UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Ingegneria Industriale

SCUOLA DI DOTTORATO DI RICERCA IN INGEGNERIA INDUSTRIALE
INDIRIZZO: INGEGNERIA CHIMICA
CICLO XXV

LATENT VARIABLE MODELING APPROACHES TO ASSIST THE IMPLEMENTATION OF QUALITY-BY-DESIGN PARADIGMS IN PHARMACEUTICAL DEVELOPMENT AND MANUFACTURING

Direttore della Scuola: Ch.mo Prof. Paolo Colombo
Coordinatore d'indirizzo: Ch.mo Prof. Alberto Bertucco
Supervisore: Ch.mo Prof. Massimiliano Barolo

Dottorando: Emanuele Tomba

Foreword

The realization of the work included in this Dissertation involved the intellectual and financial support of many people and institutions, to whom the author is very grateful.

Most of the research activity that led to the results reported in this Dissertation has been carried out at CAPE-Lab, Computer-Aided Process Engineering Laboratory, at the Department of Industrial Engineering of the University of Padova (Italy), under the supervision of Prof. Massimiliano Barolo and Prof. Fabrizio Bezzo. Part of the work was carried out at Pfizer Worldwide R&D, Groton, CT (U.S.A) during a 6-month stay under the supervision of Dr. Salvador García-Muñoz, and part represents a collaboration with Dr. Simeone Zomer and Dr. John Robertson from GlaxoSmithKline, Harlow (U.K.). Furthermore, a segment of the reported experimental results have been obtained in collaboration with Prof. Daniele Marchisio and Prof. Antonello A. Barresi from the Department of Applied Science and Technology of Politecnico di Torino (Italy).

Financial support to this study has been provided by the University of Padova and by "Fondazione Ing. Aldo Gini", Padova (Italy).

All the material reported in this Dissertation is original, unless explicit references to studies carried out by other people are indicated. In the following, a list of the publications stemmed from this project is reported.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS (published or in press)

Tomba, E., M. De Martin, P. Facco, J. Robertson, S. Zomer, F. Bezzo and M. Barolo (2013). General approach to aid the development of continuous pharmaceutical processes using multivariate statistical modeling – An industrial case study. *Int. J. Pharm.*, in press. DOI: 10.1016/j.ijpharm.2013.01.018.

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

Tomba, E., P. Facco, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Combining fundamental knowledge and latent variable techniques to transfer process monitoring models between plants. *Chemom. Intell. Lab. Syst.*, **116**, 67-77.

Facco, P., E. Tomba, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Transfer of process monitoring models between different plants using latent variable techniques. *Ind. Eng. Chem. Res.*, **51**, 7327-7339.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS (submitted or in preparation)

Tomba E., M. Barolo and S. García-Muñoz (2013). *In-silico* product formulation design through latent variable model inversion. *In preparation*.

Facco, P., M. Largoni, E. Tomba, F. Bezzo and M. Barolo (2013). Transfer of process monitoring models between plants: batch systems. *In preparation*.

Tomba E., N. Meneghetti, P. Facco, T. Zelenkova, D.L. Marchisio, A.A. Barresi, F. Bezzo and M. Barolo (2013). Product transfer between different plants through latent variable model inversion. Submitted to *AIChE J.*

Tomba, E., P. Facco, F. Bezzo and S. García-Muñoz (2012). Exploiting historical databases to design the target quality profile for a new product. Submitted to *Ind. Eng. Chem. Res.*

CONTRIBUTIONS IN REFEREED CONFERENCE PROCEEDINGS (published or in press)

Meneghetti, N., E, Tomba, P. Facco, F. Lince, D.L. Marchisio, A.A. Barresi, F. Bezzo and M. Barolo (2013). Supporting the transfer of products between different equipment through latent variable model inversion.

Accepted for presentation in: *ESCAPE 23, 23rd European Symposium on Computer Aided Process Engineering*, June 9-12, 2013, Lappeenranta (Finland).

Tomba, E., S. García-Muñoz, P. Facco, F. Bezzo and M. Barolo (2012). A general framework for latent variable model inversion to support product and process design. *Computer Aided Chemical Engineering 30*, (I.D.L. Bogle and M. Fairweather, Eds.), Elsevier, Amsterdam (The Netherlands), p.512-516.

CONTRIBUTIONS IN UNREFEREED CONFERENCE PROCEEDINGS

Tomba, E., M. Barolo and S. García-Muñoz (2012). Inversion of latent variable regression models to support product and process development. *AIChE 2012 Annual Meeting*, October 28-November 2, Pittsburgh, PA (U.S.A.).

Tomba, E., P. Facco, F. Bezzo, M. Barolo and S. García-Muñoz (2012). Impiego di dati storici per lo sviluppo di nuovi prodotti: applicazioni all'industria farmaceutica. *Convegno GRICU 2012. Ingegneria chimica: dalla nanoscala alla macroscala*, September 16-19, Montesilvano (PE, Italy).

Tomba, E., M. De Martin, P. Facco, F. Bezzo, S. Zomer, J. Robertson and M. Barolo (2012). Metodi statistici multivariati in supporto allo sviluppo di processi farmaceutici in continuo: un'applicazione industriale. *Convegno GRICU 2012. Ingegneria chimica: dalla nanoscala alla macroscala*, September 16-19, Montesilvano (PE, Italy).

Tomba, E., P. Facco, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Trasferimento tra impianti diversi di modelli per il monitoraggio di processo. *Convegno GRICU 2012. Ingegneria chimica: dalla nanoscala alla macroscala*, September 16-19, Montesilvano (PE, Italy).

Tomba, E., P. Facco, F. Bezzo, S. García-Muñoz and M. Barolo (2011). Transferring monitoring models between different scales through multivariate statistical techniques. *AIChE 2011 Annual Meeting*, October 17-21, Minneapolis, MN (U.S.A.).

Abstract

With the introduction of the Quality-by-Design (QbD) initiative, the American Food and Drug Administration and the other pharmaceutical regulatory Agencies aimed to change the traditional approaches to pharmaceutical development and manufacturing. Pharmaceutical companies have been encouraged to use systematic and science-based tools for the design and control of their processes, in order to demonstrate a full understanding of the driving forces acting on them. From an engineering perspective, this initiative can be seen as the need to apply modeling tools in pharmaceutical development and manufacturing activities.

The aim of this Dissertation is to show how statistical modeling, and in particular latent variable models (LVMs), can be used to assist the practical implementation of QbD paradigms to streamline and accelerate product and process design activities in pharmaceutical industries, and to provide a better understanding and control of pharmaceutical manufacturing processes.

Three main research areas are explored, wherein LVMs can be applied to support the practical implementation of the QbD paradigms: process understanding, product and process design, and process monitoring and control. General methodologies are proposed to guide the use of LVMs in different applications, and their effectiveness is demonstrated by applying them to industrial, laboratory and simulated case studies.

With respect to **process understanding**, a general methodology for the use of LVMs is proposed to aid the development of continuous manufacturing systems. The methodology is tested on an industrial process for the continuous manufacturing of tablets. It is shown how LVMs can model jointly data referred to different raw materials and different units in the production line, allowing to understand which are the most important driving forces in each unit and which are the most critical units in the line. Results demonstrate how raw materials and process parameters impact on the intermediate and final product quality, enabling to identify paths along which the process moves depending on its settings. This provides a tool to assist quality risk assessment activities and to develop the control strategy for the process.

In the area of **product and process design**, a general framework is proposed for the use of LVM inversion to support the development of new products and processes. The objective of model inversion is to estimate the best set of inputs (e.g., raw material properties, process parameters) that ensure a desired set of outputs (e.g., product quality attributes). Since the inversion of an LVM may have infinite solutions, generating the so-called null space, an optimization framework allowing to assign the most suitable objectives and constraints is used to select the optimal solution. The effectiveness of the framework is demonstrated in an industrial particle engineering problem to design the raw material properties that are needed to produce granules with desired characteristics from a high-shear wet granulation process. Results show how the framework can be used to design experiments for new products design. The analogy between the null space and the Agencies' definition of design space is also demonstrated and a strategy to estimate the uncertainties in the design and in the null space determination is provided.

The proposed framework for LVM inversion is also applied to assist the design of the formulation for a new product, namely the selection of the best excipient type and amount to mix with a given active pharmaceutical ingredient (API) to obtain a blend of desired properties. The optimization framework is extended to include constraints on the material selection, the API dose or the final tablet weight. A user-friendly interface is developed to aid formulators in providing the constraints and objectives of the problem. Experiments performed industrially on the formulation designed *in-silico* confirm that model predictions are in good agreement with the experimental values.

LVM inversion is shown to be useful also to address product transfer problems, namely the problem of transferring the manufacturing of a product from a source plant, wherein most of the experimentation has been carried out, to a target plant which may differ for size, lay-out or involved units. An experimental process for pharmaceutical nanoparticles production is used as a test bed. An LVM built on different plant data is inverted to estimate the most suitable process conditions in a target plant to produce nanoparticles of desired mean size. Experiments designed on the basis of the proposed LVM inversion procedure demonstrate that the desired nanoparticles sizes are obtained, within experimental uncertainty. Furthermore, the null space concept is validated experimentally.

Finally, with respect to the **process monitoring and control** area, the problem of transferring monitoring models between different plants is studied. The objective is to monitor a process in a target plant where the production is being started (e.g., a production plant) by exploiting the data available from a source plant (e.g., a pilot plant). A general framework is proposed to use LVMs to solve this problem. Several scenarios are identified on the basis of the available information, of the source of data and on the type of variables to include in the model. Data from the different plants are related through subsets of variables (common variables) measured in both plants, or through plant-independent variables obtained from conservation balances (e.g., dimensionless numbers). The framework is applied to define the process monitoring model for an industrial large-scale spray-drying process, using data available from a pilot-scale process. The effectiveness of the transfer is evaluated in terms of monitoring performances in the detection of a real fault occurring in the target process. The proposed methodologies are then extended to batch systems, considering a simulated penicillin fermentation process. In both cases, results demonstrate that the transfer of knowledge from the source plant enables better monitoring performances than considering only the data available from the target plant.

Riassunto

La recente introduzione del concetto di *Quality-by-Design* (QbD) da parte della *Food and Drug Administration* e delle altre agenzie di regolamentazione farmaceutica ha l'obiettivo di migliorare e modernizzare gli approcci tradizionalmente utilizzati dalle industrie farmaceutiche per lo sviluppo di nuovi prodotti e dei relativi processi produttivi. Scopo dell'iniziativa è di incoraggiare le industrie stesse all'utilizzo di procedure sistematiche e basate su presupposti scientifici sia nella fase di sviluppo di prodotto e processo, che nella fase di conduzione del processo produttivo stesso. A tal proposito, le Agenzie hanno definito paradigmi e linee guida per agevolare l'implementazione di queste procedure in ambito industriale, favorendo una migliore comprensione dei fenomeni alla base dei processi produttivi, in maniera da assicurare un controllo stringente sulla qualità dei prodotti finali, in termini di proprietà fisiche, ma soprattutto di efficacia e sicurezza per i pazienti.

Da un punto di vista ingegneristico, il *Quality-by-Design* può essere visto come il tentativo di introdurre principi di modellazione in ambiti di sviluppo e di produzione farmaceutica. Questo offre enormi opportunità all'industria farmaceutica, che può beneficiare di metodologie e strumenti ormai maturi, già sperimentati in altri settori industriali maggiormente inclini all'innovazione tecnologica. Allo stesso tempo, non va tralasciato il fatto che l'industria farmaceutica presenta caratteristiche uniche, come la complessità dei prodotti, le produzioni tipicamente discontinue, diversificate e in bassi volumi e, soprattutto, lo stretto controllo regolatorio, che richiedono strumenti dedicati per affrontare i problemi specifici che possono sorgere in tale ambiente. Per questi motivi, vi è l'esigenza di concepire metodologie che siano adeguate alle peculiarità dell'industria farmaceutica, ma al tempo stesso abbastanza generali da poter essere applicate in un'ampia gamma di situazioni.

L'obiettivo di questa Dissertazione è dimostrare come la modellazione statistica, e in particolar modo i modelli a variabili latenti (LVM, *latent variable models*), possano essere utilizzati per guidare l'implementazione pratica dei principi fondamentali del *Quality-by-Design* in fase di sviluppo di prodotto e di processo e in fase di produzione in ambito farmaceutico. In particolare, vengono proposte metodologie *generali* per l'impiego di modelli a variabili latenti nelle tre aree principali sulle quali l'iniziativa del *Quality-by-Design* si fonda: il miglioramento della comprensione sui processi, la progettazione di nuovi prodotti e processi produttivi, e il monitoraggio e controllo di processo. Per ciascuna di queste aree, l'efficacia della modellazione a variabili latenti viene dimostrata applicando i modelli in diversi casi studio di tipo industriale, di laboratorio, o simulati.

Per quanto riguarda il miglioramento della **comprensione sui processi**, nel Capitolo 3 è proposta una strategia generale per applicare LVM nello sviluppo di sistemi di produzione in continuo. L'analisi è applicata a supporto dello sviluppo di un processo industriale continuo di produzione di compresse su scala pilota. La procedura si basa su tre fasi fondamentali: *i*) una fase di gestione dei dati; *ii*) una fase di analisi esplorativa; *iii*) una fase di analisi globale. Viene mostrato come i parametri dei modelli costruiti a partire dai dati del processo possano essere interpretati sulla base di principi fisici, permettendo di identificare le principali forze motrici che agiscono sul sistema e di ordinarle a seconda della loro importanza. Questo può essere utile per supportare una valutazione dei rischi necessaria a definire una strategia di controllo per il processo e per guidare la sperimentazione fin dalle prime fasi dello sviluppo. In particolare, nel caso studio considerato, la metodologia proposta individua nel processo utilizzato per macinare le particelle di principio attivo e nella sezione nella quale il principio attivo è formulato le principali fonti di variabilità entranti nel sistema con effetto sulle proprietà fisiche del prodotto finale. Dall'analisi globale, è mostrato come l'utilizzo di modelli a variabili latenti a blocchi multipli permetta di individuare le unità del processo più critiche e, all'interno di ciascuna di esse, le variabili più critiche per la qualità del prodotto. Inoltre questi modelli si dimostrano particolarmente utili nell'identificare le traiettorie lungo le quali il processo si muove, a seconda delle proprietà delle materie prime e dei parametri di processo utilizzati, fornendo così uno strumento per garantire che l'operazione segua la traiettoria designata.

Nell'ambito della **progettazione di nuovi prodotti e processi**, l'efficacia dei modelli a variabili latenti è dimostrata nel Capitolo 4, dove è proposta una procedura generale basata sull'inversione di LVM per supportare lo sviluppo di nuovi prodotti e la determinazione delle condizioni operative dei rispettivi processi di produzione. L'obiettivo della procedura proposta è quello di fornire uno strumento atto a dare un'adeguata formalizzazione matematica, in termini di inversione di LVM, al problema di progettazione, secondo gli obiettivi e i vincoli che il problema stesso può presentare.

Dal momento che l'inversione di LVM può avere soluzioni multiple, vengono individuati quattro possibili problemi di ottimizzazione, tramite i quali effettuare l'inversione. L'obiettivo dell'inversione del modello è di stimare le condizioni ottimali in ingresso al sistema (in termini, per esempio, di caratteristiche delle materie prime o di parametri di processo) che assicurino di raggiungere la qualità desiderata per il prodotto in uscita. La procedura è applicata con successo in un caso studio industriale, per la determinazione delle proprietà delle materie prime in ingresso a un processo di granulazione a umido, con l'obiettivo di ottenere in uscita granuli con determinate caratteristiche di qualità.

È inoltre esaminato il concetto di spazio nullo, lo spazio cioè cui appartengono tutte le soluzioni di un problema di inversione di LVM, che corrispondono ad uno stesso insieme di variabili desiderate (proprietà del prodotto) in uscita. In particolare, si dimostra come la definizione di spazio nullo presenti diverse caratteristiche comuni alla definizione di spazio di progetto (*design space*) di un processo, stabilita dalle linee guida delle Agenzie di regolamentazione, e come lo spazio nullo possa essere utilizzato al fine di una identificazione preliminare dello spazio di progetto. Al fine di avere una misura sull'affidabilità delle soluzioni del problema di inversione, viene proposta una strategia per stimarne le incertezze.

Sono inoltre presentate alcune soluzioni per affrontare questioni specifiche relative all'inversione di LVM. In particolare, si propone una nuova statistica (P^2) da utilizzare per la selezione del numero di variabili latenti da includere in un modello utilizzato per l'inversione, in modo tale da descrivere adeguatamente l'insieme dei regressori, oltre a quello delle variabili in uscita. In aggiunta, dato che a causa delle possibili incertezze del modello non è assicurato che la sua inversione fornisca una soluzione che consenta di ottenere le proprietà desiderate per il prodotto, è proposta una strategia per sfruttare la struttura di covarianza dei dati storici per selezionare nuovi profili di qualità per il prodotto, in modo da facilitare l'inversione del modello. Gli approcci proposti sfruttano i parametri del modello e i vincoli imposti per la qualità del prodotto per stimare nuovi insiemi di proprietà, per i quali l'errore di predizione del modello è minimo. Questo agevola l'inversione del modello nel fornire le proprietà del prodotto desiderate, dal momento che queste possono essere assegnate come vincoli rigidi al problema di ottimizzazione.

Nel Capitolo 5 la procedura presentata al Capitolo 4 per l'inversione di LVM è applicata per progettare la formulazione di nuovi prodotti farmaceutici, in cui l'obiettivo è di stimare i migliori eccipienti da miscelare con un dato principio attivo e la loro quantità in modo da ottenere una miscela di proprietà adeguate per la fase di compressione. La procedura proposta al Capitolo 4 è ampliata al fine di includere i vincoli per la selezione dei materiali e di considerare gli specifici obiettivi che un problema di formulazione può presentare (per esempio, massimizzare la dose di principio attivo, o minimizzare il peso della compressa finale). L'inversione del modello è risolta come problema di programmazione non lineare misto-intera, per il quale è sviluppata un'interfaccia utente che consenta ai formulatori di specificare gli obiettivi e i vincoli che il problema di formulazione da risolvere può presentare. La metodologia proposta è testata in un caso studio industriale per progettare nuove formulazioni per un dato principio attivo. Le formulazioni progettate *in-silico* sono preparate e verificate sperimentalmente, fornendo risultati in linea con le predizioni del modello.

Nel Capitolo 6 è presentata una diversa applicazione della procedura generale per l'inversione di LVM presentata al Capitolo 4. Il caso studio riguarda un problema di trasferimento di prodotto, in cui l'obiettivo è di ottenere nanoparticelle di diametro medio predefinito, tramite un processo di precipitazione con anti-solvente in un dispositivo obiettivo. La metodologia sfrutta i dati storici disponibili da esperimenti effettuati su un dispositivo di riferimento di diversa dimensione da quello obiettivo, e sullo stesso dispositivo obiettivo ma con una diversa configurazione sperimentale. Un modello di tipo joint-Y PLS (JY-PLS) è inizialmente utilizzato per correlare dati di diversa origine (per dispositivo e configurazione sperimentale). Quindi, la procedura presentata al Capitolo 4 viene impiegata per invertire il modello JY-PLS al fine di determinare le condizioni operative nel dispositivo obiettivo, che assicurino l'ottenimento di nanoparticelle di diametro medio desiderato. La convalida sperimentale conferma i risultati ottenuti dall'inversione del modello. Inoltre gli esperimenti consentono di convalidare sperimentalmente il concetto di spazio nullo, dimostrando come diverse condizioni di processo stimate lungo lo spazio nullo consentano effettivamente di ottenere nanoparticelle con le medesime dimensioni medie.

