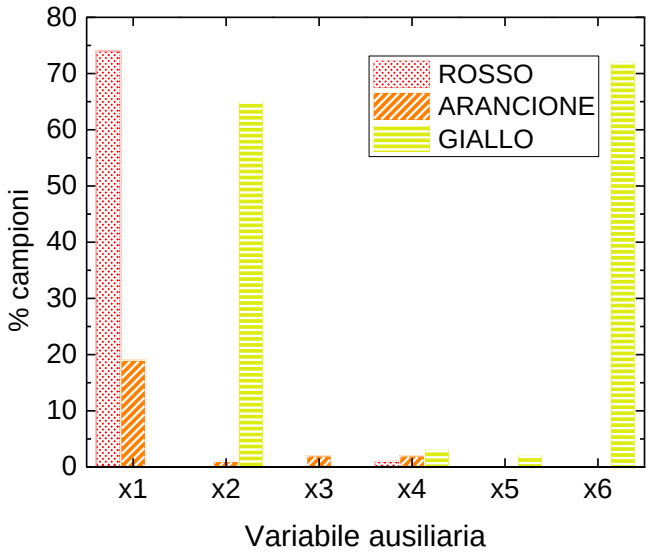


Figura 4.1. *Esempio 1a: numero di campioni (batch) in cui ogni variabile ausiliaria è stata classificata come ' ROSSO', 'ARANCIONE' o 'GIALLO'.*

4.1.3 Esempio 1b: modifica di Y_{sx}

In questo secondo esempio, viene forzata la presenza di un PMM modificando il valore della costante di resa (Y_{sx}, contenuta nella variabile ausiliaria x_5), come descritto in §3.2.2. Nuovamente, si utilizza un set di variabili ausiliarie definito in §3.1.

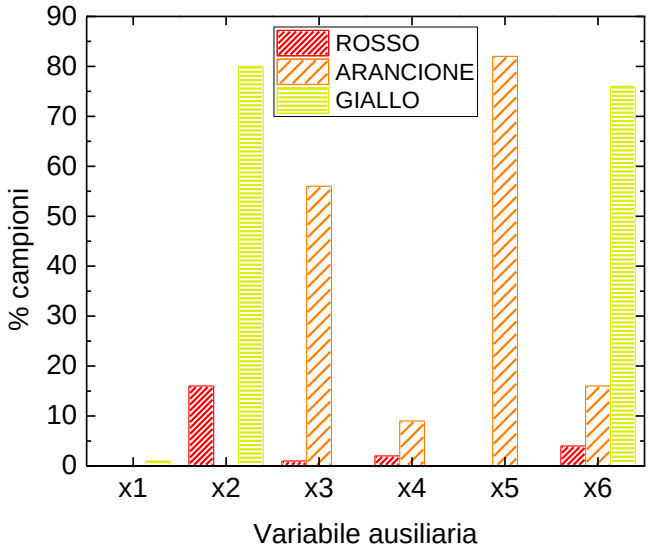


Figura 4.2. *Esempio 2a: numero di campioni (batch) in cui ogni variabile ausiliaria è stata classificata come ' ROSSO', 'ARANCIONE' o 'GIALLO'.*

I risultati dell'analisi, riportati in Figura 4.2, non offrono una chiara indicazione sulle cause del mismatch come nell'esempio precedente. Infatti, anche se la variabile x_5, contenente il

parametro modificato, viene classificata in più dell'80% dei batch come 'ARANCIONE', altre variabili ausiliarie sono classificate, con percentuali sebbene poco significative (~15%), come 'ROSSO', ovvero come maggiormente responsabili del mismatch. Inoltre anche la terza variabile ausiliaria è classificata come 'ARANCIONE' per una percentuale di campioni analoga a x_6.

Si sottolinea che a essere classificate per il maggior numero di campioni come 'ROSSO', 'ARANCIONE' o 'GIALLO' sono le variabili ausiliarie x_2, x_3, x_5 e x_6, che sono fortemente correlate, in quanto sono prodotti dei termini del modello per la concentrazione di biomassa. Sebbene l'utilizzo dei coefficienti di correlazione parziale venga effettuato proprio evitare che la correlazione tra variabili influenzate dal mismatch mascheri il contributo della vera causa di errore nel modello a principi primi, l'uso di coefficienti di correlazione parziale del primo ordine (che quantificano la correlazione tra due variabili, controllandone una terza) può non essere sufficiente per eliminare la correlazione rispetto ad altre variabili del set di dati. In futuro, può risultare quindi opportuno condurre la diagnosi con coefficienti di correlazione parziale di ordini superiori al primo, che evidenzino la correlazione tra due variabili controllando più variabili terze invece di una sola. Un'altra possibile soluzione è quella di ricorrere ad un set alternativo di variabili ausiliarie per condurre la diagnosi:

$$\begin{aligned} x_1 &= K_l a \cdot \left(C_l^{*} - C_l\right) & x_2 &= \mu \\ x_3 &= \mu_{pp} & x_4 &= K \cdot C_p \\ x_5 &= C_x \cdot \left(\frac{x_2}{Y_{x/s}} + \frac{x_3}{Y_{p/s}} + m_x\right) & x_6 &= C_x \cdot \left(\frac{x_2}{Y_{x/o}} + \frac{x_3}{Y_{p/o}} + m_o\right) \end{aligned} \tag{4.3}$$

Rispetto alle variabili ausiliarie richiamate nella (4.1), in questo caso le velocità specifiche di crescita della biomassa e di produzione di penicillina, μ e μ_{pp}, nelle variabili ausiliarie x_2 e x_3 non sono moltiplicate per la concentrazione di biomassa. Ne risultano due variabili ausiliarie che non sono addendi presenti nelle equazioni di modello. Ciò tuttavia permette di ridurre la correlazione di tali variabili ausiliarie con le variabili x_5 e x_6 e può risultare importante nell'applicazione di questa tecnica di diagnosi, in quanto è possibile ridurre ulteriormente il contributo della variabile originale di concentrazione della biomassa alla correlazione tra le variabili ausiliarie, nel caso una di queste sia indicata come responsabile del PMM. In Figura 4.3 è riportato il risultato della diagnosi: la variabile x_5 è evidenziata come fortemente legata al PMM, con il 70% dei campioni in cui è classificata come 'ROSSO'.

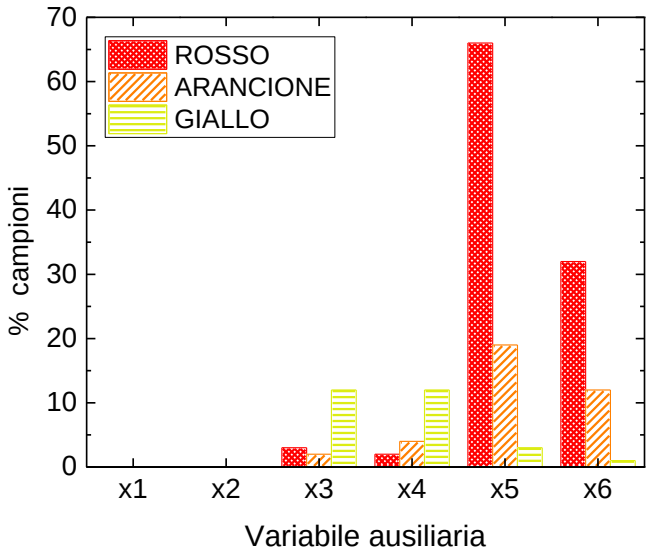


Figura 4.3. *Esempio 1b: numero di campioni (batch) in cui ogni variabile ausiliaria è stata classificata come 'ROSSO',' ARANCIONE' o 'GIALLO', applicando il metodo di diagnosi con un nuovo set di variabili ausiliarie.*

In questo caso l'ipotesi di normalità è verificata per 57 distribuzioni di coefficienti di correlazione parziale su 60. Anche se il risultato mostra chiaramente che ad essere identificata come prima causa del PMM è la variabile ausiliaria x_5, che contiene il parametro che è stato modificato, anche la sesta variabile ausiliaria, fortemente correlata con la precedente, viene classificata per un numero significativo di campioni come 'ROSSO' e 'ARANCIONE'.