La sezione finale di questa Dissertazione propone l'applicazione di LVM a supporto del **monitoraggio e controllo di processo** in operazioni farmaceutiche. In particolare, nel Capitolo 7, è affrontato il problema del trasferimento di modelli per il monitoraggio di processo tra impianti diversi. In questo caso il problema è di assicurare che l'operazione in un impianto obiettivo sia sotto controllo statistico fin dai primi istanti di funzionamento dell'impianto, sfruttando la conoscenza disponibile (in termini di dati) da altri impianti. È proposta una procedura generale basata su LVM per far fronte a questo tipo di problemi. La procedura identifica cinque diversi scenari, a seconda del tipo di informazioni disponibili (solo dati di processo o sia dati di processo sia conoscenza di base sul processo), della provenienza dei dati disponibili (solo dall'impianto di riferimento o sia dall'impianto di riferimento sia dall'impianto obiettivo) e dal tipo di variabili di processo considerate per la costruzione del modello (solo variabili comuni tra gli impianti o sia variabili comuni sia altre variabili). Per modellare in maniera congiunta i dati disponibili da impianti diversi, sono utilizzate analisi delle componenti principali (PCA) o modelli di tipo JY-PLS, a seconda che, per la costruzione del modello di monitoraggio, si considerino solo variabili comuni tra gli impianti (nel caso PCA), o sia variabili comuni sia altre variabili (nel caso JY-PLS).
Le metodologie proposte sono verificate nel trasferimento di modello per il monitoraggio di un processo industriale di atomizzazione, dove il riferimento è un impianto su scala pilota, mentre l'impianto obiettivo è un'unità produttiva su scala industriale. Le prestazioni in fase di monitoraggio del processo su scala industriale sono soddisfacenti per tutti gli scenari proposti. In particolare, è dimostrato come il trasferimento di informazioni dall'impianto di riferimento migliori le prestazioni del modello per il monitoraggio dell'impianto obiettivo.
Le procedure proposte sono inoltre applicate in uno studio preliminare per il trasferimento di sistemi di monitoraggio in processi discontinui, considerando come caso studio un processo simulato di fermentazione per la produzione di penicillina, in cui sono simulati due impianti differenti per scala e configurazione. Le prestazioni del sistema di monitoraggio indicano che, anche in questo caso, considerare nella costruzione del modello i dati disponibili dalle

operazioni nell'impianto di riferimento rende il sistema più efficiente nella rilevazione delle anomalie simulate nell'impianto obiettivo, rispetto a considerare nel modello di monitoraggio i soli (pochi) dati disponibili dall'impianto obiettivo stesso.

Table of contents

List of acronyms

API	=	active pharmaceutical ingredient
AR	=	alarm rate
BFI	=	brittle fracture index
BIP	=	block importance in the projection
CCV	=	common cause variability
CFD	=	computational fluid dynamics
CIJM	=	confined impinging jets mixer
CPP	=	critical process parameter
CQA	=	critical-to-quality attribute
DAE	=	differential algebraic equation
DEM	=	discrete element method
DoE	=	design of experiments
ESS	=	error sum of squares
FDA	=	Food and Drug Administration
FFC	=	flow function coefficient
HC	=	hard constraint
ICH	=	International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use
IR	=	infra-red
JY-PLS	=	joint-Y projection to latent structures
LDPE	=	low-density polyethylene
LOD	=	loss on drying
LPLS	=	L-shaped projection to latent structures
LV	=	latent variable
LVM	=	latent variable model
LVRM	=	latent variable regression model
MB	=	multi-block
MB-PLS	=	multi-block projection to latent structures
MBJY-PLS	=	multi-block joint-Y projection to latent structures

MCC	=	microcrystalline cellulose
MINLP	=	mixed-integer nonlinear programming
MJY-PLS		multiway joint-Y projection to latent structures
MPCA	=	multiway principal component analysis
MPLS	=	multiway projection to latent structures
MRT	=	microwave resonance technology
MSPC	=	multivariate statistical process control
MW	=	molecular weight
NIPALS	=	nonlinear iterative partial least squares
NIR	=	near infra-red
NMR	=	nuclear magnetic resonance
NOC	=	normal operating conditions
ODE	=	ordinary differential equation
OPLS	=	orthogonal projection to latent structures
PAT	=	process analytical technology
PBM	=	population balance model
PC	=	principal component
PCA	=	principal component analysis
PCL	=	poly-ε-caprolactone
PCR	=	principal component regression
PDE	=	partial differential equation
PID	=	proportional-integral-derivative
PIMS	=	particle imaging measurement system
PLS	=	projection to latent structures
PRESS	=	prediction error sum of squares
PSD	=	particle size distribution
PSE	=	process systems engineering
QbD	=	quality-by-design
QTPP	=	quality target product profile
RMSECV	=	root mean squared error of cross-validation
SC	=	soft constraint
SIMPLS	=	statistically inspired modification of projection to latent structures

SPE	=	squared prediction error
SPE$_i$	=	squared prediction error for sample *i*
$\mathrm{SPE}_{(1-\alpha)\mathrm{lim}}$	=	$(1-\alpha)\%$ squared prediction error confidence limit
SVD	=	singular value decomposition
TD	=	time to detection
TSS	=	total sum of squares
VIP	=	variable importance in the projection
WSPLS	=	weighted-scores projection to latent structures
wtd	=	*weighted temperature difference*

Chapter 1

Motivation and state of the art

This Chapter provides an overview of the background and the motivations of this Dissertation. First, the *Quality-by-Design* (QbD) initiative for the pharmaceutical industry is introduced from a regulatory point of view. Then, the significance of this concept and the opportunities it gives for the process systems engineering community are pinpointed and discussed. Finally, the role and the importance of latent variable models in the implementation of QbD paradigms are highlighted, providing the objectives of the Dissertation and a roadmap to its reading.

SECTION A – OVERVIEW OF REGULATORY ISSUES

1.1 The Quality-by-Design (QbD) initiative

Despite being perceived as on the cutting edge for its social impact and for the innovation and relevance of the manufactured products, the pharmaceutical industry has been traditionally based on experienced and strict procedures not only for product and process development but even for product manufacturing. This situation has been partly due to the rigid regulatory environment, which strongly contributed to prevent improvements and innovation in the manufacturing technologies. According to the regulatory agencies, like the American Food and Drug Administration (FDA), it seemed more important to manufacture drugs precisely to the required specifications in order to protect the patients' safety, rather than latching on the latest in manufacturing trends. This contributed to spread the belief that using tried-and-true systems and operating with traditional accepted manufacturing procedures, which ensured to produce drugs with very targeted specifications, would have served as basic requirements from the regulatory point of view. As a consequence, pharmaceutical companies felt stimulated to invest their money in finding and marketing new drugs, rather than in revamping development procedures or manufacturing facilities. The result is that the pharmaceutical industry, even inventing futuristic new drugs, still relies on manufacturing techniques that lag far behind those of potato-chip and laundry-soap makers (Aboud and Hensley, 2003).

The little emphasis set by the pharmaceutical industry toward manufacturing technologies and process efficiency increased the economical efforts of companies to ensure high quality standards for the products. Pharmaceutical manufacturing has always been able to achieve reasonable product quality, but at the price of high percentages of rejected products due to process inefficiency.

In 2003, a survey of the Wall Street Journal on the state of the pharmaceutical industry quantified the percentage of product scraps due to manufacturing shortcomings in between 5% and 10% of the produced medicine. This was contrasted with the 0.0001% of the semiconductor industry. Another measure of the impact of the manufacturing system deficiencies is the number of drug recalls for quality reasons: in 2002 the FDA counted 354 prescription-drug recalls, up from 248 in 2001 and 176 in 1998. However, the manufacturing expenses to ensure those percentages of discards accounted for 36% of the total industry's costs, more than double than the share of research and development, and almost as much as the 41% devoted to marketing and administrative costs (Aboud and Hensley, 2003).

If pharmaceutical manufacturing has been dramatically affected by the lack of modernization and efficiency, pharmaceutical development activities have been hardly less so. In general, pharmaceutical development includes all the activities dealing with the transformation of one or more active pharmaceutical ingredients (APIs) into the final drug ready for the commercialization. These activities involve three main steps: product design, process design and technology transfer.

Product design includes all the activities that ensure that the designed product meets the needs for which it is intended, in terms of safety, efficacy and marketing. This involves the choice of the product form (solid, granulated, inhaled, etc.), the product formulation (namely, the choice in terms of type and amount of the materials to be mixed with the API), and the selection of the packaging materials. Process design includes the identification of the unit operations that should be used to manufacture the desired product and the definition of the process operating conditions. Both the product and the process activities must be repeatedly refined to ensure that they are sufficiently reliable for the technology transfer phase. This phase includes the scale-up from laboratory (via pilot plant) to the manufacturing scale for mass production; this is not limited to the process operating conditions, but involves all the technologies tested and implemented in the small scale plants (e.g., sensors, analyzers, etc.). Manufacturing is the culmination of this complex set of activities, which FDA has called *industrialization process* (FDA, 2004a). This process has been recognized to be the weak link in the path from scientific discovery of the API to the commercialization of the final product (IBM, 2005). Nevertheless, the FDA has acknowledged to have unintentionally contributed to this situation, by transforming its relation with companies in a strict oversight rather than an interactive collaboration, preventing companies to invest in the modernization of the industrialization process. The result is that product and process development in the pharmaceutical industry has often been highly inefficient and in general burdensome in terms of time and resources employed. Furthermore, the lack of updating in the procedures has progressively led the pharmaceutical development to become an art rather than a scientific and technical activity. The main consequence of these issues is that a long time is needed for a new product to be approved and launched into the market, which has an unavoidable impact on the product market price.

Attention towards the inefficiencies in pharmaceutical development and manufacturing dramatically increased in the last decade, due to the particular situation the pharmaceutical industry has been coping with. While in the past, the discovery activities were particularly fruitful for the companies, ensuring an adequate number of new products in the pipeline for approval and allowing them to cover the high expenses due to the long developmental times and inefficiencies after their commercialization, in recent years the number of discoveries has experienced a progressive decrease. The FDA approved only 15 new molecular entities in 2010, while 19 were approved in 2009 and 21 in 2008. In 1996, the new molecular entities approved were 53 (Mullard, 2011). New discoveries have been limited to modifications of existing products or mainly related to the drug delivery systems, rather than to new products. In the meantime, many patents of blockbuster products have (or have already) approached their expiration, while companies' pipelines miss worthy substitutes. This contributed to increase the pressure on the bigger pharmaceutical research companies, which have to face the increasing competition with generic drug makers.

Considering this technical and economic background, the regulatory agencies have tried to respond to the industry needs by focusing the attention towards the modernization of the pharmaceutical development and manufacturing apparatus, with the ultimate objectives of providing tools that can be exploited to improve the efficiency in both the development and the manufacturing stages, and can lead to a return and an advantage in terms of time, resources and competition.

For this reason, in recent years the FDA launched several initiatives in order to reform its relations with pharmaceutical companies (FDA, 2004b, 2004c). The principal aim of these initiatives was to encourage a broad change in the way industry develops and makes its products. The FDA acknowledged that the state of the rigid regulatory framework was the main cause for companies to not invest in innovation and high technology in manufacturing. In fact, as mentioned earlier, companies preferred to keep their processes frozen, as every change in the process technologies would have required new submissions to the agencies for change approval, with subsequent production delays.

Taking inspiration from the experiences of different industries (e.g. automotive, semiconductors, etc.), the FDA introduced the concept of *Quality by Design* (QbD), namely a new approach to pharmaceutical development and manufacturing, which had the purpose of favoring an efficient and flexible environment to produce reliably high quality products, without extensive regulatory oversight (Winkle, 2007). The QbD initiative encourages companies to the adoption of systematic science-based tools, rather than fixed traditional procedures. The ultimate objective of this approach is to promote product and process understanding in pharmaceutical development, in order to increase manufacturing flexibility and process robustness (i.e., the ability of the process to tolerate variability of materials and changes in the process and equipment without negative impact on product quality). According to the QbD philosophy, the quality of a product cannot be assessed at the end of the product development activity or after manufacturing, but must be "built into" the product and ensured *since its design*, through a thorough mechanistic understanding of the relations between the quality of the product and the parameters that have an impact on it.

In general, a QbD approach to pharmaceutical development must be scientific, risk-based, holistic and proactive (Winkle, 2007). In other words, to achieve a full understanding of the several sources of variability affecting a pharmaceutical product (and impacting its quality) through the raw materials and the process, companies are invited to apply mathematical and physical tools, that can describe quantitatively the relations between variables. This would help to identify the most critical variables for the quality of the product and to rank them according to the *risk* that their variations affect the product quality.

1.2 QbD paradigms for pharmaceutical development

The QbD guidelines identify and define different elements of the new QbD-based approach to pharmaceutical development. These elements, which should be integral parts of a QbD application, are proposed in order to inspire a practical implementation of QbD. Table 1.1 lists the main regulatory agencies' documents that introduce and define the QbD paradigms, with the main contribution they provide.

1.2.1 Critical-to-Quality Attributes and risk assessment

Since the ultimate scope of the initiative was to improve the control of the pharmaceutical companies on the quality of its manufactured products, the first important step of QbD is the definition of what is meant for *product quality*.

According to the guidelines of the *International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use* (ICH), which brings together the regulatory authorities of Europe, Japan and United States with experts from the pharmaceutical industry, quality is defined as the suitability of either a drug substance or drug product for its intended use (ICH, 1999). Now, the quality of a pharmaceutical product has to take into account the safety and the efficacy of the drug, together with the product characteristics related (for example) to the route of administration, the dosage form, bioavailability, strength and stability. The summary of these characteristics, which have to be achieved to ensure the desired quality, forms the so-called *Quality Target Product Profile* (QTPP; ICH, 2009).

Table 1.1. *Main regulatory agencies' documents introducing and defining the QbD paradigms.*

Document	Contribution
ICH (1999)#	Defines the concept of quality and assists in the establishment of global specifications for new drug substances or drug products.
FDA (2004a)	Defines the *industrialization process* as the set of activities related to product design, process design and technology transfer. Acknowledges that problems in these steps routinely derail or delay development programs.
FDA (2004b)	Outlines the QbD concept and summarizes initiatives to encourage science-based policies and innovation in pharmaceutical development and manufacturing. Proposes risk assessment as a tool to evaluate the impact if variations in process inputs for product quality.
FDA (2004c)	Introduces the Process Analytical Technology (PAT) framework. Defines process understanding, critical-to-quality attributes and critical process parameters and identifies PAT tools. Introduces the real time release concept.
ICH (2005) #	Defines the concept of risk for pharmaceutical quality and provides principles and examples of tools for risk assessment and management.
ICH (2008) #	Describes a model for an effective quality management system throughout the lifecycle of the product. Outlines the control strategy and continual improvement concepts.
ICH (2009) #	Provides an overview of QbD in pharmaceutical development. Defines most of the QbD paradigms (quality target product profile, critical-to-quality attributes, risk assessment, design space, control strategy), providing guidelines of their implementation and submission in technical documents.
ICH (2010) #	Questions & Answers session in order to facilitate the implementation of the Q8/Q9/Q10 guidelines. Provides several clarifications and the regulatory perspective mainly focused on QbD topics as design space, real time release testing and control stratgy
ICH (2011) #	Provides a guide for ICH Q8/Q9/Q10 guideline implementation, with emphasis on criticality identification, control strategy, design space and process validation. Introduces the use of modeling as a tool to implement QbD at every stage of development. Categorizes models and provides an ouline for their implementation, validation and verification.

# ICH documents are recommended for adoption to the regulatory bodies of European Union, Japan and United States. The same documents can be found in the adopted form as regulatory agencies' guidances.

The product characteristics that are identified as having an impact on the QTPP are defined *critical-to-quality attributes* (CQAs; FDA, 2004c). These include the physical, chemical, biological or microbiological properties or characteristics that have been demonstrated to ensure the desired product quality, if within an appropriate limit, range or distribution. According to this definition, the CQAs are associated not only to the product, but also to the raw/input materials used in product formulation (e.g. excipients), intermediates (in-process materials) and to the APIs. Product and process development in a QbD framework should be guided by the CQAs of the drug product derived from the QTPP and/or prior knowledge

(ICH, 2009). Accordingly, the FDA suggests to identify the CQAs of the input materials (APIs and excipients) or intermediates through a systematic procedure based on experiments, to assess the extent to which their variation impact on the quality of the product.

Other than CQAs, the ICH guidelines indicate the manufacturing process as another main source of variability for product quality. In particular, among the process parameters, namely those settings of the process that can be manipulated at the beginning or during the operation, the ones whose variability is demonstrated to have an impact on one or more CQAs are defined as *critical process parameters* (CPPs). CPPs need to be monitored or controlled to ensure the desired product quality (ICH, 2009).

To identify CQAs and CPPs, the FDA suggests to implement procedures based on the evaluation of the risk that considering or not an attribute as a CQA or a process parameter as a CPP has on the final product quality. The risk is linked to the impact that a variation in a material or intermediate attribute or in a process parameter has on product CQAs. This procedure of risk evaluation is called quality *risk assessment* (FDA, 2004b), which is defined as the "qualitative or quantitative process of linking the likelihood of occurrence and severity of harm" (ICH, 2005). Risk assessment "is typically performed early in the pharmaceutical development process, and is repeated as more information becomes available and greater knowledge is obtained" during the development and the manufacturing of the product (ICH, 2009). In the ICH Q9 guide (2005), a series of quality risk management tools are indicated to support the risk assessment phase in selecting and ranking the quality attributes (including material attributes) and/or process parameters that should be further evaluated or controlled within appropriate ranges to ensure the desired product quality (e.g. failure mode effect analysis, fault tree analysis, hazard operability analysis) and to manage the identified risks.

The results of the risk assessment procedure should be a list of potential parameters selected on the basis of prior knowledge, scientific first principles and experimentation. This list can be refined further through experimentation to determine the significance of individual variable and potential interactions. Once the significant parameters are identified, they can be further studied to achieve a higher level (possibly mechanistic) of process understanding. It is important to note that CQAs and CPPs can evolve throughout the product lifecycle, from the initial development through marketing and until the product discontinuation (ICH, 2009).

1.2.2 Design space

The risk assessment and process development experiments can lead to an understanding of the linkage and effect of process parameters and material attributes on product CQAs, and also help identifying the variables and their ranges within which consistent quality can be achieved. These process parameters and material attributes can thus be selected for inclusion in the *design space* of the process (ICH, 2009). The ICH Q8 guideline defines the design space as "the multidimensional combination and interaction of input variables (e.g. material attributes) and process parameters that have been demonstrated to provide assurance of quality". The design space concept is one of the fundamental paradigms on which the QbD framework is based, and its description is expected to be one of the results of the pharmaceutical development investigation according to a QbD approach.

The design space concept introduces a revolution for pharmaceutical development and in the relation between pharmaceutical companies and regulatory agencies. When a design space is established for a manufacturing process, working within the design space is not considered as a change. Only movements out of the design space are considered to be a change and would normally initiate a regulatory post approval change process. This philosophy drastically changes the classical way agencies used to supervise pharmaceutical development, when every change in the process had to be communicated for evaluation and approval.

The design space is therefore considered as the final achievement of process understanding in the development of new products and processes. A process is generally considered well understood when (FDA, 2004c):

- all critical sources of variability are identified and explained;
- variability is managed by the process;
- product quality attributed can be accurately and reliably predicted over the design space established for materials used, process parameters, manufacturing environmental and other conditions. The ability to predict reflects a higher degree of process understanding.

However, a design space can be updated over the lifecycle of the product as additional knowledge is gained.

The ICH guidelines provide general indications on how to describe and establish a design space in different situations, leaving the initiative to the companies on the most appropriate tools to employ. As emphasized by the definition, the design space has a *multivariate* nature, suited to explore not only the effect of the *single* material attributes or process parameters, but also their *interactions* and *combined* effects. For this reason a design space *cannot* be expressed as a combination of proven acceptable ranges, namely ranges of the process parameters, obtained for each single parameter while keeping the other constant, for which the operation resulted in producing a product meeting the relevant quality criteria. This is due to the fact that experiments to define proven acceptable ranges would be univariate, thus lacking an understanding of the interactions between process parameters and material attributes. Hence, the definition of the design space requires performing *multivariate* experiments which can highlight possible parameter interactions. Nonetheless, a design space can still be described in terms of ranges of material attributes and process parameters, but also in terms of more complex mathematical relationships, time dependent functions, or as a combination of variables such as components of a multivariate model (ICH, 2009).