4.2 Caso 2

Anche se si ottengono risultati molto promettenti considerando un sistema semplificato, in cui vengono analizzate solo le misure disponibili all'inizio e alla fine del processo, è interessante analizzare le possibili estensioni di questo secondo metodo di diagnosi utilizzando un intervallo più ampio dei dati disponibili, che includono le informazioni relative alla dinamica del processo. Infatti, considerando l'evoluzione nel tempo del processo, ci si aspetta di ottenere maggiori informazioni riguardo la natura del PMM e l'effetto di questo sulle variabili del processo coinvolte.

Per facilitare l'analisi, ancora una volta non viene considerata la parte iniziale batch del processo, dove le variabili in uscita sono soggette a forti variazioni, ma si considera solo la fase fed-batch del processo. A tal scopo, viene effettuato un troncamento dei set di dati ($\mathbf{X}_{dati,II}$ e $\mathbf{X}_{dati,M}$) considerando un istante del processo, uguale per tutti i batch, a partire dal quale il processo è sicuramente per tutti in modalità fed-batch, seguendo la logica adottata anche in §3.3. Inoltre, l'analisi viene condotta usando le sole variabili originali, cioè la concentrazione di substrato C_s, di ossigeno disciolto, C_{ox}, di biomassa, C_x, e di penicillina, C_p:

$$x_1 = C_s \qquad x_2 = C_{ox}$$
$$x_3 = C_x \qquad x_4 = C_p \tag{4.4}$$

L'analisi rispetto alle variabili originali permette di ottenere delle informazioni iniziali utili al fine di verificare l'efficacia del trattamento dei dati descritto nel paragrafo seguente utilizzando un set di variabili semplificato.

I set di dati analizzati sono le due matrici di processo e modello, $\underline{\mathbf{X}}_{\text{dati,П}}$ e $\underline{\mathbf{X}}_{\text{dati,M}}$ di dimensioni [N×K×T], dove T è il numero di campionamenti di ogni variabile per gli N batch nel tempo. I campioni dei due set di dati sono generati usando le combinazioni di variabili in entrata definite in §4.1.1.

4.2.1 Trattamento dei dati

La procedura di diagnosi proposta da Rato e Reis (2015) nell'ambito del controllo statistico di processo, prevede che i coefficienti di correlazione assumano una distribuzione normale affinché possano essere definiti dei limiti di confidenza basati sulla deviazione standard della distribuzione. Per generare una distribuzione normale per ogni coefficiente di correlazione parziale, è possibile ricavare un vettore di coefficienti di correlazione parziale a partire da ogni sezione di dimensioni [N×K] della matrice di dati di dimensioni [N×K×T], considerando per ogni variabile e per ogni batch, le misurazioni ad ogni istante di tempo t.

Nel caso studiato è necessario considerare che i coefficienti di correlazione parziale riflettono la dinamica del processo a causa dell'autocorrelazione che caratterizza le variabili originali e della correlazione incrociata tra le diverse variabili nel tempo, per cui non presentano una distribuzione normale.

Per estendere l'applicazione della procedura di diagnosi in questo secondo caso, sono state quindi testate alcune soluzioni basate su un'appropriata trasformazione dei dati che permetta di limitare l'autocorrelazione e correlazione incrociata delle variabili analizzate. Le soluzioni proposte si basano sulla procedura suggerita da Rato e Reis (2014) per risolvere questo tipo di problematiche, che utilizza la decomposizione di Cholesky, adattata da Press *et al*. (2007). Tale decomposizione è stata utilizzata da Rato e Reis (2015) per trasformare opportunamente le variabili misurate di sistemi dinamici continui sottoposti a monitoraggio.

In particolare, considerando una matrice di dati, le cui colonne rappresentano le K variabili misurate e le righe rappresentano le osservazioni, la trasformazione di Cholesky opera una regressione della k-esima variabile rispetto a quelle precedenti, a seconda dell'ordine in cui le variabili compaiono nella matrice dei dati.

Sebbene siano state testate diverse soluzioni per applicare la trasformazione delle variabili originali, in tutti i casi la decomposizione di Cholesky consiste nella fattorizzazione della matrice di covarianza $\mathbf{\Sigma}$ in una matrice triangolare inferiore $\mathbf{L}$ tale che (Press *et al*., 2007):

$$\mathbf{\Sigma} = \mathbf{L} \cdot \mathbf{L}^{\mathrm{T}} \quad . \tag{4.5}$$

Successivamente, la matrice triangolare inferiore può essere utilizzata per ottenere una matrice **U** di variabili decorrelate:

$$\mathbf{U} = \left(\mathbf{X} - \mathbf{m}\right) \cdot \mathbf{L}^{-1} \quad , \tag{4.6}$$

dove **X** è la matrice dei dati e **m** il vettore delle medie delle variabili. La trasformazione operata nella (4.6) tuttavia è valida solo per sistemi lineari in stato stazionario. Per applicare tale trasformazione a sistemi dinamici non lineari, come quello del Caso studio in esame, è stata utilizzata da Rato e Reis (2014) una trasformazione alternativa, in cui si considera una matrice dei dati estesa che include, rispetto all'istante di tempo corrente (0) anche i valori delle variabili calcolate per *j* istanti di tempo precedenti:

$$\tilde{\mathbf{X}} = \left[\mathbf{X}(j) \ldots \mathbf{X}(1)\ \mathbf{X}(0)\right] \quad , \tag{4.7}$$

dove $\mathbf{X}(j)$ è la matrice di dati di dimensione $[N \times K]$ calcolata *j* istanti di tempo precedenti rispetto all'istante attuale. Sulla matrice estesa è applicata la trasformazione di Cholesky (4.6), per ottenere poi un nuovo set di variabili decorrelate rispetto al tempo:

$$\tilde{\mathbf{\Sigma}} = \tilde{\mathbf{L}} \cdot \tilde{\mathbf{L}}^{\mathbf{T}} \tag{4.8}$$

$$\tilde{\mathbf{U}} = \left(\tilde{\mathbf{X}} - \tilde{\mathbf{m}}\right) \cdot \tilde{\mathbf{L}}^{-1} \quad . \tag{4.9}$$

Considerando la matrice $\tilde{\mathbf{U}}$, solo le ultime *K* colonne sono quelle di interesse, e costituiscono i valori delle variabili calcolate all'istante di tempo *t* considerato, decorrelate rispetto a sé stesse e alle altre variabili calcolate negli istanti di tempo precedenti. Allo scopo di applicare tale procedura al set di dati sono state considerate diverse soluzioni, in cui il set di dati di modello è quello ottenuto con il primo PMM parametrico su ($K_l \cdot a$).

4.2.1.1 Soluzione 1

La prima soluzione proposta per analizzare i dati di un sistema dinamico non lineare si basa sulla procedura proposta da Rato e Reis per l'utilizzo dei coefficienti di correlazione parziale per il monitoraggio di sistemi continui. In questo caso viene considerato un solo batch, generato da uno specifico set di condizioni operative, per il quale sono misurate *K* variabili di processo in ogni istante di tempo (da $t = 1$ a $t = T$). I coefficienti di correlazione calcolati in tale set di

dati, generati da una sola combinazione di input riflettono unicamente la variabilità nel tempo delle variabili.

A partire da questo set di misure, una prima possibile trasformazione delle variabili è la seguente:

1. vengono considerate P misurazioni consecutive nel tempo per definire una matrice di *lag* $\mathbf{X}_{lag}$ di dimensioni [P×K];
2. si seleziona un numero L di matrici di *lag* tali per cui, per l=1 $\mathbf{X}_{lag} = \mathbf{X}(t\text{-}PL{:}t\text{-}P(L\text{-}1))$, per l=2 $\mathbf{X}_{lag} = \mathbf{X}(t\text{-}PL{:}t\text{-}P(L\text{-}1))$…e per l=L $\mathbf{X}_{lag} = \mathbf{X}(t\text{-}P{:}t)$. L'insieme delle L matrici forma una matrice estesa $\tilde{\mathbf{X}}$, di dimensioni [P×LK];
3. si calcola la matrice di covarianza $\tilde{\mathbf{\Sigma}}$ matrice estesa $\tilde{\mathbf{X}}$, che viene utilizzata per applicare alla matrice estesa la trasformazione di Cholesky, ricavando una matrice **U** dalle ultime K righe della matrice estesa $\tilde{\mathbf{U}}$. Si ottiene quindi un numero τ di matrici **U**, con $\tau = T/P\text{-}L$+1;
4. per ogni matrice **U** viene calcolato un vettore di coefficienti di correlazione parziale, ottenendo τ elementi per ogni distribuzione di coefficienti; infine viene applicata la procedura diagnostica.