ICH specifies that a design space can be developed also for formulations only. In this case it should be described in terms of compositions rather than components, and consists of ranges of excipient amount and their physicochemical properties, based on an enhanced knowledge over a wider range of material attributes. Formulation adjustments within the design space depending on material attributes does not need a submission in a regulatory post approval change (ICH, 2010), unlike changes in the formulation components.

For a manufacturing process, the agencies leave the applicant with the choice to establish independent design spaces for one or more unit operations, or to establish a single design space that spans multiple unit operations in a line. While a separate design space for each unit operations is often simpler to develop, a design space that spans the entire process can provide more operational flexibility. In general, when describing a design space, the applicant should consider the type of operational flexibility desired. A design space can be developed at any scale, but the applicant should justify the relevance of a design space developed at small or pilot scale to the proposed production scale manufacturing process, and discuss the potential risks in the scale up operation. In general, if a process design space has to be applicable to multiple operational scales, the design space should be described in terms of relevant scale-independent parameters (ICH, 2009).

1.2.3 Control strategy and real time release testing

To ensure that a manufacturing process is maintained within the boundaries described by the design space, the definition of an appropriate *control strategy* is required. It should be noted that the term *control* does not usually refer to the traditional engineering understanding of process control. In fact, according to the regulatory agencies, the control strategy is defined as "a planned set of controls, derived from current product and process understanding, that ensures process performance and product quality. The controls can include parameters and attributes related to drug substance and drug product materials and components, facility and equipment operating conditions, in-process controls, finished product specifications and the associated methods and frequency of monitoring and control" (ICH, 2008). These controls should be based on product, formulation and process understanding and should include, at a minimum, control of the sources of variability that can impact the product quality. Understanding these sources of variability and their impact on downstream processes or processing, in-process materials and drug product quality can provide the opportunity to shift controls upstream and minimize the need for end product testing (ICH, 2009). The objective is therefore to design a system able to compensate for the variability entering the system (e.g. through the raw materials) in an adaptable manner to deliver consistent product quality. This would enable an alternative manufacturing system paradigm, where the variability of the input materials could be less tightly constrained, as the process is designed to be responsive to that variability.

Enhanced product understanding of product performance can justify the use of alternative approaches to determine that a product (intermediate or final) is meeting its quality attributes.

The use of such alternatives could support *real time release testing*, namely “the ability to evaluate and ensure the quality of in-process and/or final product *based on process data*” (measured material attributes, process parameters). For example, the uniformity of unit dose performed in-process (e.g., using weight variation coupled with near infrared – NIR – assay) can provide real time release testing and an increased level of quality assurance compared to the traditional end-product testing (ICH, 2009). From this point of view, the real time release concept is introduced with the aim of reducing or eliminating slow end product testing, by ensuring a real-time assurance of quality.

In summary, a control strategy can include, but it is not limited to, the following (ICH, 2009):

- control of input material attributes (e.g. APIs, excipients, primary packaging materials), based on understating of their impact on processability or product quality;
- product specification(s);
- controls for unit operations that have an impact on downstream processing or product quality;
- in-process or real-time release testing in lieu of end-product testing;
- a monitoring program for verifying prediction models performances (e.g. through full product testing at regular intervals).

The control strategy should facilitate feedback/feedforward controls and appropriate corrective/preventive actions for the manufacturing process. It must be underlined that, as mentioned above, in the ICH documents the control strategy is intended both to control product specifications and for the control of unit operations. These two purposes have however a completely different meaning from a practical point of view: if the control strategy aims at narrowing the region determined by product specifications, in order to ensure a robust product quality, the objective of the unit operation control is that of providing tools to respond to the variability entering the process, thus widening the acceptance region for the variables in input to the manufacturing process (e.g., raw material attributes) and subsequently the process design space. The qualitative difference between a design space established with and without defining an appropriate control strategy and its linkage to the product specification space is exemplified in Figure 1.1. In this case, for the sake of simplicity the design space is represented in both the input variable space and the product specification space, even if the latter would be the mathematical image of the former. Moreover, in both cases they are considered as subsets of a wider *knowledge space*, represented by the historical or experimental gained knowledge. As can be seen form Figure 1.1, the design space with the control strategy implemented is much wider than a design space without controls. The corresponding region around the product specifications (called *control space*) is accordingly much narrower than the region corresponding to the design space without controls. This is something commonly known in control engineering, namely the fact that to achieve tighter control of the final quality variables, one needs to accept more variation in the manipulated variables. In this way the variability is transferred from the product specifications to the manipulated variables (Bruwer and MacGregor, 2008), i.e. from where it “hurts” to where it does not.

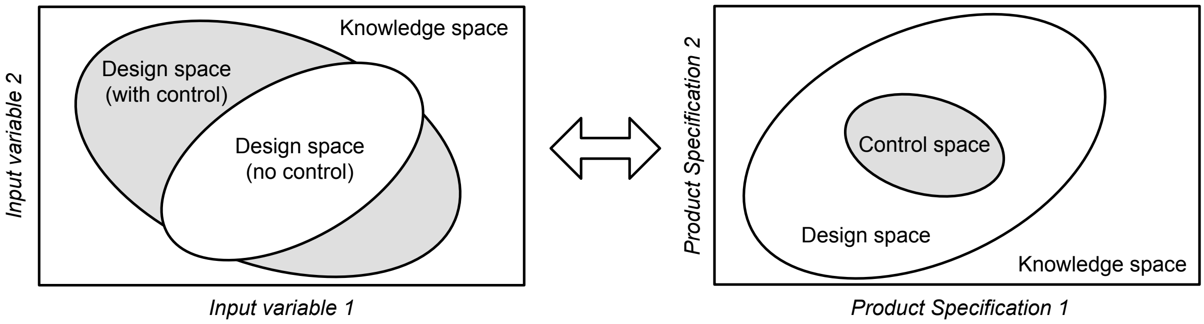


Figure 1.1. *Schematic of the relation between the design space with and without control strategy and the product specifications (adapted from Bruwer and MacGregor, 2008).*

The design space and the control strategy should be verified and improved over the lifecycle of the product, especially when new knowledge is gained. For this reason, *continuous process verification* tools should be applied by the companies to monitor the process and make adjustments to the process and/or to the control strategy. Continuous process verification is an approach to process validation that includes the continuous monitoring and evaluation of manufacturing process performance (ICH, 2009). It can enhance the evaluation of the manufacturing process if it provides substantially more information on process variability and control. Continuous process verification can utilize in-line, on-line or at-line monitoring or controls to evaluate process performance, which are based on product and process knowledge and understanding. Monitoring can also be combined with feedback loops in order to adjust the process to maintain output quality. The advantage of using continuous process verification is that it provides the foundation for a robust process performance and product quality monitoring system, increasing in the meanwhile product and process knowledge and facilitation of continual improvement opportunities for process and product quality. This would provide a higher assurance of an ongoing state of control (through the adoption of appropriate statistical tools), enabling the earlier detection of manufacturing-related problems and trends, and contributing to the verification of the design space (ICH, 2010).

1.2.4 Process Analytical Technology (PAT)

Building quality into products rather than testing it at the end of the manufacturing process implies that a comprehensive understanding of the characteristics of the drug (chemical, physical, pharmacological, pharmacokinetic, etc.), of the design and selection of the product components, and of the design of the manufacturing process and quality assurance is achieved. To reach this level of comprehension and develop well understood processes that are able to ensure consistently the predefined product quality, appropriate tools need to be employed, in order to measure and analyze effectively the relevant data. To this end, the FDA introduced in 2004 the *process analytical technology* (PAT) framework. According to the agency definition, PAT is "a system for designing, analyzing and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring product quality". It is important to note that the term *analytical* in PAT is viewed broadly to include chemical, physical, microbiological, mathematical and risk analysis conducted in an integrated manner (FDA, 2004c).

Through the PAT initiative, the FDA indicates the tools to be considered for an effective innovation in development, manufacturing and quality assurance. In particular, the objective of PAT is to provide support to clarify on a scientific basis typical issues that are likely to be encountered in development and manufacturing studies: for example, which the effects of product components on quality are, what sources of variability are more critical for the product, or how the process is able to manage variability.

In general, PAT includes all those tools that can provide an effective and efficient mean for acquiring valuable information to facilitate process understanding, continuous improvement through process and product monitoring and development of control and risk-mitigation strategies. In the PAT framework, these tools can be categorized according to the following (FDA, 2004c):

- multivariate tools for design, data acquisition and analysis;
- process analyzers;
- process control tools;
- continuous improvement and knowledge management tools.

The multivariate tools category includes all the multivariate mathematical approaches, such as statistical design of experiments, response surface methodologies, process simulation and pattern recognition tools, in conjunction with knowledge management systems, which allow to gain scientific understanding of the relevant multi-factorial relationships between formulation, process, and quality attributes. It includes also the means to evaluate the applicability of this knowledge to different scenarios. When used appropriately, these tools "enable the identification and evaluation of product and process variables that may be critical to product quality and performance". They may also "identify potential failure modes and mechanisms and quantify their effects on product quality" (FDA, 2004c).

Process analyzers include all the tools committed to the collection of data from the process. These measurements can be obtained at-line, i.e. by removing, isolating and analyzing the sample in proximity to the process stream; on-line, i.e. by diverting the sample from the manufacturing process and returning it to the process stream after the measurement; in-line, i.e. by keeping the sample inside the process stream, while the measurement can be made invasively or not. Process analyzers are identified as useful tools to generate data not only for process understanding, but especially for real-time control and product quality assurance during manufacturing. Process analyzers generate typically large volumes of data. For this reason, multivariate methodologies are indicated to extract critical process knowledge that can be related to product and process quality and used for process monitoring, control and end point determination. The design and installation of the analyzers on the process equipment is also identified as a critical step, as it must be ensured that the collected data are relevant and representative of process and product attributes. For this reason, the installation of process analyzers should be done after risk analysis, in order to ensure that it does not adversely affect product or process quality (FDA, 2004c).

The process control tools include all the "process monitoring and control strategies intended to monitor the state of a process and actively manipulate it to maintain a desired state. Strategies should accommodate the attributes of input materials, the ability and reliability of process analyzers to measure CQAs, and the achievement to process end points to ensure consistent quality of the output materials and the final product". Multivariate Statistical Process Control (MSPC) is advocated as a feasible and valuable tool to realize the full benefit of these (often real time) measurements. In a PAT framework, the process should be continually monitored, evaluated and adjusted using in-process measurements, tests and controls in order to guarantee continuous quality assurance. This represents a way to demonstrate process validation.

Finally, the Agency encourages the adoption of PAT as continuous improvement tools, which enable a continuous learning through the data collected and analyzed over the lifecycle of the product. Approaches that support the acquisition of knowledge from these data would be valuable for manufacturing and facilitate the communication with the Agency on a scientific basis.

On the basis of PAT tools and principles, the design and optimization of drug formulations and manufacturing processes within the PAT framework can include the following steps (FDA, 2004c):

- identify and measure CQAs and CPPs;
- design a process measurement system to allow real time (or near real time) monitoring of all CQAs, using direct or indirect analytical methods;
- design process control strategies that provide adjustments to ensure control of all critical attributes;
- develop mathematical relationships between product CQAs and material CQAs and process parameters.

The combination of assessed material attributes and process controls constitutes the PAT component of real time release. According to the FDA guidelines, process understanding, control strategies plus on-, in- or at- line measurement of CQAs that relate to product quality provides a scientific risk-based approach to justify how real time quality assurance is at least equivalent to, or better than, laboratory-based testing on collected samples. This can serve as the basis for real time release of the final product (FDA, 2004c).

Table 1.2. *Comparison between traditional and QbD-based approaches to pharmaceutical development and manufacturing (ICH, 2008).*

Aspect	Traditional approach	QbD-based approach
Pharmaceutical development	– Empirical – Typically univariate experiments	– Systematic, relating mechanistic **understanding** of material CQAs and CPPs to product CQAs – Multivariate experiments – Establishment of **design space** – **PAT** tools utilized
Manufacturing process	– Fixed – Validation based on initial full-scale batches – Focus on optimization and reproducibility	– Adjustable within **design space** – Lifecycle approach to validation – Focus on **control strategy** and robustness – Use of statistical process control
Process control	– In-process tests for go/no go decisions – Off-line analysis	– **PAT** tools utilized with appropriate feedforward and feedback control strategies – Process operations tracked and trended to support **continual improvement**
Product specifications	– Primary means of quality control – Based on batch data available	– Part of the overall quality **control strategy** – Based on desired product performance (safety and efficacy)
Control strategy	– Drug product quality controlled mainly by intermediate and end product testing	– Drug product quality ensured by risk-based **control strategy** – Quality controls shifted upstream, with the possibility of **real time release**
Lifecycle management	– Reactive (i.e., problem solving and corrective action)	– Proactive action – **Continual improvement** facilitated

1.2.5 QbD implementation in pharmaceutical development and manufacturing

QbD provides an enhanced approach to pharmaceutical development and manufacturing, based on scientific and engineering principles for assessing and mitigating risks related to poor product quality and process performances. In pharmaceutical development, the objective of QbD is the achievement of a scientific understanding of how input material factors and manufacturing process factors affect product quality. The level of achieved understanding is the basis for a robust design of the product formulation and of an effective and efficient manufacturing process. In manufacturing, the main objective of QbD is to provide systems able to assure in real time that the product meets the quality requirements. This implies that the process should have the capability to identify and respond to disturbances entering the system.

In Table 1.2, a comparison between the strategy outlined by the QbD paradigms and the traditional approaches is summarized for some of the key aspects of pharmaceutical development and manufacturing (ICH, 2009). As can be seen, unlike the traditional approach that is substantially empirical, the implementation of the QbD paradigms in pharmaceutical development has to be based on a mechanistic understanding of the driving forces acting on the process, both in the developmental phase and in the manufacturing phase. This mechanistic understanding can be achieved only through the execution of appropriate multivariate experiments in which the relevant process inputs are excited to register their impact on the output (response variables). Indeed, process understanding achieved through designed experiments and from manufacturing data has real business, other than scientific, value.

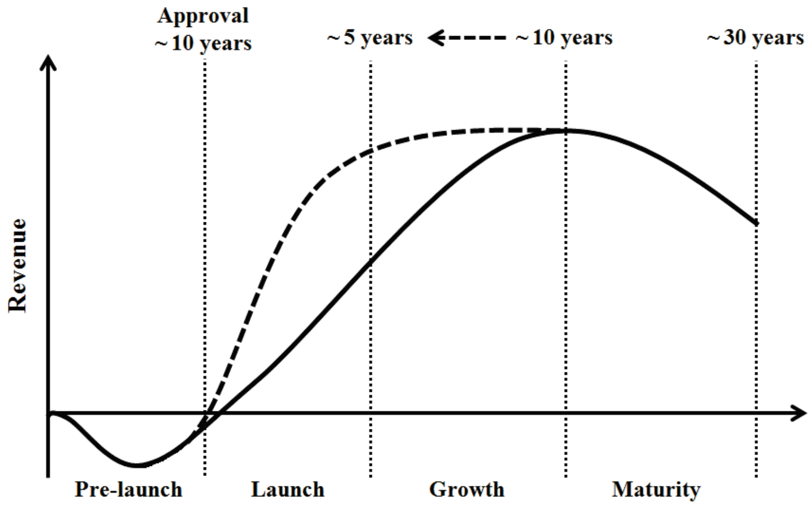


Figure 1.2. *Revenue trend for a drug product during its lifetime, if a traditional (solid line) or a QbD-based approach (dashed line) were used for pharmaceutical development and manufacturing (adapted from IBM, 2005).*

Figure 1.2 reports the trend of the total revenues a drug product brings, from the discovery to the patent expiration in a traditional pharmaceutical development and manufacturing framework (solid line) (IBM, 2005). As can be seen, after the pre-launch phase, in which investments in research and development are needed and which usually lasts around ten years, the product is launched and drug sales increase the revenues. Since very often companies launch their products before the manufacturing process is fully optimized, in the first year or two there still are not revenues. Afterwards, product sales start to increase, until reaching a peak usually ten years after the product launch. Sales then may decrease, when the product is mature, e.g., for competition reasons. IBM estimated that improving new product and process development to design robust manufacturing processes through a QbD-based approach prior to the launch of new products could help reducing the period from launch to peak sales by as much as five years, thus unlocking an enormous amount of added value (dashed line in Figure 1.2). As an example, a drug with peak annual sales of US$1 billion was estimated to generate an extra US$1.6 billion over its lifetime.

In order to exploit the scientific and economic benefits of QbD, appropriate methodologies need to be conceived and/or implemented to support the design of experiments, the analysis of the data, the extraction of information needed for process understanding, the definition of the control strategy. From an engineering point of view, these issues can be addressed by resorting to appropriate modeling tools. The pharmaceutical companies can therefore benefit strongly from modeling tools, which are now mature and widely used in other industries (e.g., chemical, petrochemical, polymer, consumer goods, energy) and need only to be adapted to the needs and constraints the pharmaceutical environment is subject to.

Stemming from the QbD paradigms, the practical implementation of QbD in the development of new pharmaceutical products can go through the following steps (Winkle, 2007; Yu, 2008):

1. Define the desired performances of the product and identify the CQAs.
2. Design the product formulation and the manufacturing process in order to meet the CQAs.
3. Understand the impact of materials attributes and process parameters on product CQAs.
4. Identify and control the sources of variability due to the raw materials and the manufacturing process.
5. Continually monitor and improve the manufacturing process in order to assure consistent product quality.

The first three steps of the above-mentioned roadmap can be seen as part of the pharmaceutical development activities, while the last two are mainly related to pharmaceutical manufacturing. Appropriate modeling tools can enter in each step of the presented procedure, thus supporting the practical implementation of QbD paradigms.

SECTION B – OVERVIEW OF RESEARCH ISSUES

1.3 QbD and modeling

Most of the QbD paradigms described in the FDA guidelines (design space, control strategy, PAT) can be understood as the application of Process Systems Engineering (PSE) to the development and manufacturing of pharmaceutical products (García-Muñoz and Oksanen, 2010). The QbD aim is to improve process efficiency and quality control on the manufactured products through a scientific understanding of the relationships between the variables impacting the quality. The description of these relationships can be mathematically formulated in a model, linking input variables (raw materials CQAs, CPPs) with product CQAs. The trends highlighted by QbD have therefore opened the route towards a new concept of product development, which has to be model-based, rather than experience based, and integrated with the process development. Modeling can be used in every stage of pharmaceutical development and manufacturing, and is mainly intended to enhance process understanding and predict the behavior of a system under different conditions (ICH, 2011). As a consequence, models can be used to support development activities, in order to accelerate the launch of new products in the market, but also to improve the productivity and to control the product quality in manufacturing environments. PSE plays then a central role for the pharmaceutical industry by providing the tools to address simultaneously process design, the design space definition, and process monitoring and control using PAT.

The use of modeling to support QbD implementation has been encouraged by the regulatory agencies, which distinguish between different categorizations of models (ICH, 2011). For the purpose of regulatory submissions, an important factor for categorization is the model contribution in assuring the quality of the product. Accordingly, models can be distinguished in low, medium and high impact. Low-impact models are typically the ones used to support product and/or process development (e.g., formulation optimization); medium-impact models can be useful in assuring the quality of the product but are not the sole indicators of product quality (e.g., most design space models, many in-process controls); high-impact models are those whose prediction is a significant indicator of the quality of the product (e.g., a chemometric model for product assay). For the purpose of implementation, models can also be categorized on the intended outcome of the model. Within each of these categories, models can be further classified (as above) in low, medium and high impact in assuring product quality. As an example, different categories based on the intended use of the model are models to support process design (usually low or medium-impact models as those for formulation optimization, process optimization, design space determination and scale-up), models to support analytical procedures (mainly chemometric models based on data generated by PAT, which are usually high-impact, especially if used for release testing), models for process monitoring and control (medium or high-impact models as multivariate statistical process control models for continuous process verification or models for feedforward process control).

PSE provides many tools for model development and application, and the pharmaceutical sector has the opportunity to benefit strongly from these mature tools. In general, mathematical models can be derived from first principles reflecting physical laws (e.g., mass and energy balances, heat transfer relations, etc.), from data (data-based models), from previous knowledge or from their combinations. Aside from the kind of model used, modeling cannot be thought as a stand-alone activity, but needs to be fully integrated with an experimental strategy. The benefit of using modeling during development should then be seen in reduced experimentation and reduced developmental resources. Accordingly, modeling would be the tool that allows both to guide smart decisions about fit for purpose experimentation and to provide more process understanding, by formalizing in mathematical terms the relationships between variables. This implies that the critical-to quality input variables have been identified and included in the model equations, thus accounting for their importance.