4.2.1.2 Soluzione 2

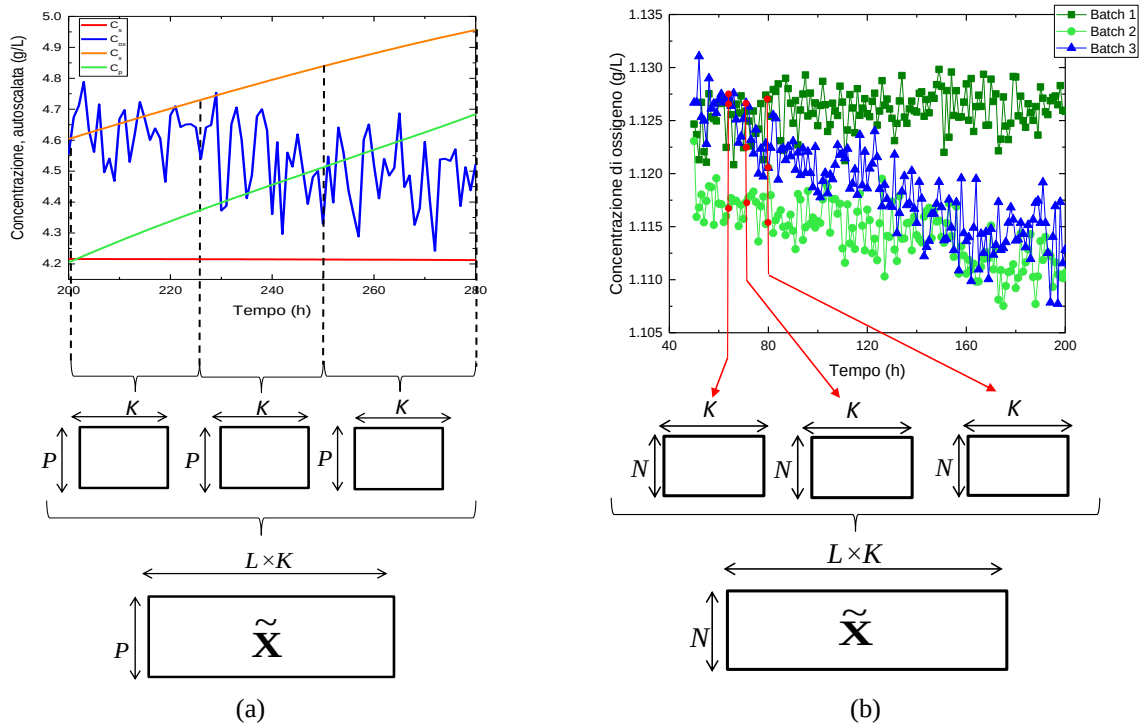
Si considera un numero N di campioni, contraddistinti da diverse condizioni iniziali per i quali vengono misurate K variabili di processo per un numero T di istanti di tempo. Questi dati sono raccolti in un set dinamico di dati **X**. A differenza della prima soluzione, i coefficienti di correlazione sono calcolati considerando non solo la variabilità nel tempo, ma anche le differenti condizioni iniziali (campioni) con cui sono stati generati i valori delle variabili. La seconda procedura proposta si articola come segue:

1. si seleziona un numero L di istanti di tempo successivi (*lag*), per cui sono calcolate $\tau = T - L + 1$ matrici estese di dimensioni [N×LK]:

$$\tilde{\mathbf{X}} = \left[\mathbf{X}(t-L)\ldots\mathbf{X}(t-1)\ \mathbf{X}(t)\right] \tag{4.10}$$

   in cui X($t - j$) è una matrice di *lag* di dimensioni [N×K] calcolata in un istante di tempo interno all'intervallo [($t - L$) t]. Il numero L di *lag* presenti nella matrice estesa dovrebbe essere associato al grado di autocorrelazione e correlazione incrociata delle variabili: tanto più queste sono maggiori, quante più matrici di *lag* dovrebbero essere comprese nella matrice estesa per effettuare la regressione delle variabili;
2. per ogni matrice estesa è calcolata la matrice di covarianza, utilizzata per ottenere una matrice estesa $\tilde{\mathbf{U}}$ di variabili decorrelate rispetto al tempo, secondo le (4.8) e (4.9). Ovvero, la matrice di covarianza $\tilde{\mathbf{\Sigma}}$ è calcolata a partire da una sezione di dimensioni [N×K] di **X**, relativa alle misurazioni delle variabili in tutti i campioni, in un istante di tempo;

3. da ogni matrice estesa è ricavata la matrice **U**, che comprende le sue ultime *K* variabili. Un totale di τ matrici **U**, per ognuna delle quali si può calcolare un vettore di coefficienti di correlazione parziale. L'operazione di trasformazione delle variabili avviene tanto per i dati di modello quanto per quelli di processo, usando per entrambi i set di dati le matrici di covarianza $\tilde{\boldsymbol{\Sigma}}$ i vettori delle medie **m**, calcolati per le τ matrici estese ricavate dai dati di modello.



(a) (b)

Figura 4.4. *(a): soluzione 1. Schema rappresentativo della costruzione della matrice estesa a partire da misurazioni nel tempo delle variabili di un sistema con un unico set di variabili di ingresso. (b): soluzione 2. Schema rappresentativo della costruzione della matrice estesa a partire da misurazioni nel tempo delle variabili di un set di dati con campioni calcolati a partire da differenti combinazioni di ingresso.*

4.2.1.3 Soluzione 3

Assumendo che la trasformazione di ogni matrice estesa debba essere effettuata rispetto ad una matrice di covarianza di popolazione, che rifletta la struttura di correlazione del sistema nelle condizioni operative normali (NOC), in questo caso l'approccio è lo stesso del caso precedente, ma ogni matrice estesa è sottoposta alla trasformazione con una matrice di covarianza mediata su tutte le τ matrici di covarianza. L'uso di matrici di covarianza di *lag* calcolate puntualmente e usate per la trasformazione di Cholesky porta al calcolo di coefficienti di correlazione che riflettono la struttura di correlazione localizzata all'interno della finestra di tempo del *lag*; usando invece un'unica matrice di covarianza di popolazione, $\boldsymbol{\Sigma}_{pop}$, si dovrebbero ottenere dei

coefficienti di correlazione i cui valori siano maggiormente uniformati e il cui andamento nel tempo permetta dunque di dedurre una distribuzione normale.

4.2.2 Risultati ottenuti

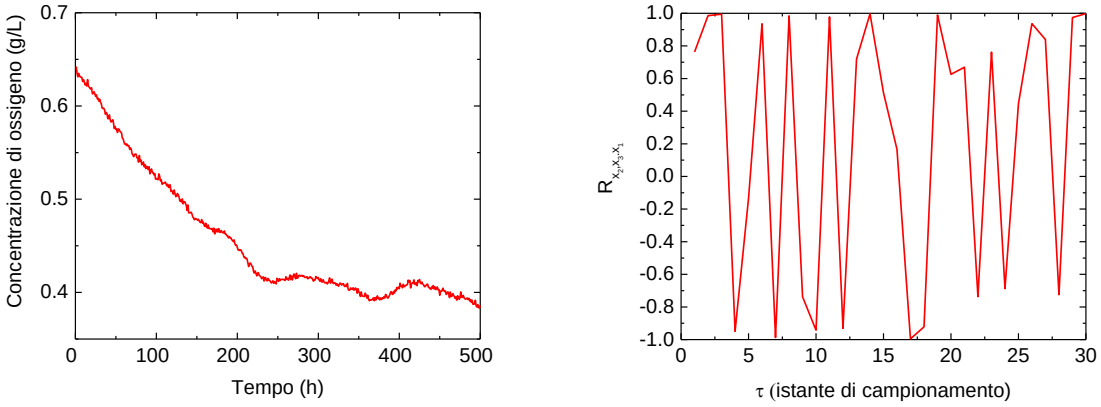
Di seguito vengono riportati i risultati ottenuti implementando i tre approcci proposti. Per ogni esempio, vengono generate una matrice di processo e una di modello, $\underline{\mathbf{X}}_{\Pi}$ e $\underline{\mathbf{X}}_{M}$ di dimensioni $[N \times K \times T]$, in cui vengono considerate solo le variabili originali del processo. L'obiettivo è di confrontare la capacità di ogni tecnica di trasformazione di eliminare le correlazioni delle variabili rispetto al tempo, in modo che i coefficienti di correlazione parziale calcolati risultino distribuiti normalmente, in un intorno rispetto alla media.