On the basis of the PAT framework (FDA, 2004c), models themselves can be considered as PAT tools. In fact, the classification of PAT tools described in §1.2.2 (multivariate tools, process analyzers, process control tools and continuous improvement tools) has defined the application fields of modeling in pharmaceutical development and manufacturing. This has opened the route for several studies employing classical PSE approaches in pharmaceutical applications, in order to demonstrate how modeling can be used to support QbD implementation (Gernaey *et al.*, 2012).

As multivariate tools for design and data analysis, several approaches based on chemometric, mechanistic and hybrid models have been proposed in recent years. Chemometric models have been widely used and are now generally accepted by the pharmaceutical community as tools for improving process knowledge especially in PAT applications (spectroscopy, image analysis, acoustic signals, etc.). The interest for multivariate data analysis methods like principal component analysis (PCA; Jackson, 1991), partial least-squares regression (PLS; Wold, 1983; Höskuldsson, 1988) and statistical design of experiments (DoE; Montgomery, 2005a) has tremendously grown after the PAT initiative, together with the diffusion of advanced characterization techniques (Hinz, 2006). Very recently, Rajalahti and Kvalheim (2011) and Pomerantsev and Rodionova (2012) have reviewed dozens of published case studies in pharmaceutics that, mainly in the last three years, combined analytical methods such as infrared (IR), near-infrared (NIR), Raman spectroscopy, hyperspectral and digital imaging, and other tools as X-ray diffraction, chromatography or nuclear magnetic resonance (NMR), with multivariate analysis tools able to analyze the lots of data these instruments allow to acquire quickly. The applications mainly range from the design of predictive models for the estimation of the API or excipient contents in tablets (Chalus *et al.*, 2005), liquids (Kim *et al.*, 2007), pellets (Mantanus *et al.*, 2010), or syrups (Ziémons *et al.*, 2010); to the characterization of polymorphs in mixtures (Blanco *et al.*, 2006); the design of models to monitor operations like crystallization (Pöllänen *et al.*, 2005), blending (Vanarase *et al.*, 2010), granulation (Halstensen *et al.*, 2006), drying (Peinado *et al.*, 2011), coating (Kucheryavski *et al.*, 2010), or end-product quality (Matero *et al.*, 2010); the prediction of physical properties for granules or tablets (Shah *et al.*, 2007); the visual characterization of product appearance (García-Muñoz and Camody, 2010b) or coating uniformity (García-Muñoz and Gierer, 2010c).

If multivariate methods have been largely employed to support analytical methods implementation and as soft sensors (Kadlec *et al.*, 2009), their use in pharmaceutical development and manufacturing with other purposes (e.g., process design and control, product transfer) is less common. Kourti (2006) provides a thorough review of the role of multivariate analysis beyond real-time analyzers and a more complete survey will be given in the next section (§1.4).

In general, to define the best process operating conditions and the possible control strategies, it would be highly desirable to have tools that allow simulating the behavior of the process *in silico*, without resorting to experimental campaigns, especially in large scale plants. Mechanistic models based on first principles enable to map the process knowledge in a series of input-output relationships, which reflect the physical behavior of the system. Although the formulation of a first-principles model requires deep knowledge of the physical phenomena occurring in a process, and adequate estimation of the model parameters is needed to use the model as a simulator and predictive tool, applications employing mechanistic models in terms of ordinary differential equations (ODEs), differential algebraic equations (DAEs) and partial differential equations (PDEs) have recently been proposed in the pharmaceutical literature (Gernaey *et al.*, 2012). Examples of mechanistic models based on ODEs can be found in Sin *et al.* (2008) for the modeling of an antibiotic production, or in the work of Zimermann *et al.* (2007) for the modeling of reaction in the synthesis of neuraminic acid.

Due to the unique challenges faced by the pharmaceutical industry, which is mainly characterized by batch productions in which solid materials are often manufactured, PDE models have been increasingly applied, especially in the form of population balance models (PBMs). These have been proposed to describe the dynamics of crystal size distribution in crystallization processes (Nagy *et al.*, 2008; Aamir *et al.*, 2010), for the description of the particle size distribution and the binder content in granulation processes (Poon *et al.*, 2009; Ramachandran *et al.*, 2009), blending operations (Boukouvala *et al.*, 2011) or milling processes (Bilgili and Scarlett, 2005), or the description of the moisture content in particles during drying (Mortier *et al.*, 2011). Other PDE applications that are increasingly being used in the pharmaceutical industry are those related to computational fluid dynamics (CFD) to simulate mixing, solid handling, separations and drying processes (Pordal *et al.*, 2002). Kremer and Hancock (2006) and Wassgren and Curtis (2006) have provided thorough reviews of the use of CFD for pharmaceutical unit operations. If the first applications of CFD were mainly oriented to the study of the flow of materials into the equipment, nowadays the trend and the challenge is toward the development of combined CFD-PBM models that can describe the change of distributed properties as a function of spatial coordinates inside a unit operation (Gernaey *et al.*, 2012). As an example, Woo *et al.* (2009) combined CFD and PBM to model an impinging jet crystallizer.

The need of technologies to describe the behavior of granular materials has contributed to resort to modeling tools able to describe the interactions between particles, other than between the particles and the fluid. Discrete element methods (DEM) are now often used to this purpose, especially for the simulation of powder mixing processes (Remy *et al.*, 2009; Dubey *et al.*, 2011). Ketterhagen *et al.* (2009) reviewed a series of applications of DEM in the pharmaceutical industry, whereas Adam *et al.* (2011) provided a specific example on the use of DEM within the definition of the design space for a blending process.

In many cases, even if a detailed mechanistic model can be written, its implementation entails a high computational burden, which prevents the use of the model in many real-time applications, such as those related to the control or optimization of the operation. For this reason, reduced-order models (Krasnyk *et al.*, 2012) and hybrid models mixing mechanistic models with a data-driven component (Doyle *et al.*, 2003) are often used. Akkisetty *et al.* (2010) reported for example the use of a neural network for representing the breakage function in a PBM describing the particle size distribution of a milled material.

The ultimate advantage of using modeling to describe pharmaceutical processes would be their implementation as control and optimization tools. Other than the computational requirements, the challenge of mechanistic models would be that of incorporating advanced online measurements (coming for example from spectroscopic instruments or online process analyzers) into the models. For these reasons, applications of advanced control methods, such as model predictive control (Hermanto *et al.*, 2011), have been limited so far. Some studies of the use of classical control theory tools for pharmaceutical processes have recently appeared in the literature (Ramachandran *et al.*, 2011; Singh *et al.*, 2012).

1.4 Latent variable models and QbD

Unlike other manufacturing industries, the pharma sector has to cope with some unique challenges in product development and manufacturing, such as a variety of production paths, multi-product low volume and mainly batch productions, the complexity of products, which are basically formulations of different raw materials (APIs, fillers, binders, disintegrants, lubricants, etc.), and, above all, a peculiar regulatory environment (García-Muñoz and Oksanen, 2010). These challenges contribute to complicate product and process design activities, because many materials and process conditions need to be tested in order to understand their impact on the final product quality.

Although mechanistic models would always be desirable to assist the implementation of the QbD paradigms, as they provide a transparent representation from first-principles of the relations between input variables (e.g., raw materials characteristics, process parameters) and product quality, the specific features of the pharmaceutical productions make their development and use particularly burdensome in most pharmaceutical development and manufacturing applications. For these reasons, pharmaceutical development has often relied on extensive experimental campaigns, aiming at generating data to increase the understanding on a process under development. As a consequence, pharmaceutical environments are usually characterized by the availability of large amounts of production and research data from development and manufacturing environments, being them obtained from designed experiments, on-going manufacturing processes, or historical products already developed. According to the QbD framework, pharmaceutical companies can benefit from a better management of these data, from which useful information for the development of new products and processes, process monitoring and control can be extracted.

There is therefore the need of developing systematic design and analysis tools that can be used throughout the variety typical of pharmaceutical productions, in order to give the opportunity for pharmaceutical development personnel to exploit optimally these data. The information extracted from these data can then drive process understanding and descision-making in product and process development, or support troubleshooting and process supervision in manufacturing environments. Latent variable models (LVMs) can represent appropriate modeling tools to better leveraging these data and respond to these needs.

LVMs are statistical models specifically designed to analyze massive amounts of (usually correlated) data. The basic idea behind LVMs is that the number of underlying forces acting on a system is much smaller than the number of available measurements taken on the system. Indeed, the forces that drive the system leave a similar signature on different measured variables, which in turn means that the measurements are correlated. LVMs enable the identification and the quantification of these driving forces thanks to the estimation of the model parameters. By combining the correlated variables, LVMs find new variables (the *latent variables*, LVs) that optimally describe the variability in the data, and can be useful in the identification of the driving forces acting on the system and responsible for the data variability. Figure 1.3 reports a geometrical interpretation of the operation performed when a LVM is built on a dataset $\mathbf{X}$ which collects 11 samples characterized by 3 measured variables x_n (n = 1,2,3).

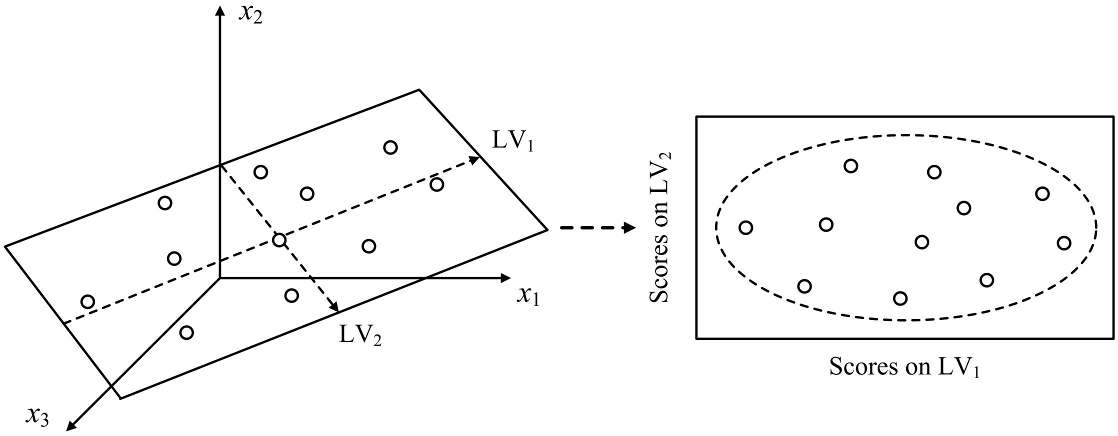


Figure 1.3. *Geometrical interpretation of the LVM built on the dataset* **X**.

As can be seen, the LVM transforms the three-dimensional **X** space into a two-dimensional space (the *latent space*) defined by the LV1 and LV2 directions. These indeed correspond to the directions along which the scattering (i.e. the variability) of the data is higher. The original **X** space can then be described by the latent space (on the right of Figure 1.3), and the projections (called *scores*) of the original variables (x_1, x_2 and x_3) on the LV space become then the new variables defining the state of the system.

There are several other advantages, other than dimensionality reduction, in using LVMs rather than the original variables to describe a system. Since LVs find the directions of maximum variability in the data, they can be easily interpreted, based on the engineering knowledge, to identify which are the driving forces acting on the system. Moreover, LVs are independent (orthogonal) and (assumed to be) normally distributed. This allows to use the probability theory to evaluate how new data are similar to the data used to build the model.

Other than modeling single spaces as in Figure 1.3, LVMs can be used to relate data from different datasets, as in latent variable regression models (LVRMs). These models are commonly associated to analytical instruments, to relate highly correlated input variables (e.g., spectroscopic variables) to response variables as product quality, as described §1.3. In general, they are powerful modeling tools in every situation in which the number of measured variables are large as compared to the number of runs/samples (Burnham *et al.*, 1996).

It must be noted that, while LVMs have found wide application as predictive tools, their importance (also in industrial applications) has not been limited to predictions. For example, they have been used for process understanding and troubleshooting (García-Muñoz *et al.*, 2003), for process operating conditions design (Jaeckle and MacGregor, 1998), for process control (Flores-Cerrillo and MacGregor, 2004), process monitoring (MacGregor and Kourti, 1995), process scale-up (García-Muñoz *et al.*, 2005) and also for product design (Muteki *et al.*, 2006) and optimization (Yacoub and MacGregor, 2004).

Details on the theoretical background behind LVMs and on the algorithms will be provided in Chapter 2. The main interest here is to show how LVMs can be feasibly used to support the *practical* implementation of QbD. Figure 1.4 provides a schematic of the steps described in §1.2.3 for the practical implementation of QbD in pharmaceutical development and manufacturing, together with some indications on how LVMs can be used. As can be seen, LVMs could have an important role in each of the QbD implementation steps reported in Figure 1.4. These steps can be summarized according to the three main objectives of QbD: product and process design, process understanding and process monitoring and control. In the following, some insights on how LVMs can be used as supporting tools in each step are presented, with a survey on reported applications of multivariate statistical approaches to face the above issues within pharmaceutical industries.

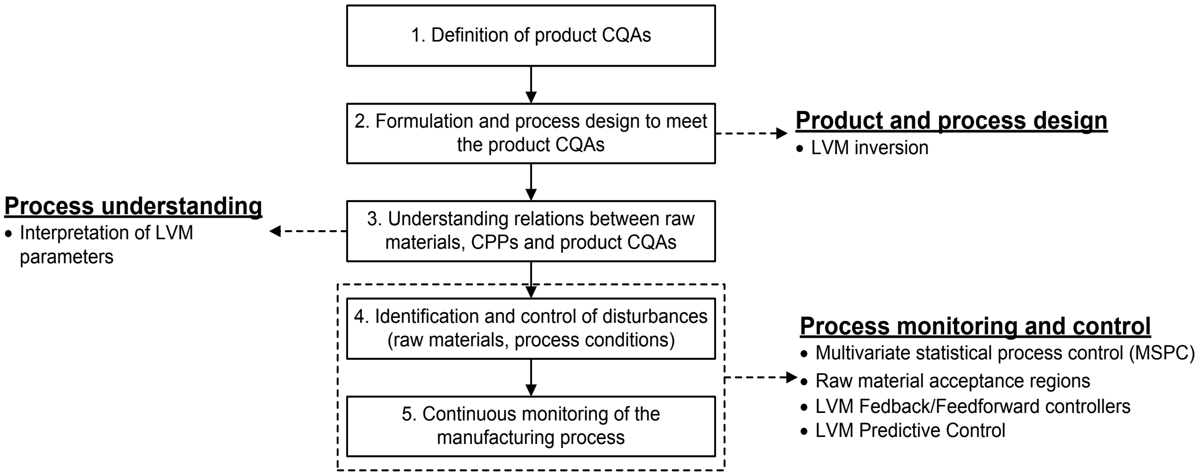

1. Definition of product CQAs
2. Formulation and process design to meet the product CQAs
3. Understanding relations between raw materials, CPPs and product CQAs
4. Identification and control of disturbances (raw materials, process conditions)
5. Continuous monitoring of the manufacturing process
Product and process design
• LVM inversion
Process understanding
• Interpretation of LVM parameters
Process monitoring and control
• Multivariate statistical process control (MSPC)
• Raw material acceptance regions
• LVM Fedback/Feedforward controllers
• LVM Predictive Control

Figure 1.4. *Schematic of the QbD implementation steps, with the indication of the possible use of LVMs in each step (adapted from Yu, 2008).*

1.4.1 Product and process design

The use of multivariate analysis in pharmaceutical development for product and process design has considerably increased after the introduction of the QbD framework. Contributions which show the application of multivariate statistical analysis to support the QbD implementation can be divided in three main categories: those based on DoE and response surface models, those combining design of experiments with LVMs, and those based on the inversion of LVMs built on historical databases.

1.4.1.1 Design of experiments

Prior to the QbD initiative, applications of statistical models for product and process design in pharmaceutical industry were limited to DoE tools (Montgomery, 2005a), with the aim of optimizing product formulations or processes. Gabrielsson *et al.* (2002) reviewed several applications of DoE and multivariate analysis in pharmaceutical applications, acknowledging that, at that time, DoE was very common in pharmaceutical development, especially for formulation design and product optimization. As an example, Campisi *et al.* (1998) used an experimental mixture design for the optimization of the theophylline solubility in a four-component blend. Ramabali *et al.* (2001) used a DoE strategy to generate data for modeling and optimizing a fluid bed granulation process.

After the introduction of the QbD initiative, the identification of the design space of the process has become the ultimate objective of the product/process design activity for pharmaceutical companies. A general guidance on the topic can be found in the work of Lepore and Spavins (2008), who outlined a series of approaches and steps for the development of a design space. Clearly, several different approaches have been proposed regarding the way a design space should be identified. Most of the industrial case studies in this field apply statistical DoE to explore the knowledge space and identify the regions within which parameter values are demonstrated to ensure the desired product CQAs. As a matter of example, am Ende *et al.* (2007) applied a QbD approach based on risk assessment and DoE performed on the parameters identified as CPPs, to define the design space for an API manufacturing process (Torcetrapib). A similar strategy was used by Burt *et al.* (2011) who exploited DoE results to develop a hybrid model to guide the development of the design space for a drug manufacturing process. In the work of Kapsi *et al.* (2012), an orthogonal design space for a compression-mix blending unit operation was developed by overlaying design spaces obtained for different product CQAs. Zacour *et al.* (2012a) constructed a tolerance-based design space from the combination of the response surface models of the single product CQAs, which reflected the probability of a given parameters combination to give CQAs within specifications.

1.4.1.2 Design of experiments and latent variable models

In the review of Gabrielsson *et al.* (2002) it was acknowledged that few examples were available on the use of multivariate data analysis methods in pharmaceutical applications, compared to DoE. After QbD was proposed by the FDA, multivariate analysis methods as LVMs have started to creep in pharmaceutical development environments, often coupled with DoE, with the main purpose of facilitating the choice of the parameters to include in a DoE analysis (e.g., on the basis of a PCA), or to discover the relationships between the input variables (design parameters or measured variables) and responses (e.g., using a PLS model), especially when the product CQAs were multivariate.

Bergman *et al.* (1998) presented a strategy based on sequential design of experiments and multivariate analysis (namely, PCA and PLS models) to optimize a multi-step process involving a granulation and a tabletting operation. Gabrielsson *et al.* (2003) used a strategy based on PCA to choose, within a large database, the excipients to test in a screening experimentation to define a pharmaceutical tablet formulation (multivariate design). The PLS model built between the excipient properties and the responses obtained from the experiments was later validated and used to design the formulation to give the desired tablet properties (Gabrielsson *et al.*, 2004). Lundsted-Enkel *et al.* (2006) reported a similar approach for a product formulation development, in which they underlined the usefulness of PCA for the analysis of the excipient databases, as interpretation of the material behavior was greatly improved, thus facilitating the choice of the excipient for formulation. Andemichael *et al.* (2009) were able to identify which batches were acceptable from an impurity point of view from a PCA on the IR spectra obtained from an API fermentation process, thus allowing to set specifications for process development in the synthesis of an antibiotic.

Another advantage in the use of multivariate models, which is often emphasized, is the possibility of analyzing several measured response variables using a single model. Andersson *et al.* (2007) demonstrated it with reference to an industrial case study related to the early development of a tablet formulation. In their work, the authors showed how a PLS model was able to cluster the response variables according to their correlation, thus guiding the experiments for the improvement of the model and for its subsequent use to guide the formulation design. Multivariate statistical analysis tools have also been used in the establishment of a design space to study the relationships among variables processed by DoE and those which are only measured. Huang *et al.* (2009) applied an approach based on DoE to perform the experiments, and then they used PCA and PLS to evaluate the impact of materials CQAs and CPPs on manufacturability and final product CQAs, with the aim of establishing the design space for a tablet manufacturing line involving high shear wet granulation, milling, blending, compression and coating units in small scale.

As PCA and PLS allow to model efficiently highly correlated data like those coming from online process measurements, they have often been used to include them in the analysis for the design space establishment. Streefland *et al.* (2009) used PCA and PLS to model data obtained from a DoE performed on a bacterial vaccine cultivation process. The techniques allowed to consider DoE parameters, online process measurements, NIR data, process variables related to process evolution and product CQAs in a unique model, in order to identify the design space for the process. In the study by Thirunahari *et al.* (2011), a design space for a batch cooling crystallization process was established using orthogonal PLS (OPLS; Trygg and Wold, 2002) and PCA to analyze attenuated total reflectance Fourier transform IR and Raman spectra, with the aim of favoring the desired form in a polymorphic system. Lourenço *et al.* (2012) reported a QbD study applied to an industrial pharmaceutical fluid bed granulation process. The authors acknowledged the usefulness of multivariate analysis (in the form of LVMs) in extracting information from the historical available datasets of the industrial process to increase process knowledge and guide risk assessment. This allowed to establish a better design space for the pilot scale process. Furthermore, the importance of multivariate analysis in finding correlations among different kinds of available data (process measurements, spectra) was emphasized, as well as the usefullness in identifying process patterns useful for process monitoring and control. Zacour *et al.* (2012b) used a strategy based on DoE and PAT to develop the design space for a fluid bed dryer, considering both raw material CQAs and CPPs. In the presented case study, a programmable logic controller was implemented to control the operation on the basis of the predictions obtained from a hybrid first-principles/PLS model (Zacour *et al.*, 2012c). This somehow agrees with the framework proposed by MacGregor and Bruwer (2008) for the development of design and control spaces for pharmaceutical operations. As stated by the authors, the design space in raw materials and in process parameters must be developed jointly, as changes in either one would affect the other. Furthermore, it is important to consider the control system the manufacturing plant will use while defining the design space, as changes in the control procedure would change the design space.

1.4.1.3 Latent variable model inversion

Other than the applications described above, LVMs, such as PCA and PLS, can have a prominent role in setting up a product and process design environment under a QbD framework, by analyzing the data from historical experiments and especially by exploiting data available from already developed products. If a LVRM relating raw material CQAs, CPPs (e.g., input variables) and product CQAs (response variables) is designed from historical data, it can be used to support product or process design or even to integrate them. Once the product CQAs have been defined (step 1 in Figure 1.4), the LVRM can be used to assist product and process design by using LVRM inversion technologies (Jaeckle and MacGregor, 1998 and 2000a; García-Muñoz *et al.*, 2006 and 2008). A LVRM will enforce the relationships drawn from the historical data or the performed experiments to calculate the optimal sets of raw materials and/or raw material properties (in case of product design), the optimal process operating conditions (in case of process design) or both in order to obtain a desired product, as the process output. In this way, a LVM can be used to guide the experimentation in developmental studies or for the definition of the process design space, which, as proposed by Kourti (2006) and demonstrated by García-Muñoz *et al.* (2010), can be defined directly in the LV space. Indeed, the analysis of historical data has also been suggested by ICH (2009) as a tool that can contribute to the establishment of a design space. Furthermore, if data from different plants are available, LVRMs can be inverted to support the transfer of products between different plants, namely to estimate the process conditions in a new plant in order to manufacture a product already developed in a reference plant (Jaeckle and MacGregor, 2000b). This is typical for example of process scale-up and is considered a highly risky and burdensome activity in pharmaceutical manufacturing.

Further details on LVRM inversion for product and process design will be provided in Chapter 4 of this Dissertation. In general, very few applications of model inversion for product/process design have appeared so far in the pharmaceutical literature. Very recently, some industrial case studies have been presented which applied LVRM inversion for the modeling and optimization of a tablet manufacturing line in which data from different formulations and unit operations (namely roller compactor and tablet press) were considered (Liu *et al.*, 2011a). A similar approach was proposed by Yacoub *et al.* (2011a) for the robust design and optimization of a whole manufacturing line involving wet granulation, drying, blending, compression and coating. In that work, it was shown how the LVRM identified the variable to manipulate in order to make the process robust to the raw material variability and lead to the introduction of a feedback control loop in the process to control in-process variables. Other contributions have shown how LVRM inversion could be used to support the scale-up of a roller compaction process (Liu *et al.*, 2011b), and to de-risk the scale-up of a challenging operation such as high-shear wet granulation, by using an optimized PLS model on a small scale plant to estimate the end point for the large scale operation, in order to respond to the API lot-to-lot variability (Muteki *et al.*, 2011).

Many other contributions have been proposed by practitioners and academia, to define the design space of a process based on data from experiments. To account for uncertainties in model parameters and for correlation between responses at assigned operating conditions, Bayesian approaches have been proposed (Peterson, 2008; Peterson and Yahyah, 2009) in alternative to traditional approaches, such as desirability functions or overlapping contours of different response surface models, commonly used in DoE (Stockdale and Cheng, 2009). Feasibility analysis techniques have recently been proposed to consider uncertainties in the model parameters and the process feasibility when developing design spaces using data-based approaches (Boukouvala *et al.*, 2010; Boukouvala and Ierapetritou, 2012).

1.4.2 Process understanding

As discussed in §1.2.1, process understanding involves all the activities related to the identification and management of the critical sources of variability affecting the product and process quality. In her review on the role of multivariate methods to implement PAT, Kourti (2006) emphasizes the role of multivariate analysis in pharmaceutical development for process understanding, by suggesting that tremendous insight into the process can be derived from data-based models like LVMs.

LVMs are in fact efficient tools in which the relationships among measured variables are transparent. Therefore, as mentioned earlier, LVMs parameters can be interpreted from first principles, enabling a deep understanding of the process and of the factors affecting a manufacturing operation. This type of modeling offers a tremendous advantage over other empirical/data-based models also from a regulatory point of view: as stated by the FDA (2004c), the predictive ability of a model has to reflect a higher degree of process understanding. The level of understanding achievable when black-box models (e.g., neural networks, expert systems, artificial intelligence) are used is usually lower than an LVM, due to the lack of transparency in the mechanics behind the predictions, even if the former can be more efficient than the latter for prediction (due to their nonlinear mapping capabilities). As a consequence, LVMs are preferable and accepted by the regulatory agencies to demonstrate process understanding in file submissions for approval. For all these reasons, LVM parameter interpretation has been indicated in Figure 1.4 as a useful tool for process understanding, to identify relations between raw material CQAs, CPPs and product CQAs. The consent demonstrated by the agencies toward these techniques has contributed to increase the number of published works, concerning mainly industrial case studies, which apply LVMs for process understanding.

Westerhuis and Coenegracht (1997) pioneered the use of a multi-block PLS model (MB-PLS, MacGregor *et al.*, 1994) to improve the interpretability of the model parameters and understand the critical variables in a two-step process consisting of granulation and tabletting. The MB-PLS model allowed both to include in the analysis the measurements on the intermediate product (the granules, which had a strong correlation with the final tablets) and to segregate the group of input variables (raw material compositions, granulator outputs, process parameters) in blocks, in order to study separately the influence of both groups on the tablet properties.

Soh *et al.* (2008) applied LVMs (PCA and PLS) to understand and model the effects of raw material properties (different grades of lactose and microcrystalline cellulose) and process parameters on the granule and ribbon properties obtained in a roller compaction process. Maltesen *et al.* (2008) reported an industrial case study in which PCA was used to identify the most important parameters and to find correlations between dependent and independent variables in a spray-drying process of insulin intended for pulmonary administration. In the work of Verma *et al.* (2009) multivariate regression techniques were used to identify the most critical parameters that affected nanosuspension preparation through microfluidization. Norioka *et al.* (2011) showed that multivariate statistical methods were useful to extract cause-effect relationships that allowed to understand on a scientific basis the process parameters more affecting the average and variance of the response variables in a solid form manufacturing process. A recent study of Dumarey *et al.* (2011) shows that analyzing data from an experimental design with OPLS improved the interpretability of the model. The authors applied this technique to enhance understanding on a roller compaction process, in which different grades of microcrystalline cellulose were tested at different process settings.

Oftentimes, multivariate methods are used to support the implementation of novel analytical technologies, which allow to improve understanding on a given process. An example is given by Lourenço *et al.* (2011), who used a microwave resonance technology (MRT) to monitor a fluidized bed granulation. The analysis through multiway PCA (MPCA; Nomikos and MacGregor, 1994) and multiway PLS (MPLS; Nomikos and MacGregor, 1995) of the MRT data allowed to discover a seasonality effect that affected the final granule size. Moreover, the use of a PLS model demonstrated that a relation between the particle size and the MRT measurements could be quantitatively established. This was found essential for the improvement of the process. In a similar study, Saerens *et al.* (2012) used in-line NIR spectroscopy to gain understanding on the polymer-drug interactions in a pharmaceutical hot-melt extrusion. The authors showed that NIR spectra indicated the presence of amorphous API and of hydrogen bonds between the polymer and the drug, thus demonstrating that the technology could be used to monitor the solid state behavior of the system, other than for determining the API concentration.

Other examples have been reported on the use of multivariate methods and LVMs to improve process understanding in technology transfer and process scale-up. Portillo *et al.* (2008) used analysis of variance to understand the impact of different blending parameters on the mixing rate in different scale blenders, in order to support the scale-up of a batch mixing process; the blender size was considered as a parameter in the experimental design. Kirdar *et al.* (2008) used MPCA and MPLS to identify scale-up differences and process parameters interactions that adversely impacted cell culture performances and product attributes in a biopharmaceutical application. García-Muñoz and Settell (2009) showed how PLS model parameters could be interpreted from first principles, allowing to identify the driving forces acting on a spray drying process. In the same study, they used a joint-Y PLS model (JY-PLS; García-Muñoz *et al.*, 2005) to understand the relationships between variables at multiple scales (pilot and commercial), identifying similarities that can be useful during scale-up.

Despite being used with reference to process or product development to clarify the impact of different input variables on the product or process quality, process understanding has also a significant role in product manufacturing, for example for process troubleshooting (García-Muñoz *et al.*, 2003) or root-cause analysis. This usually requires an offline analysis of the process data. An example can be found in the case study presented by García-Muñoz *et al.* (2009), who applied a JY-PLS model to determine the root cause for a bias found during the development of a multivariate calibration model for a NIR instrument coupled to a batch dryer. The proposed technique allowed to model jointly data from different batches with the laboratory data, while the interpretation of the model parameters was essential to isolate the cause of the observed drift, which was due to the nearness of the probe to the heating system port. Thanks to a MPCA model, Thomassen *et al.* (2010) were able to identify in the cultivation media the source of operational variation in the production of inactivated polio vaccine model. The obtained results led to an optimization in media preparation, resulting in a more robust composition. Furthermore, they acknowledged that a PLS analysis on the manufacturing data could not help in defining correlations between CPPs and product CQAs, being the former run at set points within strictly controlled ranges by the control system.

1.4.3 Process monitoring and control

The last two steps reported in Figure 1.4 pertain properly to pharmaceutical manufacturing and in particular to the use of LVMs for process monitoring and control. With the introduction of the QbD initiative and the PAT framework, the number of case studies employing analytical technology for process monitoring and control has tremendously increased (Chew and Sharratt, 2010). This has also been due to the fact that a process development activity cannot be considered under a QbD framework if an appropriate control strategy ensuring that the process is moving inside the design space has not been defined. In this context, LVMs such as PLS have found a wide range of applicability, especially when coupled to instruments to relate analytical measurements to product variables (e.g., concentration, moisture, particle size, etc.). In §1.3 some examples of these applications for process monitoring, end point determination and online product quality verification have already been reported. Chen *et al.* (2011) recently reviewed issues and challenges of multivariate statistical models in spectroscopic applications for real time process control, and described a practical system to enhance the robustness of closed loop control systems which include PAT instruments.

Other than multivariate calibration, LVMs can be directly used to analyze process measurements for process monitoring and control. As indicated in step 4 of Figure 1.4, LVMs identified from historical process data (e.g., experimental design procedures) can be used to implement feedback or feedforward controllers, aiming at identifying and responding to possible disturbances entering the system. One of the first examples applying LVMs to control a pharmaceutical process was presented by Westerhuis *et al.* (1997). The authors proposed a strategy to control a two-step process formed by wet granulation and tabletting, in which the process variables of the tabletting step could be adjusted depending on the granule properties to obtain tablets of desired properties. The control scheme was based on a grid of process parameter combinations, corresponding to different values for the tablet properties. García-Muñoz *et al.* (2010) proposed a feedforward controller based on a PLS model to compensate a high-shear wet granulation process for the observed changes in the properties of the incoming materials. The implementation of the controller on the process contributed to widen the range of acceptance of incoming materials and demonstrated to be useful in defining an integrated design space accounting for the network of complex relations between raw materials, process conditions and product quality. A similar exercise was recently proposed by Muteki *et al.* (2012), who implemented a feedforward controller in a tablet manufacturing process, involving blending, dry granulation, milling and tabletting. The model the controller relied on was built by considering different lots of raw materials: for different combinations of lots, the controller calculated the best parameters to operate the process in order to obtain a tablet of desired attributes. The rationale of the use of LVMs as control tools is based again on LVM inversion, where the model is inverted to estimate the manipulated variable values in order to ensure a desired set point or trajectory for the controlled variables, which often corresponds to the desired product quality. This rationale is the basis for LV model predictive control (Flores-Cerillo and MacGregor, 2004 and 2005; Wan *et al.*, 2012).

Given the statistical nature of LVMs, they can be employed for multivariate statistical process control (MSPC) in online process monitoring (step 5 in Figure 1.4). This is a well-known and applied use of LVMs in several types of industry (Kourti, 2005). Several studies in this field have recently been carried out also in the pharmaceutical industry. For example, García-

Muñoz and Settell (2009) used a PCA model built on common-cause variability data from an industrial spray drying process, to monitor the operation. They demonstrated how the model was effective in promptly detecting and identifying a fault that was occurring during the process. Burggraeve *et al.* (2011) developed a LVM procedure to monitor online a fluid bed granulation process. The strategies helped in identifying batches that gave poor quality products while the process was ongoing. Zomer *et al.* (2010) have showed how LVMs can be helpful in monitoring the development process itself, in a continuous quality verification framework. Multivariate tools are suggested to be used to review periodically the data as more knowledge is acquired during development, allowing to review the design space, change raw material CQAs or CPP values in the agreed design space, if needed.

The idea behind MSPC to establish multivariate "limits" to delineate operating regions, can be used to establish specifications for raw materials, as highlighted in step 5 of Figure 1.4. Defining an acceptance space for raw materials is fundamental for the pharmaceutical industry, where the number of materials employed in a formulation can become extremely high and affect the product quality only due to lot-to-lot variability. This was identified as a critical step also in the framework for the definition of design and control spaces proposed by MacGregor and Bruwer (2008). García-Muñoz (2009) has shown how a LVM can be used to relate data from different raw materials, scales and unit operations, while decoupling the effects of each contribution onto the product CQAs. This allows to define multivariate specifications for raw materials, independently of the process or scale in which they are used. In this way, a LVM can be used to support quantitatively the design space definition, separating the variability brought by the process operation, the control system and the raw materials.

1.5 Objectives of the research

Despite the number of studies on the application of modeling in pharmaceutical development and manufacturing has increased considerably in the last decade, in most published contributions *tailored* solutions to specific problems have been provided. However, there is a strong need to conceive *general* modeling strategies to support the implementation of QbD in the pharmaceutical industry. This need, together with the regulatory framework described earlier, provide the background and the motivation this Dissertation is intended for. The main objective of the research presented in this Dissertation is to demonstrate how LVMs can be feasibly used in a wide range of applications to assist the practical implementation of QbD paradigms into pharmaceutical development and manufacturing.

Stemming from the steps indicated in Figure 1.4, this Dissertation proposes different strategies based on LVMs to address many of the issues commonly encountered when developing new products and processes. The proposed strategies are of particular interest for the pharmaceutical community as they respond to the requirements dictated by the QbD initiative paradigms. The Dissertation presents innovative and *general* procedures for latent variable modeling within a pharmaceutical industry setting, and shows how LVMs can be used to support process understanding, product and process design (including the transfer of technology between different plants), and process monitoring and control.

LVMs are becoming popular among pharmaceutical experts as PAT tools or as modeling tools for analytical instruments. This Dissertation aims at demonstrating that LVMs can be used to assist any phase of the development of a pharmaceutical product, of its manufacturing and during the operation for commercial production. LVMs are shown to be extremely useful to *i*) optimally analyze and exploit historical data from known products or processes or from designed experiments, in order to draw understanding on the system under study, *ii*) accelerate development steps by guiding experiments for the achievement of the desired product properties and process performances, *iii*) find normal operating conditions that ensure the correct process operation and identify possible anomalies and actions to take for their correction.

Novel approaches are presented in this Dissertation (in the form of general procedures) to address problems commonly encountered in product and process development and manufacturing, for which either solutions have not yet been proposed, or the proposed solutions are very tailored to the specific application. The main application areas of the procedures proposed in this Dissertation and the innovative contributions they provide are summarized in the following.

- **Supporting process understanding**, in particular with reference to the design of continuous pharmaceutical processes. In the last few years, the use of continuous processing has greatly attracted the attention of pharmaceutical companies, traditionally based on batch processes (Plumb, 2005). For this reason, there is the need to have efficient tools to streamline the design of continuous pharmaceutical processes. The objective is therefore to provide a general framework for the use of LVMs to deal with the relevant features of continuous processes, such as the presence of different unit operations in the same manufacturing line, the possibility of disturbances entering different sections of the process, the propagation of these disturbances across the units and their effect on the final product. The framework should manage efficiently developmental data, extract information from them that are useful for process understanding in order to identify the main sources of variability and to establish a plantwide design space and control strategy.
- **Supporting the design of new products and processes**, and the transfer of a new developed product between the plant where it has been developed and a different manufacturing plant. Development environments are usually characterized by the presence of datasets that keep track of raw materials and product characterization, historical or new experiments designed for product or process development. Strategies that could exploit

efficiently these data to guide the experimentation for the design of new products and new processes would be needed. LVMs provide a useful tool to this purpose, through the implementation of appropriate model inversion technologies. However, since experimentation has to deal with many constraints, due for example to the used raw materials and to the operating limits, the inversion problem has to consider these constraints. The objective is therefore to provide a general framework for LVRM inversion that can deal with the several types of constraints commonly encountered in pharmaceutical development. The framework aim is to give a *systematic* tool that exploits the historical knowledge to suggest the optimal experiments to perform in product and formulation design, process design and in the transfer of products between different plants (e.g., scale-up), with the ultimate scope of accelerating development to reduce the time-to-market for new products.

- **Supporting process monitoring activities during technology transfer**, namely transferring models for process monitoring between different plants, in order to ensure that a process be in control since its start-up. From the perspective of process monitoring, one of the most common issues when transferring technologies between a reference plant and a target plant is that the target plant is desired to be under MSPC as soon as possible. An issue therefore arises on whether it is possible to transfer a monitoring model developed in the reference plant, where most of the experimentation can be carried out, to the target plant, when usually experiments are limited (e.g., when scaling-up a production). The objective is therefore to provide a general framework based on the use of LVMs to transfer knowledge between plants with the aim of having an efficient monitoring model for the target plant available since its start-up.

The effectiveness of the general procedures proposed in this Dissertation is demonstrated by applying each of them to experimental case studies. The next section presents a roadmap to the Dissertation and clarifies how the above-mentioned issues and objectives fit in the framework of Figure 1.4 for the practical implementation of QbD.

1.6 Dissertation roadmap

The research work carried out in this Dissertation focuses around the three milestones which the QbD initiative is founded on: process understanding, product and process design, and process monitoring and control (Figure 1.5). The relationship between these three milestones is not necessarily sequential: process understanding can be improved as more knowledge is gained during process design and during manufacturing (and this is the meaning of the continual process verification and improvement paradigms described in §1.2.1). However, by considering the three milestones from a modeling perspective and considering that the objective of the Dissertation is to support LV model-based design and process control, process understanding is believed to precede design, as a model is considered as the result of a process understanding activity.