4.2.2.1 Risultati ottenuti: soluzione 1

Alcuni problemi emersi nell'applicazione di tale soluzione proposta sono i seguenti.

1. Il tempo di campionamento Δt definito nel simulatore non può essere troppo basso, cioè la frequenza di campionamento non può superare una certa soglia (Δt = 0.2 h). Sebbene l'aumento della frequenza di campionamento permetterebbe di disporre di più elementi nei vettori di coefficienti di correlazione parziale, agevolando così la generazione di una distribuzione di coefficienti parziali normale, ciò impedisce la triangolarizzazione della matrice di covarianza, dato che questa risulta non definita positiva, ovvero presenta autovalori negativi o nulli. In questo caso è dunque necessario ridurre la frequenza di campionamento, a scapito di un numero esiguo di elementi nei vettori dei coefficienti di correlazione (per esempio, per P = 15 istanti di tempo e L = 3 *lag* si ottengono solamente 22 elementi per ogni distribuzione dei coefficienti di correlazione).
2. Un secondo problema è rappresentato dal rumore che alcune variabili presentano nell'andamento nel tempo (in particolare la concentrazione di ossigeno e, in modo meno marcato, la concentrazione di substrato). L'effetto di tale rumore non viene eliminato con la trasformazione delle variabili, influenzando l'evoluzione nel tempo della maggior parte dei coefficienti di correlazione parziale calcolati per le τ matrici $\mathbf{U}$.

In Figura 4.5*a* viene riportato l'andamento della concentrazione di ossigeno nel corso della simulazione nella fase fed-batch del processo. Come si nota, tale valore è affetto da un rumore di misurazione, probabile causa dell'andamento fortemente oscillatorio dei coefficienti di correlazione calcolati (Figura 4.5*b*).



(a) (b)

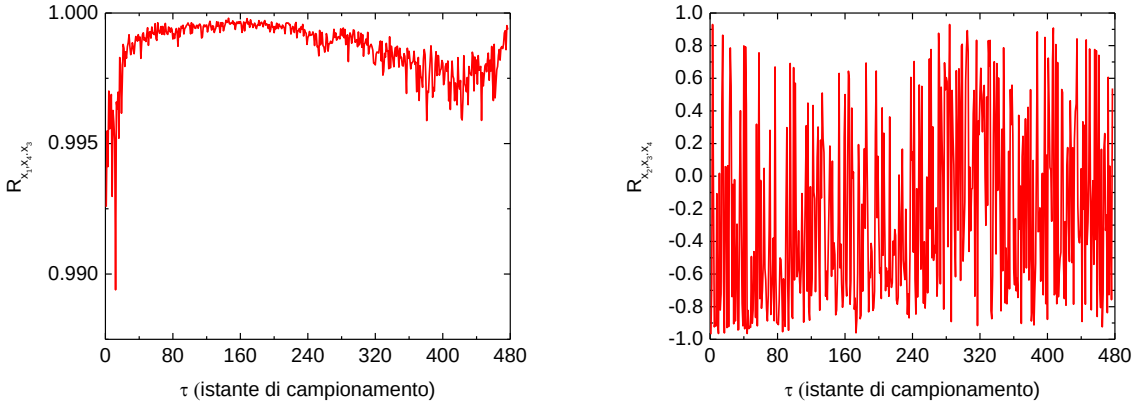
Figura 4.5. Soluzione 1. *(a): andamento nel tempo della concentrazione di ossigeno disciolto, usando variabili originali non sottoposte alla trasformazione; (b); andamento, per gli istanti di campionamento definiti dalla trasformazione delle variabili, del coefficiente di correlazione parziale in cui concentrazione di ossigeno e di biomassa sono correlate, controllando la concentrazione di substrato, per i dati di modello.*

4.2.2.2 Risultati ottenuti: soluzione 2

In questo caso, i campioni sono costituiti da N batch simulati con diverse condizioni iniziali. Scelto un numero L di istanti di tempo successivi rispetto a cui effettuare la regressione, è disponibile di un numero maggiore di istanti di tempo τ per definire la distribuzione dei coefficienti di correlazione parziale, rispetto a quelli che si ottengono utilizzando il primo approccio, in cui le misurazioni nel tempo costituivano i campioni delle matrici di *lag*.

Neppure in questo caso si ottengono distribuzioni normali dei coefficienti di correlazione parziale. Inoltre, la maggior parte di essi presenta andamenti nel tempo fortemente oscillatori, con picchi tra i valori limite estremi (–1 e 1). In Figura 4.6 sono riportati due esempi antitetici degli andamenti nel tempo osservati per due coefficienti di correlazione parziale. Il primo (Figura 4.7a) rappresenta l'andamento desiderato, in quanto i valori nel tempo sono compresi in un intervallo molto ristretto, mentre il secondo (Figura 4.7b), presenta un andamento molto oscillatorio, in un intervallo ampio di valori. Dal punto di vista ingegneristico, un tale andamento non ha significato fisico, perché indica che la correlazione tra due variabili (eliminando il contributo di una terza) cambia repentinamente nel tempo, sia in valore assoluto che in verso. Un'analisi puntuale dei coefficienti di correlazione generati rivela che la maggior parte dei coefficienti di correlazione, in cui la concentrazione di ossigeno compare come variabile correlata e anche come variabile controllata, presentano andamenti simili a quelli riportati in Figura 4.6*b*. Al contrario, invece, se questa variabile non è presente, gli andamenti dei coefficienti presentano un andamento pressoché costante nel tempo, come in Figura 4.6*a* (dove infatti si riporta l'andamento del coefficiente in cui concentrazione di substrato e penicillina sono correlati, controllando la concentrazione di biomassa).

Per risolvere tale problema, è stato modificato il numero di *lag* per il calcolo delle matrici estese assumendo che, aumentando il numero di matrici presenti nella matrice estesa (cioè includendo un numero maggiore di istanti di tempo considerati per la regressione delle variabili), l'effetto di decorrelazione rispetto al tempo delle variabili interessi finestre di tempo maggiori. Inoltre, tale scelta dovrebbe contribuire a limitare i problemi dovuti dalla presenza di un forte rumore nelle variabili di concentrazione di ossigeno (soprattutto) e substrato.



(a) (b)

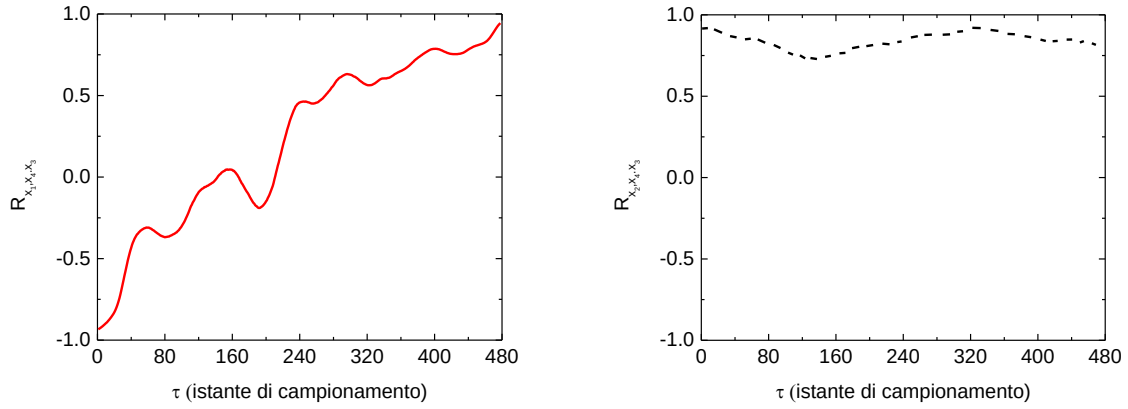
Figura 4.6. Soluzione 2. *(a): andamento, per gli istanti di campionamento definiti dalla trasformazione delle variabili, del coefficiente di correlazione parziale in cui concentrazione di substrato e di penicillina sono correlate, controllando la concentrazione di biomassa, per i dati di modello.; (b); andamento del coefficiente di correlazione parziale in cui concentrazione di ossigeno e di biomassa sono correlate, controllando la concentrazione di penicillina per i dati di modello.*

Tuttavia, non si sono riscontrati rilevanti miglioramenti neppure applicando questa soluzione: è da sottolineare che oltre un certo numero di matrici di *lag*, non è possibile l'applicazione del metodo di Cholesky, poiché la matrice di covarianza della matrice estesa ottenuta non risulta essere definita positiva.

4.2.2.3 Risultati ottenuti: soluzione 3

In questo caso vengono utilizzati gli stessi set di dati usati per il secondo approccio, ma calcolando, una volta definito il numero L di *lag*, un'unica matrice di covarianza estesa, media di tutte le τ matrici di covarianza, che riflette una struttura di correlazione del modello mediata nel tempo, in modo da ottenere valori maggiormente uniformi dei coefficienti di correlazione.
Dai risultati emerge che tale approccio non permette di eliminare l'effetto della dinamica del processo sui coefficienti di correlazione; tuttavia, non si osservano più gli andamenti fortemente oscillatori ottenuti nei due casi precedenti, ma un'evoluzione nel tempo ben definita. In Figura 4.7 è riportato l'andamento, a diversi istanti di campionamento, di due coefficienti di correlazione: come si può vedere per il coefficiente di correlazione tra substrato e penicillina, controllando la biomassa (Figura 4.7*a*), l'andamento nel tempo, sebbene non presenti forti

oscillazioni ad alta frequenza, non permette di ottenere una distribuzione normale. Il coefficiente di correlazione tra concentrazione di ossigeno e di penicillina, controllando la concentrazione di biomassa, presenta invece un andamento approssimativamente costante nel tempo come nel caso del coefficiente di correlazione tra concentrazione di ossigeno e di penicillina, controllando la concentrazione di substrato. Esso presenta inoltre una distribuzione fortemente tendente alla normalità.



(a) (b)

Figura 4.7. Soluzione 3. *(a) andamento, per gli istanti di campionamento definiti dalla trasformazione delle variabili, del coefficiente di correlazione parziale in cui concentrazione di substrato e di penicillina sono correlate, controllando la concentrazione di biomassa, per i dati di modello; (b) andamento del coefficiente di correlazione parziale in cui concentrazione di ossigeno e di penicillina sono correlate, controllando la concentrazione di biomassa, per i dati di modello.*

Per verificare eventuali miglioramenti dei profili nel tempo dei coefficienti di correlazione, è quindi stata eseguita un'analisi di sensitività, rispetto al numero L di matrici di *lag* incluse in una matrice estesa e al tempo di campionamento Δt scelto per la simulazione dei dati, in modo da valutare se la scelta di diversi valori per questi due parametri permetta di ottenere coefficienti di correlazioni parziale che presentano una dinamica meno accentuata e quindi un andamento maggiormente costante nel tempo.

Le due analisi di sensitività sono le seguenti:

- il numero L di matrici di *lag* viene aumentato mantenendo costante la frequenza di campionamento (intervallo di campionamento Δt = 0.5 h);
- l'intervallo di campionamento Δt viene aumentato con un numero costante di matrici di *lag* (L = 10), ovvero considerando dei dataset con differenti numeri di campionamenti.

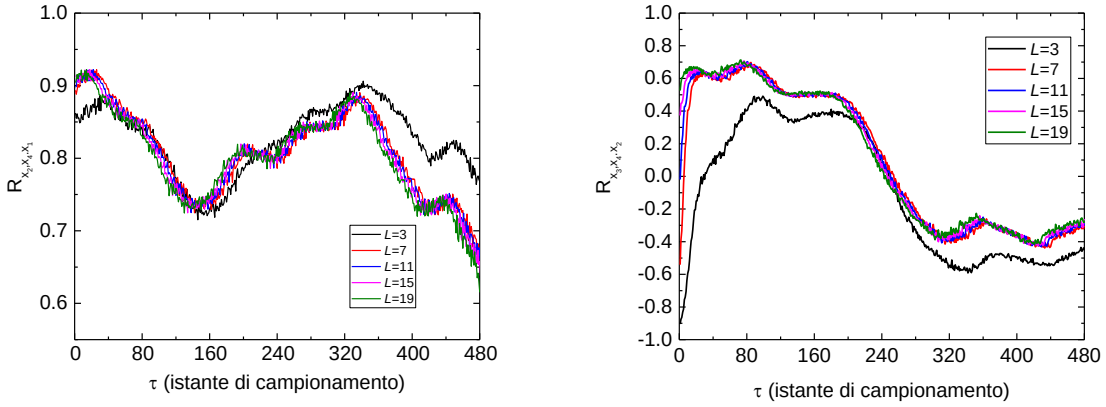
Sia nel primo che nel secondo caso, ci si aspetta che aumentando il numero di matrici di *lag* o diminuendo la frequenza di campionamento in modo che la regressione delle variabili coinvolga un maggiore numero di istanti di tempo, la rimozione dell'autocorrelazione e correlazione incrociata delle variabili risulti più efficace.

Sono stati considerati due coefficienti di correlazione parziale: il primo, calcolato tra concentrazione di ossigeno e di penicillina, controllando quella di substrato, presenta

un'evoluzione nel tempo in un intervallo di valori ristretto (tra circa 0.9 a 0.65); il secondo, calcolato tra concentrazione di biomassa e penicillina, controllando quella di ossigeno, presenta una chiara evoluzione e variazione nel tempo (tra –1 a 0.6).

In Figura 4.8 è mostrata l'analisi di sensitività degli andamenti nel tempo dei coefficienti di correlazione rispetto al numero di matrici di *lag* incluse nelle matrici estese.

Si può notare solo un parziale miglioramento (cioè un andamento nel tempo maggiormente costante) passando da tre a sette matrici di *lag* incluse nella matrice estesa; per valori superiori di *L* l'andamento nel tempo dei coefficienti di correlazione di fatto non cambia, mostrando quindi una scarsa sensitività della trasformazione oltre un certo numero di istanti di tempo compresi nella regressione delle variabili.

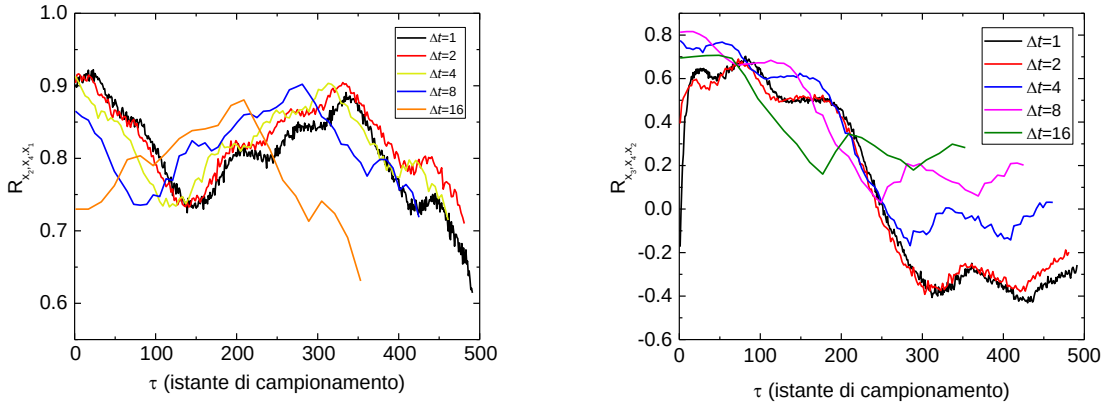


(a) (b)

Figura 4.8. *Soluzione 3: andamenti dei coefficienti di correlazione al variare del numero di matrici di lag considerate per la regressione. (a): andamento, per gli istanti di campionamento definiti dalla trasformazione delle variabili, del coefficiente di correlazione in cui concentrazione di ossigeno e di penicillina sono correlate, controllando la concentrazione di substrato, per i dati di modello; (b): andamento del coefficiente di correlazione in cui concentrazione di biomassa e di penicillina sono correlate, controllando la concentrazione di ossigeno, per i dati di modello.*

Considerando l'analisi di sensitività rispetto alla frequenza di campionamento, è necessario considerare che, aumentando l'intervallo di campionamento, diminuisce anche il numero di campioni nel tempo e il numero di istanti di tempo τ, anche se il tempo nominale della simulazione rimane invariato: è per tale motivo che gli andamenti dei coefficienti di correlazione in Figura 4.9 si fermano ad istanti di tempo diversi.