Given this background, the general procedures described in §1.5 can be categorized as responding to common needs in process understanding, product and process design activities, and process monitoring and control. Figure 1.5 represents these three categories, reporting for each of them the specific applications which will be presented in the Dissertation. As a consequence, after the description of the LVM techniques and of the statistical modeling background in Chapter 2, the following Chapters of the Dissertation can be divided according to those three research areas.

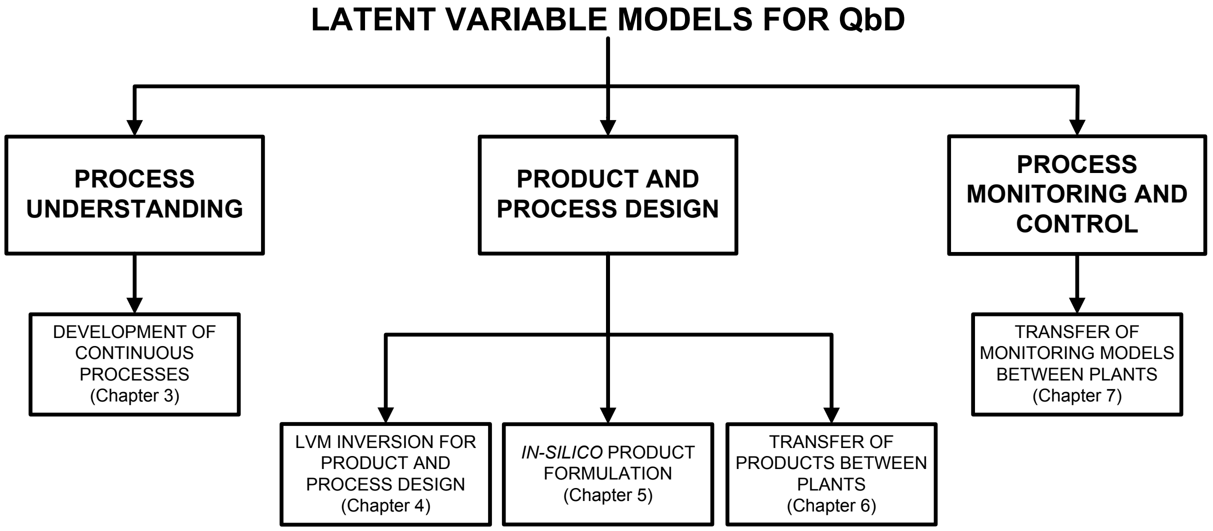

LATENT VARIABLE MODELS FOR QbD
PROCESS UNDERSTANDING
PRODUCT AND PROCESS DESIGN
PROCESS MONITORING AND CONTROL
DEVELOPMENT OF CONTINUOUS PROCESSES (Chapter 3)
LVM INVERSION FOR PRODUCT AND PROCESS DESIGN (Chapter 4)
IN-SILICO PRODUCT FORMULATION (Chapter 5)
TRANSFER OF PRODUCTS BETWEEN PLANTS (Chapter 6)
TRANSFER OF MONITORING MODELS BETWEEN PLANTS (Chapter 7)

Figure 1.5. *Dissertation roadmap with reference to the framework for the QbD practical implementation given in Figure 1.4.*

With respect to process understanding, a general approach is presented for the application of LVMs to aid the development of continuous manufacturing processes. This study will be presented in Chapter 3 and is applied to an industrial case study concerning a continuous tablet manufacturing pilot line involving different unit operations. Data on experiments performed considering raw materials that underwent different pre-treatments and different process settings were available. A framework is proposed based on a data management, an exploratory analysis and a comprehensive analysis steps. These steps are shown to be useful to understand respectively the main driving forces acting on each unit operations and how these propagate their effect into the system, impacting on the process performances and the product properties.

The central part of the Dissertation (Chapter 4, 5 and 6) is focused on the use of LVM inversion to assist product and process design and product transfer (Figure 1.5). In Chapter 4, the theory on LVRM inversion is revisited and a general framework is proposed, which can deal with different objectives and constraints commonly encountered in development activities. The framework aim is to formulate and solve the most appropriate inversion problem depending on the constraints that are provided by the user. For this reason, the inversion problem is formulated as an optimization problem. The inversion output will be a set of conditions with which the experiment should be performed in order to obtain the desired product. The framework effectiveness is tested for the design of the raw material properties to obtain granules of desired characteristics, assuming that a high-shear granulation process is used. Data on an industrial process (Vemavarapu *et al.*, 2007) are used to build the model and test the results. It is shown how the QbD definition of design space is linked to the mathematical concept of *null space*, which is intended as the multivariate space of the input conditions all corresponding, according to the model, to the same product properties (Jaeckle and MacGregor, 1998).

In Chapter 5, the framework for LVRM inversion is extended to an *in-silico* formulation design case study. Here, the objective is to provide an automatic tool for the formulation scientists in order to select the best excipients and their amount, which have to be mixed with a given API in a tablet formulation to obtain a blend of desired characteristics. The LVRM inversion problem is then adjusted to the specific objectives of the formulation problem (e.g., the maximization of the dose per tablet) and integrated with logical constraints, accounting for the material selection. A software with a user-friendly interface is developed to allow specifying the different objectives and constraints the user may have in terms of desired product properties or materials to be employed. The software solves a mixed-integer nonlinear programming problem (MINLP, Quesada and Grossman, 1992) and returns the formulation to be tested in order to obtain the desired blend.

In Chapter 6, the LVRM inversion framework proposed in Chapter 3 is applied to support the transfer of a product between different plants. The solution strategy is applied to an experimental case study concerning a nanoparticle manufacturing process through solvent displacement, using static mixers (Lince *et al.*, 2009). The objective is to obtain nanoparticles of desired mean size in a new static mixer, by exploiting the information acquired from the experiments performed in a static mixer of different size. An appropriate modeling technique (JY-PLS; García-Muñoz *et al.*, 2005) is implemented to relate data of different plants. The inversion is then performed with the aim of estimating the process conditions for the new mixer in order to obtain the desired mean nanoparticle size. This case study is also used to verify experimentally the theoretical concept of *null space*. The examples proposed in Chapter 4, 5 and 6 aim at demonstrating how LVRM inversion can be feasibly used as a systematic and science-based tool to accelerate experimentation in several phases of pharmaceutical development.

Finally, Chapter 7 addresses one of the most challenging issues in the design of systems for process monitoring and MSPC, namely the problem of transferring monitoring models between different plants (Figure 1.5). A framework is proposed for the use of LVMs to address this problem. The framework distinguishes between approaches merely based on data obtained online from the process, from approaches that combine the use of online measurements with the use of simple process knowledge coming from conservation laws (e.g., mass or energy balances). The different strategies are implemented to design a monitoring system for an industrial production-scale continuous spray-drying process, transferring knowledge acquired from a pilot-scale plant. The methods are then tested in the detection of a real fault. The framework is also extended to a simulated case study concerning a biopharmaceutical process for penicillin production via batch processing. The proposed examples show how the framework can be feasibly used to implement continuous quality assurance programs in pharmaceutical product manufacturing, where usually limited data are available if new productions are being started.

Chapter 2

Latent variable modeling background

This Chapter provides a general overview of the statistical and mathematical techniques that are applied in this Dissertation. First, a background on latent variable models (LVMs) is presented. Techniques like PCA and PLS are discussed from both an algorithmic and a practical point of view, together with advanced LVM tools that will be used in the following Chapters. Furthermore, the concepts of multivariate control charts and LVM inversion are introduced, and the fundamentals for their determination are provided.

2.1 Latent variable models

A latent variable model (LVM) is a statistical model that relates a set of *N* manifest (i.e. observable) variables to a set of *A* latent variables (LVs) which are unobservable and *explain* the dependence relationships between the manifest variables. LVs are found as linear combinations of the manifest variables in order to "summarize" their information content in an appropriate way according to the objective of the analysis (Varmuza and Filzmoser, 2009). Therefore, in order for an LVM to be useful, *A* should be significantly smaller than *N*.

Observed manifest variables are usually organized in a dataset $\mathbf{X}\ [I \times N]$, in which the *N* variables have been observed per *I* observations (or samples), or distinguished in a dataset $\mathbf{X}$ of regressors and a dataset $\mathbf{Y}\ [I \times M]$ of response variables.

In the first case, the objective of an LVM analysis is to explain the correlation structure of the *N* variables, in order to understand the relationships among them. Principal component analysis (PCA; Jackson, 1991) is one of the most useful techniques to this end.

In the second case, the objective of an LVM analysis is to explain the cross-correlation structure of the variables in $\mathbf{X}$ and in $\mathbf{Y}$, in order to study and quantify the relationships between regressors and response variables. To this purpose, a latent variable regression model (LVRM) as projection to latent structures (PLS; Wold *et al.*, 1983) can be used. In both cases, the manifest variables space is transformed into a lower-dimensional LV space: the coordinates of the *I* samples on this LV space provide a compressed representation of the observations, while the directions of the LV space provide a corresponding representation of the manifest variables. The general objectives of the different LVM techniques are therefore the same: *i*) data reduction and *ii*) data interpretation. In the following, theoretical and practical aspects of some LVM techniques are discussed.

2.1.1 Principal component analysis (PCA)

Principal component analysis (PCA; Jackson, 1991) is a multivariate statistical method that allows summarizing the information embedded in a dataset $\mathbf{X}$ of correlated variables, by projecting the data through a linear transformation onto a new coordinate system of latent orthogonal variables, which optimally capture the variability of the data and the correlation among the original manifest variables. Each of these coordinates identifies a latent direction in the data and is called principal component (PC).

To find the directions of the new coordinate system, a combination of the original variables in $\mathbf{X}$ is found which maximizes the variance of the data projections. For one PC, the optimization problem is represented by Eq.(1.2)[1].

$$\begin{aligned}&\max_{\mathbf{p}}\left(\mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}\right)\\&\text{subject to}\quad \mathbf{p}^{\mathrm{T}}\mathbf{p}=1\end{aligned}\quad , \tag{2.1}$$

where $\mathbf{p}$ is the $[N\times 1]$ vector of the combination coefficients, called *loadings*, that represent the director cosines of the PC, i.e. the latent direction of maximum variance in the data. The original data can be projected onto the PC direction, by obtaining a vector $\mathbf{t}$ $[I\times 1]$ of the coordinates into the PC space, called *scores*:

$$\mathbf{t}=\mathbf{X}\mathbf{p}\quad . \tag{2.2}$$

As a consequence, the problem in (2.1) can be reformulated as in (2.3), representing the maximization of the score vector length (Burnham *et al.*, 1996):

$$\begin{aligned}&\max_{\mathbf{p}}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}\right)\\&s.t.\ \ \mathbf{t}=\mathbf{X}\mathbf{p}\\&\qquad \mathbf{p}^{\mathrm{T}}\mathbf{p}=1\end{aligned}\quad . \tag{2.3}$$

The analytical solution of this problem is readily obtained from its optimality conditions (López-Negrete de la Fuente *et al.*, 2010) and is represented by the following eigenvalue problem:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}=\lambda\mathbf{p}\quad , \tag{2.4}$$

[1] In this Dissertation, the superscript $^{\mathrm{T}}$ attached to a vector/matrix is used to indicate its transpose.

where $\mathbf{p}$ corresponds to the eigenvector of the covariance matrix of $\mathbf{X}$ ($\mathbf{C}=\mathbf{X}^{\mathrm{T}}\mathbf{X}$), corresponding to the eigenvalue λ. λ is a measure of the variance explained by the product $\mathbf{tp}^{\mathrm{T}}$, namely the amount of information embedded in the model by the calculated PC.

The model loadings of a PCA model can therefore be determined from the eigenvector decomposition of the matrix $\mathbf{C}$, from which N $\mathbf{p}_n$ eigenvectors are determined. It results that, geometrically, the loadings are orthonormal:

$$\begin{cases} \mathbf{p}_n^{\mathrm{T}}\mathbf{p}_r = 0 & \text{for } n \neq r \\ \mathbf{p}_n^{\mathrm{T}}\mathbf{p}_r = 1 & \text{for } n = r \end{cases} \quad \text{with } n,r = 1,\ldots,N \quad . \tag{2.5}$$

Furthermore, the score vectors are orthogonal, as results from Eq.(2.2), Eq.(2.4) and Eq.(2.5), and have length equal to the eigenvalue associated to the corresponding PC:

$$\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_r = \mathbf{t}_n^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{t}_r = 0 \quad \text{for } n \neq r \quad \text{with } n,r = 1,\ldots,N \quad , \tag{2.6}$$

$$\mathbf{p}_n^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}_n = \mathbf{t}_n^{\mathrm{T}}\mathbf{t}_n = \lambda_n \quad \text{with } n = 1,\ldots,N \quad . \tag{2.7}$$

Eventually, when all the PCs have been determined, the dataset $\mathbf{X}$ can be viewed as the sum of the outer products of the N pairs of scores-loadings vectors:

$$\mathbf{X} = \sum_{n=1}^{N} \mathbf{t}_n \mathbf{p}_n^{\mathrm{T}} \quad . \tag{2.8}$$

Given the equivalence with the eigenvalue problem in Eq.(2.4), PCs are ordered according to the variance of the original dataset $\mathbf{X}$ they capture. When the columns (i.e. variables) of $\mathbf{X}$ are correlated, the $\mathbf{X}$ matrix is not full rank, and can be represented with a number A of PCs, such that $A \ll N$. If two or more original variables in $\mathbf{X}$ are correlated, they identify a common direction of variability. This direction can be described by a unique PC if a PCA analysis is performed. A single PC will therefore capture the variability of all the variables which are correlated along the direction identified by the PC. One of the most important contributions of PCA is therefore that it allows to describe the original dataset $\mathbf{X}$ with a lower number of variables, by projecting the data in $\mathbf{X}$ from the hyperspace of the original variables to the low-dimensional latent space of the PCs. As a consequence, the decomposition of $\mathbf{X}$ reported in Eq.(2.8) can be described by two terms: the sum of the outer products of the scores and loadings on the first A PCs of the models and the sum of the outer products of the scores and loadings vectors on the last $(N-A)$ PCs:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \sum_{a=A+1}^{N} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} \quad . \tag{2.9}$$

On this basis, only the first A PCs can be used to build a PCA model on $\mathbf{X}$ and to effectively describe its variability. The first A score vectors can be collected in the columns of a matrix of scores $\mathbf{T}$ $[I \times A]$, whose rows include the projections of the data samples of matrix $\mathbf{X}$. Analogously, the loadings of the first A PCs form the columns of a loading matrix $\mathbf{P}$ $[N \times A]$:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad , \tag{2.10}$$

where $\mathbf{E}$ is the $[I \times N]$ matrix generated by the last $(N - A)$ PCs of the model, which includes the residuals, if $\mathbf{X}$ is reconstructed using only the first A PCs:

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \quad , \tag{2.11}$$

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \quad . \tag{2.12}$$

Figure 2.1 reports the geometrical interpretation of the PCA model parameters, in a simplified case. A dataset $\mathbf{X}$ which collects 7 samples characterized by 2 measured variables x_n (n = 1,2), is plotted.

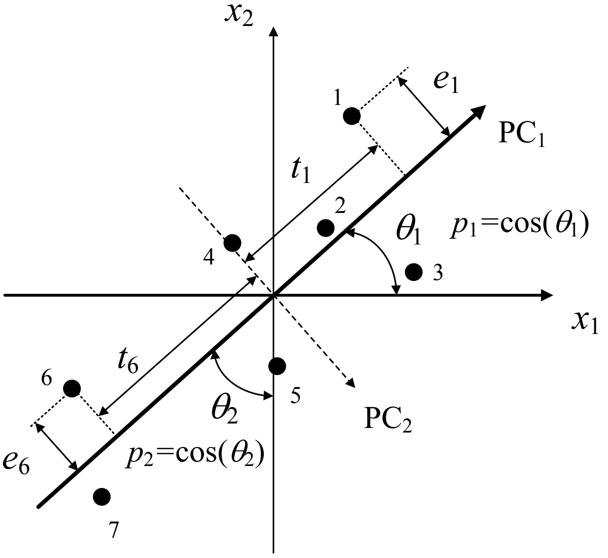


Figure 2.1. *Geometrical interpretation of the PCA scores and loadings for a dataset with 7 samples and 2 variables (x_1 and x_2).*

As it can be seen, data follow a defined trend in the (bi-dimensional) space of the original variables (x_n). If a PCA model is applied, the direction of maximum variability of the data is identified by PC1. The loadings of the model (p_1, p_2) represent the director cosines of PC1, namely the cosines of the angles between the latent directions and the axes of the original variable space. The scores represent the coordinates of the data samples of matrix $\mathbf{X}$ in the new reference system represented by PC1. The lack of representativeness of the data by the model is quantified by the residuals, represented by the perpendicular distances of the points from the line representing the first PC direction. In Figure 2.1, the second principal component that can be estimated from the data (PC2) is also reported as a dashed line. As it can be seen, PC2 is orthogonal to PC1 but accounts for a very limited variability in the data compared to PC1. In this case, it can be therefore concluded that PC1 is enough to adequately describe $\mathbf{X}$.

PCA model scores and loadings are usually plotted and interpreted to gain understanding on the similarity between different samples in the dataset (through scores) and on the correlation among the original variables (through loadings). Further details on the interpretation of the PCA scores and loading plots are provided in Appendix A. The parameters of a PCA model are usually calculated from the singular value decomposition (SVD; Meyer, 2000) of the $\mathbf{C}$ matrix or using the nonlinear iterative partial least squares algorithm (NIPALS; Wold, 1966; Geladi and Kowalski, 1986). Details on the algorithms used in this Dissertation are provided in Appendix B.

2.1.1.1 PCA data pretreatment

Before the model is developed, it is convenient to tailor the data to the analysis to be performed. For this reason data are often pre-treated, in order to better fulfill the important assumptions of the method. Pre-treatments depend on the characteristics of the available data and on the objectives of the analysis. They may include filtering, denoising, transformations, advanced scaling and data compression (Eriksson *et al.*, 2006)

In general, when dealing with datasets including many variables of different type and physical meaning (as process or development datasets) it is important that variables are weighted in a similar way, in order to exploit the PCA model to understand their importance. This can be achieved by auto-scaling, i.e. by mean-centering variables and scaling them to unit variance. Mean-centering consists in subtracting to each column $\mathbf{x}_n$ of the $\mathbf{X}$ matrix the mean value of the column itself. This is essential for the correct interpretation of the PCA model, as, if not mean-centered, principal components may identify as significant directions of variability in the data due to the differences between the variable mean values (Wise *et al.*, 2006).

The scaling to unit variance is performed by dividing each column of the $\mathbf{X}$ matrix by its standard deviation, so that the total variance of the column is equal to one. This is an essential step to make the analysis independent of the units of the variables and allows the simultaneous analysis of quantities which have different magnitudes. The scaling operation has also the advantage of partially linearizing data. Variables can undergo further scaling or weighting operations to determine a different impact of each variable on the model (Kourti, 2003). It is important to underline that when data in $\mathbf{X}$ are only mean-centered, matrix $\mathbf{\Sigma}$ represents the covariance matrix of $\mathbf{X}$, while if data are auto-scaled, it becomes the correlation matrix of $\mathbf{X}$. For this reason, correlations between variables can be identified from the loadings of a PCA model performed on auto-scaled data.

2.1.1.2 Selection of the number of PCs

An additional issue to be considered in building a PCA model is the determination of the dimensionality of the latent space of the model, namely the selection of the number of PCs to be used in the model. Several methods have been proposed in the literature to this purpose and the work of Valle *et al.* (1999) provides a thorough review and comparison of the most important ones.

In general, to select the appropriate number of PCs different issues should be considered, as the number of samples, the total variance explained, the relative size of the eigenvalues (i.e. the variance explained per component), and the subject-matter interpretations of the PCs (Johnson and Wichern, 2007). In this Dissertation three of the several available methods have been applied and are here presented:

- the scree test (Jackson *et al.*, 1991);
- the eigenvalue-greater-than-one rule (Mardia *et al.*, 1979);
- the cross-validation based on the prediction error sum of squares (Wold, 1978).