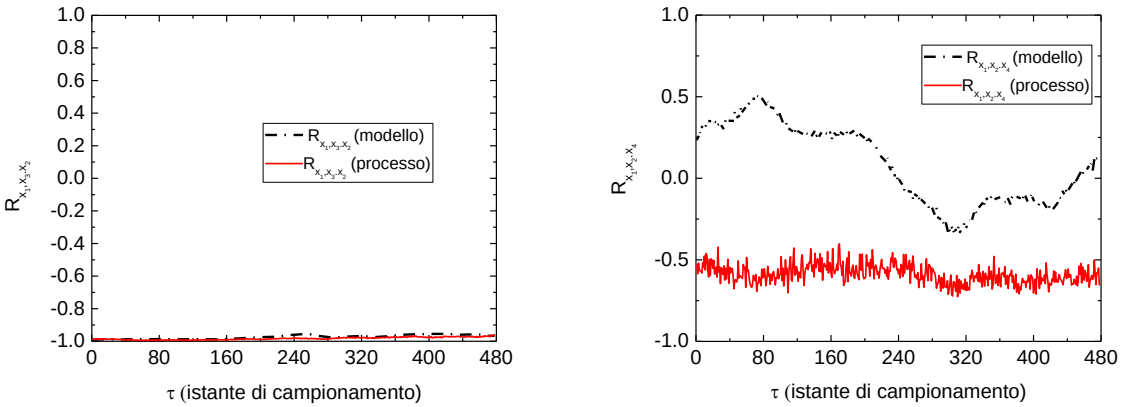
L'effetto della diminuzione della frequenza di campionamento sull'andamento, nel tempo, per il coefficiente di correlazione parziale in Figura 4.9*a*, è molto ridotto. Aumentando Δ*t* fino a 16 h, rispetto al valore nominale (0.5 h), i profili dei coefficienti di correlazione non sembrano perdere la loro dinamica. Un diverso risultato si osserva in Figura 4.9*b*: per il coefficiente di correlazione tra concentrazione di biomassa e di penicillina, controllando la concentrazione di ossigeno, che presentava una dinamica accentuata, un aumento del tempo di campionamento permette di ottenere un'evoluzione nel tempo meno influenzata dalla dinamica del processo.



(a) (b)

Figura 4.9. *Andamenti dei coefficienti di correlazione al variare del valore dell'intervallo di campionamento. (a): andamento, per gli istanti di campionamento definiti dalla trasformazione delle variabili, del coefficiente di correlazione in cui concentrazione di ossigeno e di penicillina sono correlate, controllando la concentrazione di substrato, per i dati di modello; (b): andamento del coefficiente di correlazione in cui concentrazione di biomassa e di penicillina sono correlate, controllando la concentrazione di ossigeno, per i dati di modello.*

Osservando gli andamenti dei coefficienti di correlazione nel tempo, calcolati utilizzando la procedura descritta in §4.2.1.3, si deduce che la trasformazione applicata senza usare matrici di covarianza puntualmente calcolate nelle finestre di *lag* non ha effetto nel rimuovere la correlazione nel tempo delle variabili. Ciò tuttavia può risultare utile nel caso si voglia utilizzare un altro approccio nella diagnosi del PMM, e cioè il controllo degli andamenti dei coefficienti di correlazione nel tempo tramite limiti di confidenza puntualmente definiti nel tempo.



(a) (b)

Figura 4.10. *(a): andamento, per gli istanti di campionamento definiti dalla trasformazione delle variabili, del coefficiente di correlazione parziale in cui concentrazione di substrato e di biomassa sono correlate, controllando la concentrazione di ossigeno, per i dati di modello e di processo; (b); andamento del coefficiente di correlazione parziale in cui concentrazione di substrato e di ossigeno sono correlate, controllando la concentrazione di penicillina, per i dati di modello e di processo.*

In Figura 4.10 sono riportati gli andamenti nel tempo di due coefficienti di correlazione in cui la concentrazione di ossigeno è controllata (Figura 4.10*a*) e associata (Figura 4.10*b*), utilizzando il set di dati di modello e quello di processo. Si nota che, controllando la concentrazione di ossigeno, variabile particolarmente legata al PMM, i coefficienti di correlazione calcolati per i dati di processo hanno valori nel tempo molto simili a quelli di modello. Quando invece l'ossigeno compare come variabile associata si osserva una chiara deviazione dei valori dei coefficienti di correlazione del set di dati di modello rispetto a quelli di processo. Questo risultato suggerisce che una tecnica diagnostica alternativa può essere implementata usando delle carte di controllo.

4.3. Conclusioni

La procedura di diagnosi del PMM applicata a un set di dati in cui si considerano solo i valori delle variabili in entrata al processo e i valori finali delle variabili in uscita, permette di identificare la causa del PMM, sfruttando la definizione delle variabili ausiliarie: l'analisi offre delle informazioni sulla discrepanza tra dati di processo e di modello. Un accorgimento da adottare nell'applicazione della diagnosi è la definizione di un adeguato set di variabili ausiliarie. In questo senso, la rielaborazione delle variabili ausiliarie a partire dai termini presenti nelle equazioni di modello, come visto in §4.1.3, effettivamente rende il metodo più robusto qualora non si riesca a eliminare il contributo al PMM di variabili ausiliarie fortemente correlate a quella responsabile del PMM, facendo uso dei coefficienti di correlazione parziale. Un'altra possibile soluzione da adottare è quella di usare coefficienti di correlazione di ordine superiore al primo.

La normalità delle distribuzioni dei coefficienti di correlazione, richiesta per la definizione di limiti di confidenza appropriati, è il maggiore problema inerente all'applicazione della procedura di diagnosi. Per agevolare una distribuzione normale dei coefficienti è stato necessario considerare un set di variabili di input del processo distribuiti normalmente. Tuttavia non si riesce, in questo caso, ad ottenere la normalità di tutti i coefficienti di correlazione parziale.

Per poter in modo implementare la procedura di diagnosi applicata in §4.1 a set di dati dinamici sono state utilizzate tre soluzioni, per trasformare le variabili in modo da ottenere dei profili dei coefficienti di correlazione parziale, nel tempo, in modo da poterne ricavare delle distribuzioni normali. In tal modo si sfrutta la natura dinamica del processo descritto nel caso studio per ottenere maggiori informazioni sull'entità e l'evoluzione del PMM nel corso del tempo.

La prima soluzione indagata non dà risultati soddisfacenti, in quanto la presenza di rumore sulle misure delle variabili influenza fortemente la distribuzione dei coefficienti di correlazione. Lo stesso problema è riscontrato applicando la seconda soluzione, che non dà risultati utili al fine di ottenere vettori di coefficienti di correlazione parziale distribuiti normalmente. Una possibile

soluzione a tale problema è la filtrazione delle variabili che contengono eccessivo rumore, in modo da ottenere un andamento tendenzialmente monotono delle variabili nel tempo ed evitare andamenti oscillatori dei coefficienti di correlazione.
Il terzo approccio utilizzato non permette, di fatto, di eliminare la dinamica dei coefficienti di correlazione. Applicando l'ultima soluzione proposta per la trasformazione delle variabili, ottenendo coefficienti con una dinamica ben definita e non oscillatoria, è emerso che una possibile applicazione futura è l'implementazione di carte di controllo sui coefficienti di correlazione per la diagnosi del PMM.