The scree test is an empirical and graphical procedure, which is based on the analysis of the profile of an index indicating the variability of the original data captured by the PCA model per PC (e.g., the explained variance R^2 per PC, the eigenvalues or the residual percent variance). The method is based on the idea that the variance described by the model should reach a "steady-state", when additional PCs begin to describe the variability due to random errors. When a break point is found in the curve or when the profile stabilizes, that point corresponds to the number of PCs to be included in the model. The implementation of the method is relatively easy, but if the curve decreases smoothly it can be difficult to identify an "elbow" on it.

The eigenvalue-greater-than-one rule is a simple rule for which all the PCs whose corresponding eigenvalues are lower than one are not considered in the model. The basic idea behind this method is that, if data are auto-scaled, the eigenvalue corresponding to a PC represents roughly the number of original variables whose variability is captured by the PC itself. If so, a PC capturing less than one original variable should not be included in the model. Although this method is very easy to implement and automate, in some cases PCs are discarded even if their eigenvalue is very close to one and their contribution to explain the systematic variability is significant. In these cases, it may be reasonable to lower the threshold in order to include PCs whose eigenvalue may be (slightly) lower than one.

Cross-validation (Wold, 1978) is another technique which can be employed in the determination of the number of PCs. The basic idea of cross-validation is that the number of PCs to be selected to build the model is the one for which the error in reconstructing new samples through the model is minimum. When no external validation data are available, the data in the $\mathbf{X}$ matrix itself are used to evaluate the reconstruction error (or prediction error sum of squares, *PRESS*). Different cross-validation algorithms can be employed. In general, the steps of the procedure are the following:

1. divide the $\mathbf{X}$ dataset in *G* subgroups $\mathbf{X}^g$ of *C* samples (with *g* = 1,…,*G*);
2. delete the samples in one of the $\mathbf{X}^g$ groups from the original dataset $\mathbf{X}$;
3. build a PCA model with the reduced dataset $\mathbf{X}$;
4. project the data in $\mathbf{X}^g$ in the PCA model built in step 3.;
5. compute *PRESS*$_g$ for the reconstruction of $\mathbf{X}^g$ and store it:

$$PRESS_g = \sum_{c=1}^{C}\sum_{n=1}^{N}\left(x_{c,n} - \hat{x}^g_{c,n}\right)^2 = \sum_{c=1}^{C}\sum_{n=1}^{N} e^2_{c,n} \quad , \tag{2.13}$$

   being $\hat{x}^g_{c,n}$ the reconstructed element of $\mathbf{X}^g$ in the *c*-th row and *n*-th column and $e^2_{c,n}$ the corresponding reconstruction error;
6. go back to step 1 to select the next subset until all the *G* subsets have been considered;
7. repeat the procedure by increasing the number of PCs used to build the PCA model.

By summing all the partial *PRESS* values per subgroup, eventually a total *PRESS* per PC is obtained. The evaluation of the *PRESS* profile can be useful to select the number of PC to build the model. Namely, a PC is included if it increases the predictive power of the model. Therefore, the number of PCs for which the minimum value of *PRESS* is found or for which a steady state in the profile is reached, should be considered. Relevant indices and statistical tests have also been proposed to support the analysis (Wold, 1978).

Note that if $G = I$, then $B = 1$. Therefore cross-validation deletes and reconstructs one sample at a time from the original dataset. This is analogous to a delete-1 jackknife approach (Duchesne and MacGregor, 2001). In general, cross-validation has been shown to be not reliable when autocorrelation or nonlinearities are present in the data, as happens in dynamic processes (Ku *et al.*, 1995). Therefore, unless needed for online applications, the selection of the PCs to be used in a PCA model should be based on the analysis of different criteria.

2.1.1.3 PCA diagnostics

There are several diagnostics which can be used to evaluate the performance of a PCA model. In general, model, variable and sample diagnostics can be distinguished (Eriksson *et al.*, 2001).

Among the model diagnostics, it is important to consider the amount of variability of the original data explained by the model, which is quantified by R^2 (for autoscaled data):

$$R^2 = 1 - \frac{\sum_{i=1}^{I}\sum_{n=1}^{N}\left(x_{i,n} - \hat{x}_{i,n}\right)^2}{\sum_{i=1}^{I}\sum_{n=1}^{N}\left(x_{i,n}\right)^2} = 1 - \frac{ESS}{TSS} \quad , \tag{2.14}$$

where *ESS* and *TSS* stand respectively for error sum of squares and total sum of squares. In (2.14) $\hat{x}_{i,n}$ represents the element in the *i*-th row and *n*-th column of the matrix $\hat{\mathbf{X}}$ reconstructed through the PCA model. R^2 is therefore calculated for different number of PCs included in the model and is also reported as a cumulative value (R^2_{CUM}). To evaluate the performances of the model with new samples, a similar statistic is used, exploiting the *PRESS* values calculated in cross-validation:

$$Q^2 = 1 - \frac{PRESS}{TSS} \quad . \tag{2.15}$$

As for R^2 and *PRESS*, the values of Q^2 are calculated for each PC and are usually cumulated (Q^2_{CUM}). Q^2 can be used in alternative to *PRESS* for the selection of the PCs to include in the model and can be seen as a measure of the "predictive" power of the model. Usually $Q^2 < R^2$.
The variation explained by the model for the dataset $\mathbf{X}$ can be reported also per variable included in the dataset, both in model calibration and in cross-validation. The same equations as in Eqs.(2.14)-(2.15) apply, limited to each column *n* of matrix $\mathbf{X}$. In (2.16), only the case of the explained variance per variable in calibration is reported ($R^2_{\mathrm{pv},n}$):

$$R^2_{\mathrm{pv},n} = 1 - \frac{\sum_{i=1}^{I}\left(x_{i,n} - \hat{x}_{i,n}\right)^2}{\sum_{i=1}^{I}\left(x_{i,n}\right)^2} = 1 - \frac{ESS_{\mathrm{pv}}}{TSS_{\mathrm{pv}}} \quad \text{with } n = 1,\dots,N \quad . \tag{2.16}$$

Beside diagnostics on the model performances, when a PCA model is built, it allows to calculate statistics on the data used for its calibration, in order to discover potential outliers or data that have a strong influence on the model. Two statistics are used to this purpose: the Hotelling's T^2 and the squared prediction error (SPE).
The Hotelling's T^2 statistic (Hotelling, 1933) measures the overall distance of the projections of an observation (i.e. a sample) of the $\mathbf{X}$ dataset from the PC space origin. Since each PC of the model explains a different percentage of variance of the data, the Mahalanobis distance is used to calculate it (Mardia *et al.*, 1979):

$$T_i^2 = \mathbf{t}_i^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \mathbf{t}_i = \sum_{a=1}^{A} \frac{t_{a,i}^2}{\lambda_a} \quad , \tag{2.17}$$

where $\mathbf{t}_i$ is the $[A \times 1]$ vector including the projections $t_{a,i}$ of the *i*-th observation on the *A* PCs used to build the model, while $\boldsymbol{\Lambda}$ is the $[A \times A]$ eigenvalue diagonal matrix. The T^2 statistic represents the multivariate generalization of the Student's *t*-test, and provides a check for observations adhering to multivariate normality (Eriksson *et al.*, 2001). In general, it is used to assess the deviation of an observation from the average conditions represented in the dataset. A sample with a large T_i^2 has a large influence to the model (high *leverage*) and should be handled with care: if it is well-represented by the model, the information it provides can be legitimate and useful to expand the data space and ensure the robustness of the model. The representativeness of the observation by the model is quantified through the SPE statistic:

$$\mathrm{SPE}_i = (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}}(\mathbf{x}_i - \hat{\mathbf{x}}_i) = \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_i \quad , \tag{2.18}$$

being $\mathbf{e}_i$ the $[N \times 1]$ vector of the residuals in the reconstruction of the *i*-th observation set $\mathbf{x}_i$. SPE_i measures the orthogonal distance of the *i*-th observation from the latent space identified by the model, namely it accounts for the mismatch of the model in representing $\mathbf{x}_i$: samples with a high value of SPE are characterized by a different correlation structure compared to the one described by the PCA model and, as a consequence, they are not well-represented by the model. In general, samples with high values of SPE but low values of T^2 have no influence on the model and do not provide information. Therefore, by removing them the model is unlikely to change. Nonetheless, they can be useful to establish good bounds for uncertainty. Statistical tests are established to evaluate if a sample should or not be considered an outlier based on the T^2 and SPE statistics. The discussion the confidence limits for these statistics is reported in Section 2.1.4.

Regardless of being pinpointed as an outlier, when a sample is reconstructed through a PCA model, it may be useful to identify the variables that are most responsible for its distance from the origin of the PC space or from the PC space itself. This can be done by analyzing the contributions of each variable in the $\mathbf{X}$ dataset to the T^2 and SPE statistics of the sample. In particular, the contributions to T^2 can be calculated as follows:

$$\mathbf{t}_{\mathrm{CONT},i}^{\mathrm{T}} = \mathbf{t}_i^{\mathrm{T}}\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^{\mathrm{T}} \quad . \tag{2.19}$$

$\mathbf{t}_{\mathrm{CONT},i}$ is a $[N \times 1]$ vector of the contributions of each variable to the Hotelling's T^2 statistic and can be considered a scaled version of the data within the PCA model. The formulation in (2.19) has the property that the sum of the squared elements of $\mathbf{t}_{con,i}$ gives T_i^2 for the *i*-th observation.

The contribution of each variable to the SPE_i statistic for the *i*-th sample coincides instead with the residuals in the reconstruction of the sample through the model (i.e. each single element $e_{i,n}$ of the *i*-th row of the residual matrix $\mathbf{E}$).

$$\mathrm{SPE}_{\mathrm{CONT},i,n} = e_{i,n} \quad . \tag{2.20}$$

The analysis of the variable contributions can reveal which variables mainly determine the position of a sample in the score space or out of it. This, together with physical knowledge on the system, may be useful especially when outliers are pinpointed, to understand the root cause of the problem. Procedures to calculate limits for the variable contributions have been proposed (Conlin *et al.*, 2000).

2.1.2 Projection to latent structures (PLS)

Projection to latent structures (PLS; Wold *et al.*, 1983; Höskuldsson, 1988) is a regression modeling technique which relates a dataset of regressors $\mathbf{X}$ to a dataset of response variables $\mathbf{Y}$ through the projections on their latent structures. As seen in the previous section, it is possible to represent a matrix $\mathbf{X}$ in terms of its score matrix $\mathbf{T}$. $\mathbf{T}$ can be directly related to $\mathbf{Y}$ instead of $\mathbf{X}$, since scores are orthogonal and are characterized by a high signal-to-noise ratio. This is performed in principal component regression (PCR; Geladi and Kowalski, 1986). PCR allows to solve problems due to the possible ill-conditioning of the $\mathbf{X}$ matrix, for which ordinary least square regression may not be feasible. However, PCR assumes that the main variability in $\mathbf{X}$, captured by the PC scores, is correlated to $\mathbf{Y}$, and that $\mathbf{Y}$ is full rank. Both of these assumptions are rarely satisfied, especially if $\mathbf{Y}$ is multivariate. Instead of relating the directions of maximum variability of $\mathbf{X}$ with $\mathbf{Y}$, PLS finds a transformation of the $\mathbf{X}$ data in order to maximize the covariance of its latent variables (LVs) with the $\mathbf{Y}$ dataset variables. For the first LV this is represented by the following optimization problem (Burnham *et al.*, 1996):

$$\begin{aligned}&\max_{\mathbf{w}_1}\left(\mathbf{w}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1\right)\\ &s.t.\quad \mathbf{w}_1^{\mathrm{T}}\mathbf{w}_1=1\end{aligned} \quad , \qquad (2.21)$$

where $\mathbf{w}_1$ is the $[N \times 1]$ weights vector for the first LV, representing the coefficients of the linear combination of the **X**-variables determining the PLS scores $\mathbf{t}_1$:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \quad . \qquad (2.22)$$

Note that the solution of the problem in (2.23) corresponds to the eigenvector decomposition of matrix $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$:

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1 = \lambda_1\mathbf{w}_1 \quad . \qquad (2.23)$$

In order to obtain the weight vectors for the further LVs, the problem in Eq.(2.23) may be solved iteratively using the deflated $\mathbf{X}_a$ and $\mathbf{Y}_a$ matrices. The deflation process, for $a = 1,\ldots,A-1$ being A the number of LVs to consider, is defined as:

$$\mathbf{X}_{a+1} = \left(\mathbf{I}_I - \frac{\mathbf{t}_a\mathbf{t}_a^{\mathrm{T}}}{\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a}\right)\mathbf{X}_a \qquad (2.24)$$

$$\mathbf{Y}_{a+1} = \left(\mathbf{I}_I - \frac{\mathbf{t}_a \mathbf{t}_a^{\mathrm{T}}}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} \right) \mathbf{Y}_a \quad , \tag{2.25}$$

where $\mathbf{I}_I$ is the $[I \times I]$ identity matrix. Namely, at the a-th step the reconstructions of each dataset from the a-th estimated LV are subtracted to the datasets themselves. In particular, from the second terms of Eqs.(2.24)-(2.25) it results that:

$$\mathbf{p}_a^{\mathrm{T}} = \frac{\mathbf{t}_a^{\mathrm{T}} \mathbf{X}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} \tag{2.26}$$

$$\mathbf{q}_a^{\mathrm{T}} = \frac{\mathbf{t}_a^{\mathrm{T}} \mathbf{Y}_a}{\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a} \quad , \tag{2.27}$$

where $\mathbf{p}_a$ and $\mathbf{q}_a$ represent respectively the $[N \times 1]$ and $[M \times 1]$ loadings in the reconstruction of $\mathbf{X}_a$ and $\mathbf{Y}_a$. Eventually, the datasets are decomposed and related through their latent structures:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2.28}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{2.29}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} \quad . \tag{2.30}$$

In Eqs.(2.28)-(2.30), $\mathbf{T}$ is the $[I \times A]$ score matrix, $\mathbf{P}$ and $\mathbf{Q}$ are the $[N \times A]$ and $[M \times A]$ loading matrices, while $\mathbf{E}$ and $\mathbf{F}$ are the $[I \times N]$ and $[I \times M]$ residual matrices accounting for the model mismatch. In (2.30), $\mathbf{W}^{*}$ is the $[N \times A]$ weight matrix, which is calculated from the weight matrix $\mathbf{W}$, to allow interpretation with respect to the original $\mathbf{X}$ matrix:

$$\mathbf{W}^{*} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \quad . \tag{2.31}$$

The advantage in using PLS is that it provides a model for the correlation structure of $\mathbf{X}$, a model for the correlation structure of $\mathbf{Y}$ and a model of their mutual relation. Therefore PLS is most suitable to handle reduced-rank datasets, in which highly correlated and possibly noisy data are included. More specifically, its basic assumption is that the spaces identified by $\mathbf{X}$ and $\mathbf{Y}$ have a common latent structure, which can be employed to relate them. Note that oftentimes in Eq.(2.29) of the PLS model the score matrix $\mathbf{T}$ is substituted by the $\mathbf{Y}$ space score matrix $\mathbf{U}$ $[I \times A]$, with $\mathbf{U} = \mathbf{TB}$ (called the inner relation; Geladi and Kowalski, 1986), providing than a completely bilinear structure for the PLS model. Even if this is not necessary, it is fundamental when dealing with nonlinear systems, for which the relations between $\mathbf{U}$ and $\mathbf{T}$ is nonlinear (Martens, 2001).

Figure 2.2 provides a geometrical interpretation of the PLS model: a dataset $\mathbf{X}$ $[20 \times 3]$ of regressor and a dataset $\mathbf{Y}$ $[20 \times 2]$ of response variables are considered. As can be seen, data in **X** arrange mainly on a plane, defined by two latent directions. Latent directions are identified in the **X** and in the **Y** space in order to best approximate the directions of maximum variability of the points in the original spaces and to provide a good correlation between the projections of the points themselves along these directions. As in the PCA case (Figure 2.1), the projections of the original points on these directions represent the PLS scores, while the loadings are the director cosines of the latent directions. Note that, while weights **W** are orthogonal in the **X** space, the loadings **Q** in the **Y** space may not necessarily be (Eriksson *et al.*, 2001).

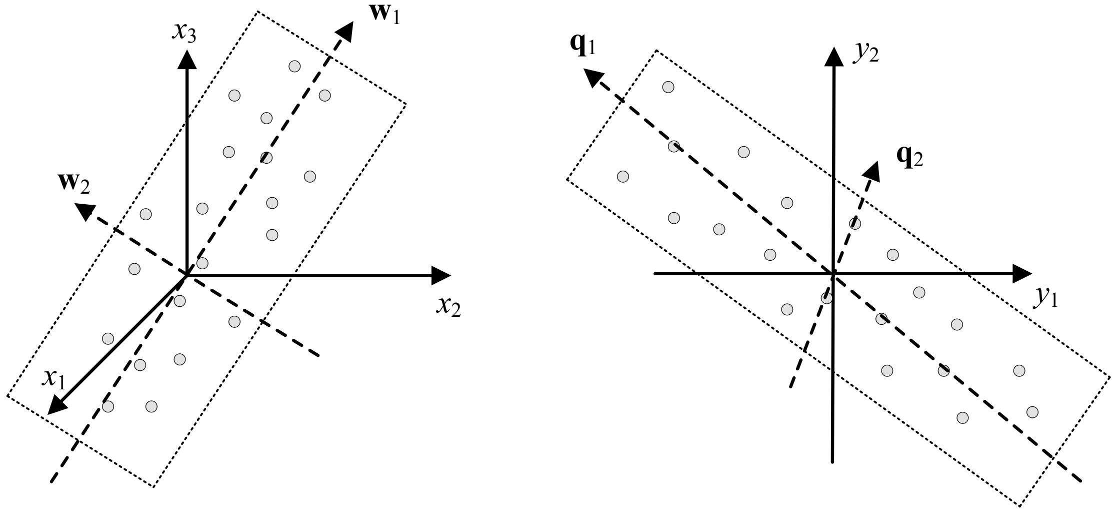


Figure 2.1. *Geometric interpretation of the PLS model decomposition in latent structures.*

As for PCA, PLS model scores, weights and loadings can be interpreted to gain understanding on the similarity between different samples and on the correlation among variables within and between datasets. Further details on the interpretation of the PLS scores and weights/loadings are provided in Appendix A. Several algorithms have been proposed in the literature to calculate the parameters of a PLS model, such as NIPALS (Wold *et al.*, 1983) and SIMPLS (de Jong, 1993). The advantage in using these algorithms instead of solving the eigenvector decomposition in Eq.(2.30) or the optimization problem in Eq.(2.29) is that they are iterative, allowing to stop after calculating a given number of LVs. Furthermore they can easily handle datasets with missing data, providing robust models. Details on the algorithms used in this Dissertation for PLS are provided in Appendix B. A thorough theoretical analysis on PLS modeling techniques can be found in the work of Höskuldsson (1988) and in the studies of Burnham and coworkers (1996, 1999a and 1999b).

2.1.2.1 Selection of the number of LVs and model diagnostics

In general, before applying a PLS model, appropriate data pre-treatments may be needed. The same considerations provided in Section 2.1.1 for PCA are valid for PLS. In particular, mean-centering and scaling to unit variance are the pre-processing methods applied throughout this Dissertation when applying PLS modeling.

Another issue related to the development of a PLS model is the selection of the number of LVs to include in the model. As in the PCA case, different methods have been proposed and compared in the literature (Li *et al.*, 2002; Wiklund *et al.*, 2007). One of the most popular methods, which is widely applied also in this Dissertation, is cross-validation. The procedure is the same described in Section 2.1.1.2 for PCA and is repeated increasing at each iteration the number of LVs included in the model, obtaining a profile of *PRESS* or alternatively the root mean square error of cross-validation per LV (*RMSECV$_a$*):

$$RMSECV_a = \sqrt{\frac{PRESS_a}{I \cdot N}} \quad . \tag{2.32}$$

In the regression case, the *PRESS* values are calculated on the basis of the predictions of the response variables in $\mathbf{Y}$. Accordingly, the number of LVs to consider should be the one for which *PRESS$_a$* (or *RMSECV$_a$*) is minimum. The profile of explained variance in cross validation (Q^2, Eq.(2.23)) provides a similar information. However, it must be underlined that a PLS model provides a model both for the $\mathbf{X}$ and for the $\mathbf{Y}$ datasets. Depending on the objectives of the analysis, it may be limiting to perform the cross-validation only on the $\mathbf{Y}$ matrix. In fact, as shown by Burnham *et al.* (1999a), if the dimensionality of the latent spaces of these two datasets is different, there could be LVs of a dataset not overlapped with the LVs of the other dataset. For this reason indices accounting for the variance captured for each dataset would be preferable to use. Further details on this issue will be given in Chapter 4.