Conclusioni

In questa Tesi è stato affrontato il problema della diagnosi del possibile disallineamento (*process/model mismatch*, PMM) tra dati di processo e le relative predizioni di un modello a principi primi. Le tecniche per la diagnosi del PMM adottate sono due: quella elaborata da Meneghetti *et al.* (2014), basata sull'analisi dei residui di una matrice ricavata da dati di processo proiettata sullo spazio di un modello PCA costruito su una matrice ricavata da dati del modello a principi primi; e una tecnica di diagnosi basata sul confronto dei coefficienti di correlazione parziale tra il set di dati di processo e quello di modello, adattata dagli studi di Rato e Reis (2015) sul monitoraggio di sistemi continui.

Tali tecniche sono state applicate per la diagnosi di un modello a principi primi che descrive un processo di fermentazione per la produzione di penicillina, sviluppato da Birol *et al.* (2002). Le misure di processo sono state generate *in silico* utilizzando tale modello, considerando i parametri nominali; le predizioni del modello sono state invece generate usando lo stesso modello a principi primi, ma forzando la presenza di un PMM tramite la modifica di due parametri differenti. In entrambi i casi le variabili originali sono state combinate in modo non lineare con i parametri del modello per formare nuove variabili, dette variabili ausiliarie, in modo da poter estrarre, da un confronto tra le strutture di correlazione dei due set di dati, informazioni sulla causa del PMM.

La prima metodologia è stata applicata sia considerando solo gli ingressi al processo e le misure finali delle variabili, sia considerando un set di dati esteso per analizzare la dinamica del processo. Nel secondo caso, sebbene in generale l'analisi dell'intera evoluzione del processo si possa rivelare dispendiosa per problemi inerenti a misurazioni delle variabili nel tempo (per esempio traiettorie non allineate, dati mancanti), nondimeno essa permette di acquisire maggiori informazioni relative alla causa del PMM e alla sua evoluzione nel tempo.

Mentre, in un primo esempio di PMM parametrico, la metodologia di diagnosi elaborata da Meneghetti *et al.* (2014) riconosce correttamente la causa del disallineamento tra i dati di processo e quelli del modello, in un secondo esempio essa non è in grado di discriminare correttamente quale tra le variabili ausiliarie sia la maggior responsabile del PMM; un miglioramento nella diagnostica si osserva tuttavia se, invece che impiegare unicamente dati relativi all'inizio e alla fine di un batch, vengono impiegate anche le relative traiettorie temporali. La scelta delle variabili ausiliarie è fatta in base alla struttura di equazioni del modello a principi primi: un obiettivo futuro inerente al miglioramento di questa tecnica è stabilire un criterio opportuno per automatizzare la scelta delle combinazioni di parametri e variabili del modello. Inoltre, la maggiore limitazione di questa tecnica diagnostica è la

presenza di una forte correlazione tra le variabili originali del modello, esistente anche tra le variabili ausiliarie che sono generate delle combinazioni delle stesse.
Una soluzione a tale problema viene suggerita con l'applicazione della seconda metodologia, attraverso un confronto delle strutture di correlazione tra i dati di processo e quelli di modello ottenuto impiegando i coefficienti di correlazione parziale. Nel caso una specifica variabile (non nota a priori) sia associata al PMM, si può esplicitarne il contributo escludendo quello di altre variabili ad essa correlate, che non sono però responsabili del PMM. La diagnosi avviene costruendo dei limiti di confidenza delle distribuzioni normalizzate dei coefficienti di correlazione parziale, calcolati dai dati di modello, e confrontandoli con le distribuzioni dei coefficienti di correlazione calcolati dai dati di processo.
La diagnosi, applicata a set di dati costituiti dagli ingressi e dagli stati finali delle variabili, offre nuovamente dei buoni risultati per il primo esempio di PMM; nel secondo caso, è necessario riformulare le variabili ausiliarie per identificare correttamente la causa del PMM. Ciò suggerisce di utilizzare, in futuro, coefficienti di correlazione parziale di grado superiore (ovvero in cui si esplicita la correlazione tra due variabili, eliminando il contributo di più variabili terze, invece di una sola), per ottenere una maggiore decorrelazione delle variabili.
Per applicare la seconda metodologia di diagnosi ad un set di dati dinamico, è necessario eliminare l'autocorrelazione delle variabili nel tempo e la correlazione incrociata con altre variabili al fine di ottenere andamenti dei coefficienti di correlazione parziale da cui possano essere ricavate distribuzioni normali. A tal scopo sono state suggerite tre diverse soluzioni che, sebbene non siano in grado di favorire la generazione di distribuzioni normali dei coefficienti di correlazione parziale, offrono comunque una buona base di partenza per l'elaborazione di diverse tecniche di diagnosi in grado di sfruttare l'evoluzione nel tempo dei coefficienti di correlazione parziale per elaborare un metodo di diagnosi basato sull'implementazione di opportune carte di controllo.

Appendice A

Scale up del bioreattore

La generazione del PMM viene considerata conseguente a un'operazione di *scale up* del reattore a partire da una scala di laboratorio ad una scala pilota. Di seguito è mostrata la procedura di dimensionamento alla scala superiore.

A.1 Procedura adottata

A partire dalla scala originale di 100 L del simulatore, viene effettuato uno *scale-down* del reattore a 10 L. Tale operazione simula il processo effettivo di *scale-up* di un processo di fermentazione da una scala di laboratorio a una scala di impianto pilota.

Una volta che un processo è realizzato con successo in scala di laboratorio, i valori delle variabili operative e delle proprietà fisiche del processo sono noti o possono essere misurati. Il processo viene poi condotto e ottimizzato in reattori di dimensione maggiore, la scala pilota, dove le condizioni operative e il regime fluidodinamico sono simili a quelli della scala industriale.

Il rapporto di *scale-up* è solitamente 1:10, anche se rapporti inferiori riducono il rischio di comportamenti inaspettati del processo a scale maggiori. Il metodo utilizzato per lo *scale-up* si basa su regole generiche.

I criteri di scale-up secondo indicazioni generali più utilizzati, e le loro percentuali di utilizzo nell'industria della fermentazione, sono: potenza specifica, P/V, costante (30%); velocità tangenziale dell'agitatore, πNT, costante (20%); coefficiente volumetrico di trasporto di massa per l'ossigeno, $K_l a$, costante (30%).

Dato che il modello di Birol *et al.* utilizza l'(2.9) per il calcolo di $K_l a$, dove non si fa riferimento alle caratteristiche reologiche del liquido, e alla geometria del reattore, lo *scale-down* viene eseguito facendo riferimento a una regola generale presentata da Garcia-Ochoa e Gomez (2008), adattata da Oldshue (1966), mantenendo costante il rapporto P/V (Tabella A.1).

Si può ricavare in tal modo il coefficiente volumetrico di trasporto di massa per l'ossigeno, da cui è possibile trovare la portata di aria in ingresso, usando la (2.9). Per la portata di alimentazione si usa una proporzione con il rapporto tra i volumi.

Per quanto riguarda i parametri di controllo, i guadagni vengono scalati usando la medesima regola, e poi tutti i parametri (compresi i guadagni già modificati) vengono sottoposti a un *tuning* basandosi sugli errori ottenuti in simulazione, ovvero un *tuning* sul campo.

Tabella A.1. *Differenti criteri di scale-up per bireattori (adattati da Oldshue, 1986).*

Variabile	Valore di riferimento al volume di modello (2 L)	Valori di riferimento alla scala pilota (20 L)			
		Criterio di *scale-up*			
		$P/V=C$	$\pi NT=C$	$Re=C$	$K_La=C$
T	1.0	2.14	2.14	2.14	2.14
P	1.0	10.0	4.80	0.50	13.8
P/V	1.0	1.0	0.48	0.05	1.38
N	1.0	0.60	0.47	0.22	0.67
NT	1.0	1.28	1.0	0.47	1.43
Re	1.0	2.75	2.15	1.0	3.07
K_La	1.0	0.77	0.55	0.19	1.0

Appendice B

Codici di calcolo

In Tabella B.1 e B.2 sono riportati i codici di calcolo e i *file* da cui sono immessi i relativi dati di input.

Tabella B.1. *Codici di calcolo per il capitolo 3.*

Codici di calcolo	Dati di input	Descrizione
`Closedlooppilot.m`	-	Codice per la generazione dei dati di input per un set di dati utilizzati per costruire un modello PCA
`Reducesamples.m`	`Result100.mat`	Codice per la selezione di un numero ridotto di campioni a partire da un set di dati di partenza
`MRLR2DKLA.m`	`Datasetpilot.mat`	Codice per la diagnosi del PMM tramite l'analisi dell'indice MRLR: primo esempio
`MRLR2DYSX.m`	`Datasetpilot.mat`	Codice per la diagnosi del PMM tramite l'analisi dell'indice MRLR: secondo esempio
`Generazionematrici3d.m`	`Datasetpilot.mat`	Codice per la generazione di set di dati dinamici
`MRLR3DKLA.m`	`Dati3dPKLA.mat` `Dati3dMKLA.mat`	Codice per la diagnosi del PMM in dinamico: primo esempio
`MRLR3DYSX.m`	`Dati3dPYSX.mat` `Dati3dMYSX.mat`	Codice per la diagnosi del PMM in dinamico: secondo esempio

Tabella B.2. *Codici di calcolo per il capitolo 4.*

Codici di calcolo	**Dati di input**	**Descrizione**
`Normsaples.m`	–	Codice per la generazione di un set di dati di input distribuiti normalmente
`NuovosamplingKLA.m`	`DSF.m`	Codice per la generazione di set di dati tridimensionali per il primo esempio di PMM
`NuovosamplingYSX.m`	`DSF.m`	Codice per la generazione di set di dati tridimensionali per il secondo esempio di PMM
`DiagnosiKLA2D.m`	`DS1KLA.m` `DS2KLA.m`	Codice per la diagnosi del PMM in seti di dati costituiti da entrate e valori finali delle uscite del processo: primo esempio
`DiagnosiYSX2D.m`	`DS1YSX.m` `DS2YSX.m`	Codice per la diagnosi del PMM in seti di dati costituiti da entrate e valori finali delle uscite del processo: primo esempio
`Datasetdinamico`	`DSF.m`	Codice per la generazione di un set di dati dinamico
`Cholratoreis.m`	`DSF.m`	Codice per la trasformazione delle variabili in un set di dati dinamico: prima soluzione
`Cholsecondasol.m`	`Xpkla.m` `Xmkla.m`	Codice per la trasformazione delle variabili in un set di dati dinamico: seconda soluzione
`Cholterzasol.m`	`Xpkla.m` `Xmkla.m`	Codice per la trasformazione delle variabili in un set di dati dinamico: terza soluzione

Riferimenti bibliografici

Birol G., C. Udney e A. Cinar (2002). A modular simulation package for fed-batch fermentation: penicillin production. *Comp.Chem. Eng,*, **26**, 1553-1565.

Burnham, A.J., R. Viveros e J.F. MacGregor (1996). Frameworks for latent variable multivariate regression. *J. Chemom.*, **10**, 31-45.

Eriksson L., E. Johansson., N. Kettaneh-Wold e S. Wold. *Multi and Megavariate Data Analysis: Principles and Applications*; Umetrics: Umea, 2001.

Garcia-Ochoa F. e E. Gomez (2009). Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnol. Adv.*, **27**, 153-176.

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

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psycol.*, **24**, 417-441.

Baiiley J.E., D.F. Ollis (1986). *Biochemical Engineering Fundamentals*. New York: McGraw Hill.

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

López-Negrete de la Fuente, R., S. García-Muñoz e L.T. Biegler (2010). An efficient nonlinear programming strategy for PCA models with incomplete data sets. *J. Chemom.*, **24**, 301-311

Mardia, K. V., J.T Kent e J. M. Bibby. *Multivariate Analysis;* Academic Press: London, UK, 1979.

Meneghetti N, P. Facco, S. Bermingham, D. Slade, F. Bezzo e M. Barolo (2015). First-Principles Model Diagnosis in Batch Systems by Multivariate Statistical Modeling. *Computer Aided Chemical Engineering*, 37, *12th International Symposium on process Systems Engineering and 25th European Symposium on Computer Aided process Engineering* (K.V. Garnaey, J.K. Huusom, R. Ganu, Eds.), Elsevier, Amsterdam (The Nederlands) 437-442.

Meneghetti N., P. Facco, F. Bezzo e M. Barolo (2014). A Methodology to diagnose Process/Model Mismatch in First Principles Models. *Ind. Eng. Chem. Res.*, **53**, 14002-14013.

Montgomery, D.C. (2005b). *Introduction to statistical quality control. 5th edition.* John Wiley & Sons, Inc., New York (U.S.A.).

Nomikos, P. e J.F. MacGregor (1994). Monitoring batch processes using multiway principal component analysis. *AIChE J.*, **40**, 1361-1375.

Oldshue J.Y. (1966) Fermentation mixing scale-up techniques. *Biotechnol. Bioeng.*; VIII,3-24.

Press W.H., S.A. Teukolsky, W.T. Vetterling e B.P. Flannery (2007). *Numerical recipes: the art of scientific computing*. 3rd ed. New York: Cambridge University Press.

Qin S.J. (2003). Statistical process monitoring: basics and beyond. *J. Chemom,* **17**, 480-502

Bajpai R. e M. Reuss (1980). A mechanistic model for penicillin production. *J. Chem Technol. Biotechnol.* **30**, 330-344.

Rato T. J. e Marco S. Reis (2014). Sensitivity enhancing transformations for monitoring the process correlation structure. *J.Process Control*, **24**, 905-915.

Rato T. J. e Marco S. Reis (2014). Non-casual data driven monitoring of the process correlation structure: A comparison study with new methods. *Comp. Chem Eng.,* **71**, 307-322.

Rato T. J. e Marco S. Reis (2015). On-line process monitoring using local measures of association: Part I – Detection Performance. *Chemom. Intell. Lab. Syst.***142**, 255-264.

Rato T. J. e Marco S. Reis (2015). On-line process monitoring using local measures of association. Part II: Design issues and fault diagnosis. *Chemom. Intell. Lab. Syst.* **142**, 266-275.

Pirt S. e R. Righelato (1967). Effect of growth rate on the synthesis of penicillin by *penicillium Chrysogenum* in batch and chemostat cultures. *Applied Microbiology*, **15**, 1284-1290.

Tomba E. (2013). Latent variable modelling approaches to assist the implementation of quality-by-design paradigms in pharmaceutical development and manufacturing. *Università degli Studi di Padova.*

Van't Riet K. (1979). Review of measuring methods and nonviscous gas–liquid mass transfer in stirred vessels. *Ind Eng Chem Process Des Dev,* **18**, 357–364

Whitman W.G. (1923). Preliminary experimental confirmation of the two-film theory of gas absorption. *Chem Metall Eng*; **29**,146–149

Wise, B.M., N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig e R. Scott Koch (2006). *PLS_Toolbox Version 4.0 for use with MATLAB™*. Eigenvector Research, Inc., Wenatchee, WA (U.S.A.).

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics*, **20**, 397-405.

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

Siti web

http://simulator.iit.edu/web/pensim/bground.html (ultimo accesso: 10/04/2016)

Ringraziamenti

Desidero ringraziare il Professor Barolo per avermi dato un'importante opportunità di apprendimento nello svolgimento del lavoro di Tesi. Un ringraziamento speciale va a Natascia, per la sua disponibilità senza riserve e i consigli di cui voglio fare tesoro.
Ringrazio anche tutti gli amici del Cape Lab per i mesi passati in compagnia.
Grazie ai miei genitori, alla mia famiglia e ai miei amici, che mi sostengono con amicizia, amore, affetti.
Grazie a te.