Once a PLS model is built, the diagnostics to evaluate its performances are the same as for the PCA model (Eqs.(2.14)-(2.20)). In this case, they can be applied to both the involved datasets. Furthermore, for a successful calibration of the PLS model and for model interpretation, it may be useful to understand which are the regressor variables that most affect the projections and that are most appropriate to build the PLS model. This can be quantified by the VIP index (variable importance in the projection; Chong and Jun, 2005), which is defined as:

$$\mathrm{VIP}_n = \sqrt{N \sum_{i=1}^{A} R_{y,a}^2 \left(w_{n,a}\right)^2 \Big/ \sum_{i=1}^{A} R_{y,a}^2} \quad , \tag{2.33}$$

where *N* is the total number of variables considered, $R_{y,a}^2$ is the variance of **Y** explained by the *a*-th LV of the model, while $w_{n,a}$ is the weight of the *n*-th variable on the *a*-th LV calculated from the PLS model. By comparing the variables VIP, variables more relevant for explaining **Y** can be identified. Since the sum of squares of all the *N* VIPs is equal to the number of terms in the model, the average VIP would be equal to 1. Variables with $\mathrm{VIP}_n \geq 1$ are therefore considered valuable predictors of the variables in **Y** (Eriksson *et al.*, 2001).

2.1.3 Other latent variable modeling techniques

In the following Sections, a description of other advanced LVMs used throughout this Dissertation is provided. In particular, multi-block PLS and Joint-Y PLS are considered.

2.1.3.1 Multi-block PLS

Multi-block PLS (MB-PLS; Wangen and Kowalski, 1989) is an extension of the PLS method to consider multiple matrices (blocks) of data in a single model. The blocks can be both regressor and response variable matrices. This technique offers the advantage of improving the interpretability of the model in all the cases in which it is more convenient to keep variables in separate blocks rather than in a whole dataset. Blocking the available data can be justified for example by the different origin of the considered data, by the presence of variables with similar meaning and in different numbers, or by the need of understanding the relationships between variables in different blocks. This made MB models particularly attractive for the analysis of process data in which, for example, data from different plant sections or different unit operations needed to be considered separately (MacGregor *et al.*; 1994; Kourti *et al.*, 1995; Westerhuis and Coenegracht, 1997).

The MB-PLS algorithm can handle many types of pathway relationships between the blocks. Blocks can be left end blocks, which only predict subsequent blocks, right end blocks, which are only predicted by preceding blocks, or interior blocks, which are predicted by other blocks to their left but also predict blocks to the right of themselves (Westerhuis *et al.*, 1998).

The typical matrix structure handled by a MB-PLS is represented in Figure 2.3 for the case in which $B=2$ regressor blocks $\mathbf{X}_\mathrm{A}$ $[I\times N_\mathrm{A}]$ and $\mathbf{X}_\mathrm{B}$ $[I\times N_\mathrm{B}]$ and one response variables block $\mathbf{Y}$ $[I\times M]$ are considered. The basic idea of MB-PLS is to find a common latent space between blocks in order to maximize the covariance of the regressor block scores and the response variable blocks, while at the same time determining the latent structures proper of each block. In such a way, not only is the relation between each regressor block and the response variables optimized, but the model is also able to represent the relationships between different blocks. The practical advantage is that, in addition to a global latent space considering all blocks, a latent space of each block is available. This can be useful for example in monitoring a process formed by different units in a line (MacGregor *et al.*; 1994).

The parameters involved in a MB-PLS model are represented in Figure 2.3, for $A=1$ LV (i.e. the first iteration of the parameter estimation algorithm). They include the weights ($\mathbf{w}_\mathrm{A}$ $[N_\mathrm{A}\times 1]$, $\mathbf{w}_\mathrm{B}$ $[N_\mathrm{B}\times 1]$), loadings ($\mathbf{p}_\mathrm{A}$ $[N_\mathrm{A}\times 1]$, $\mathbf{p}_\mathrm{B}$ $[N_\mathrm{B}\times 1]$) and scores ($\mathbf{t}_\mathrm{A}$ $[I\times 1]$, $\mathbf{t}_\mathrm{B}$ $[I\times 1]$) of each block model and the super weights ($\mathbf{w}_\mathrm{S}$ $[B\times 1]$) and super scores ($\mathbf{t}_\mathrm{S}$ $[I\times 1]$, $\mathbf{u}$ $[I\times 1]$) of the combined regressor block model. The algorithm for their determination is described in Appendix B and is based on two levels: in the sub-level each block is used in a PLS cycle with $\mathbf{Y}$ to calculate the block scores; in the super level, the block scores are then combined in the super-block $\mathbf{T}_{\mathrm{MB}}$ and a PLS cycle between $\mathbf{T}_{\mathrm{MB}}$ and $\mathbf{Y}$ is performed, to calculate the super weights and super scores. The procedure is repeated until convergence of the super scores and for the *A* LVs considered to build the model (Westerhuis *et al.*, 1998).

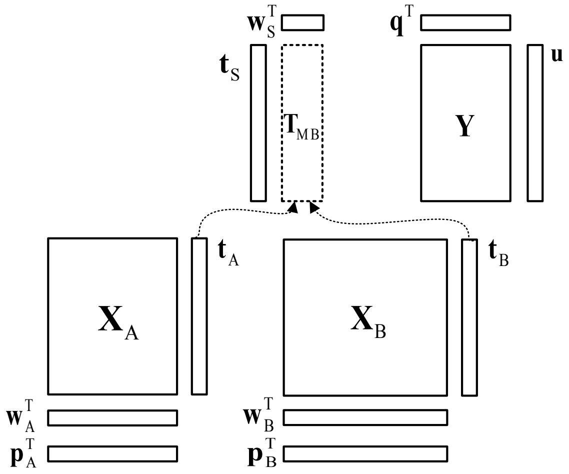

$\mathbf{w}_{\mathrm{S}}^{\mathrm{T}}$
$\mathbf{q}^{\mathrm{T}}$
$\mathbf{t}_{\mathrm{S}}$
$\mathbf{u}$
$\mathbf{T}_{\mathrm{MB}}$
$\mathbf{Y}$
$\mathbf{t}_{\mathrm{A}}$
$\mathbf{t}_{\mathrm{B}}$
$\mathbf{X}_{\mathrm{A}}$
$\mathbf{X}_{\mathrm{B}}$
$\mathbf{w}_{\mathrm{A}}^{\mathrm{T}}$
$\mathbf{w}_{\mathrm{B}}^{\mathrm{T}}$
$\mathbf{p}_{\mathrm{A}}^{\mathrm{T}}$
$\mathbf{p}_{\mathrm{B}}^{\mathrm{T}}$

Figure 2.3. *Schematic of the data structure and parameters of a MB-PLS model considering a single LV (adapted from Westerhuis et al., 1998).*

Westerhuis *et al.* (1998) have shown that it is possible to calculate the MB-PLS parameters based on the standard PLS method, if the appropriate variable scaling is applied. In particular, with reference to Figure 2.3, PLS can be applied between the auto-scaled $\mathbf{X}$ and $\mathbf{Y}$ matrices, where $\mathbf{X}$ is defined as:

$$\mathbf{X} = \left[\mathbf{X}_{\mathrm{A}}^{\mathrm{T}} \big/ \sqrt{N_{\mathrm{A}}} \quad \mathbf{X}_{\mathrm{B}}^{\mathrm{T}} \big/ \sqrt{N_{\mathrm{B}}}\right]^{\mathrm{T}} \quad . \tag{2.34}$$

In this case, the PLS scores $\mathbf{T}$ and $\mathbf{U}$ equal the MB-PLS super scores $\mathbf{T}_{\mathrm{S}}$ and $\mathbf{U}$. The block parameters can therefore be calculated from them. For each LV considered in the model and for the *b*-th block, it results that:

$$\mathbf{w}_{b} = \mathbf{X}_{b}\mathbf{u} \big/ \mathbf{u}^{\mathrm{T}}\mathbf{u} \tag{2.35}$$

$$\mathbf{t}_{b} = \mathbf{X}_{b}\mathbf{w}_{b}^{\mathrm{T}} \big/ \sqrt{N_{b}} \quad , \tag{2.36}$$

Furthermore, the super weights results from Eq.(2.44):

$$\mathbf{W}_{\mathrm{S}} = \mathbf{T}_{\mathrm{MB}}^{\mathrm{T}}\mathbf{U}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1} \quad . \tag{2.37}$$

The above equivalences between PLS and MB-PLS are valid only if there are no missing data in the datasets. Missing data in general decrease the performances of blocking methods.

For the MB-PLS model, the same considerations for the selection of the LVs as for the PLS model apply. The diagnostics used to evaluate the model are the same described above, and are extended in this case to the multiple blocks. An additional index can be used to quantify the importance of each block of variables in the projection, namely how important the variability included into a block to predict the response variables is. This index, called BIP (block importance in the projection; Yacoub *et al.*, 2011a), can be calculated as follows for the b-th block:

$$\mathrm{BIP}_b = \sqrt{B\sum_{a=1}^{A} R_{y,a}^2 \left(w_{b,a}\right)^2 \Big/ \sum_{a=1}^{A} R_{y,a}^2} \quad . \tag{2.38}$$

In Eq.(2.45) B is the number of blocks in the model, while $w_{b,a}$ the weight of the b-th block on the a-th LV calculated from the MB-PLS. A threshold at BIP = 1 can be set and the b-th block is considered significant if $\mathrm{BIP}_b \geq 1$.

2.1.3.2 Joint-Y PLS

Joint-Y PLS (JY-PLS; García-Muñoz, 2004; García-Muñoz *et al.*, 2005) is a latent variable regression model (LVRM) technique which allows to relate two or more regressor datasets (e.g., a dataset $\mathbf{X}_\mathrm{A}$ $[I \times N_\mathrm{A}]$ and $\mathbf{X}_\mathrm{B}$ $[J \times N_\mathrm{B}]$) through the joint space formed by their corresponding response variables datasets (e.g., $\mathbf{Y}_\mathrm{A}$ $[I \times M]$ and $\mathbf{Y}_\mathrm{B}$ $[J \times M]$). The basic idea of JY-PLS is that, if the correlation structure of the response variable datasets $\mathbf{Y}_\mathrm{A}$ and $\mathbf{Y}_\mathrm{B}$ is similar, a common latent space can be identified between $\mathbf{Y}_\mathrm{A}$ and $\mathbf{Y}_\mathrm{B}$. Assuming that the regressor datasets are correlated with the corresponding response variables, this common latent space (or part of it) will be spanned by (part of) the LVs of the regressor datasets. Otherwise stated, there will exist a region in the latent space of the matrix $\mathbf{Y}_\mathrm{J}$, obtained by joining the response variable datasets ($\mathbf{Y}_\mathrm{J} = \left[\mathbf{Y}_\mathrm{A}^\mathrm{T}\ \mathbf{Y}_\mathrm{B}^\mathrm{T}\right]^\mathrm{T}$), in which the LVs of the regressor datasets will be overlapped. This region can be exploited to relate the different datasets and to transfer information between them. To find this common region, the available datasets are decomposed on their latent structures in order to maximize at the same time the squared covariance between $\mathbf{X}_\mathrm{A}$ and $\mathbf{Y}_\mathrm{A}$, and between $\mathbf{X}_\mathrm{B}$ and $\mathbf{Y}_\mathrm{B}$, together with the squared joint covariance between them. For the first LV, this can be written as a variation of the PLS model objective function in Eq.(2.21):

$$\begin{aligned} &\max_{\mathbf{w}_1}\ \mathbf{w}_1^\mathrm{T} \begin{bmatrix} \mathbf{X}_\mathrm{A}^\mathrm{T}\mathbf{Y}_\mathrm{A}\mathbf{Y}_\mathrm{A}^\mathrm{T}\mathbf{X}_\mathrm{A} & \mathbf{X}_\mathrm{A}^\mathrm{T}\mathbf{Y}_\mathrm{A}\mathbf{Y}_\mathrm{B}^\mathrm{T}\mathbf{X}_\mathrm{B} \\ \mathbf{X}_\mathrm{B}^\mathrm{T}\mathbf{Y}_\mathrm{B}\mathbf{Y}_\mathrm{A}^\mathrm{T}\mathbf{X}_\mathrm{A} & \mathbf{X}_\mathrm{B}^\mathrm{T}\mathbf{Y}_\mathrm{B}^\mathrm{B}\mathbf{Y}_\mathrm{B}^\mathrm{T}\mathbf{X}_\mathrm{B} \end{bmatrix} \mathbf{w}_1 \quad , \\ &s.t. \quad \mathbf{w}_1^\mathrm{T}\mathbf{w}_1 = 1 \end{aligned} \tag{2.39}$$

where $\mathbf{w}_1$ is a $[(N_\mathrm{A} + N_\mathrm{B}) \times 1]$ common weight vector including the weights of dataset A and B ($\mathbf{w}_1 = \left[\mathbf{w}_{1,\mathrm{A}}^\mathrm{T}\ \mathbf{w}_{1,\mathrm{B}}^\mathrm{T}\right]^\mathrm{T}$). As in the PLS case, the solution of the problem in Eq.(2.39) is demonstrated to correspond to the eigenvector corresponding to the largest eigenvalue λ of the left hand matrix in Eq.(2.39) (García-Muñoz *et al.*, 2005):

$$\begin{bmatrix} \mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}}\mathbf{Y}_{\mathrm{A}}^{\mathrm{T}}\mathbf{X}_{\mathrm{A}} & \mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}}\mathbf{Y}_{\mathrm{B}}^{\mathrm{T}}\mathbf{X}_{\mathrm{B}} \\ \mathbf{X}_{\mathrm{B}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{B}}\mathbf{Y}_{\mathrm{A}}^{\mathrm{T}}\mathbf{X}_{\mathrm{A}} & \mathbf{X}_{\mathrm{B}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{B}}\mathbf{Y}_{\mathrm{B}}^{\mathrm{T}}\mathbf{X}_{\mathrm{B}} \end{bmatrix} \mathbf{w}_1 = \lambda_1 \mathbf{w}_1 \quad . \tag{2.40}$$

Once the weights in $\mathbf{w}_1$ have been calculated, the rest of the parameters for the JY-PLS model can be computed from them. As in the PLS case, the problem can be solved iteratively, using at each step the deflated versions of the considered datasets, in order to compute the parameters for all the LVs used to build the model. Eventually, the datasets are decomposed onto their latent structures:

$$\mathbf{Y}_{\mathrm{J}} = \begin{bmatrix} \mathbf{Y}_{\mathrm{A}} \\ \mathbf{Y}_{\mathrm{B}} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{\mathrm{A}} \\ \mathbf{T}_{\mathrm{B}} \end{bmatrix} \mathbf{Q}_{\mathrm{J}}^{\mathrm{T}} + \mathbf{E}_{\mathbf{Y}^{\mathrm{J}}} \tag{2.41}$$

$$\mathbf{X}_{\mathrm{A}} = \mathbf{T}_{\mathrm{A}}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}} + \mathbf{E}_{\mathbf{X}_{\mathrm{A}}} \tag{2.42}$$

$$\mathbf{X}_{\mathrm{B}} = \mathbf{T}_{\mathrm{B}}\mathbf{P}_{\mathrm{B}}^{\mathrm{T}} + \mathbf{E}_{\mathbf{X}_{\mathrm{B}}} \tag{2.43}$$

$$\mathbf{T}_{\mathrm{A}} = \mathbf{X}_{\mathrm{A}}\mathbf{W}_{\mathrm{A}}^{*} \tag{2.44}$$

$$\mathbf{T}_{\mathrm{B}} = \mathbf{X}_{\mathrm{B}}\mathbf{W}_{\mathrm{B}}^{*} \quad , \tag{2.45}$$

where $\mathbf{Q}_{\mathrm{J}}$ represent the $[M \times A]$ matrix of loadings defining the common latent space of $\mathbf{Y}_{\mathrm{J}}$, being A the number of LVs used to build the model. The meaning of the other symbols is the same as in the PLS model case. Indeed, the JY-PLS method calculate two separate PLS models for datasets A ($\mathbf{X}^{\mathrm{A}}$ and $\mathbf{Y}^{\mathrm{A}}$) and datasets B ($\mathbf{X}_{\mathrm{B}}$ and $\mathbf{Y}_{\mathrm{B}}$), whose spaces are however rotated to align with the common correlation structure of the variables in $\mathbf{Y}_{\mathrm{J}}$.

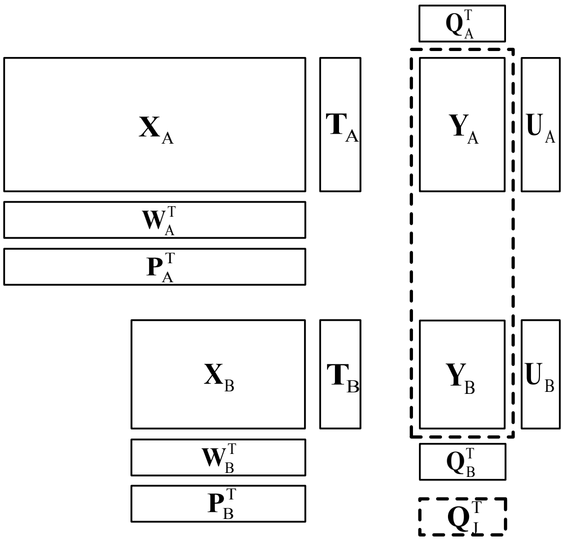

$\mathbf{Q}_{\mathrm{A}}^{\mathrm{T}}$
$\mathbf{X}_{\mathrm{A}}$
$\mathbf{T}_{\mathrm{A}}$
$\mathbf{Y}_{\mathrm{A}}$
$\mathbf{U}_{\mathrm{A}}$
$\mathbf{W}_{\mathrm{A}}^{\mathrm{T}}$
$\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}$
$\mathbf{X}_{\mathrm{B}}$
$\mathbf{T}_{\mathrm{B}}$
$\mathbf{Y}_{\mathrm{B}}$
$\mathbf{U}_{\mathrm{B}}$
$\mathbf{W}_{\mathrm{B}}^{\mathrm{T}}$
$\mathbf{Q}_{\mathrm{B}}^{\mathrm{T}}$
$\mathbf{P}_{\mathrm{B}}^{\mathrm{T}}$
$\mathbf{Q}_{\mathrm{J}}^{\mathrm{T}}$

Figure 2.4. *Schematic of the data structure and parameters of a joint-Y PLS model (adapted from García-Muñoz et al., 2005).*

Figure 2.4 reports the typical structure of the datasets used in a JY-PLS model, with the representation of the model parameters for each dataset. As can be seen, the JY-PLS model does not impose any restrictions on the number of columns in $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{B}}$, which can be different between the datasets, or on the number of observation per regressor dataset. The only restriction is that the number of columns in the response variable matrices must be the same.

Given the similarity with PLS, a modified version of the NIPALS algorithm has been proposed for the computation of the JY-PLS model parameters. The algorithm is described in Appendix B and is demonstrated to converge to the solution of the eigenvector problem in Eq.(2.40), provided that the appropriate scaling is performed on the considered matrices. In particular, the same scaling as for the MB-PLS model apply on the auto-scaled response variable datasets $\mathbf{Y}_{\mathrm{A}}$ and $\mathbf{Y}_{\mathrm{B}}$ (Eq.(2.34)). A similar scaling is performed also on the auto-scaled regressor matrices $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{B}}$ (García-Muñoz *et al.*, 2005):

$$\mathbf{X}_{\mathrm{A}} = \mathbf{X}_{\mathrm{A}} \Big/ \sqrt{I \cdot N_{\mathrm{A}}} \tag{2.46}$$

$$\mathbf{X}_{\mathrm{B}} = \mathbf{X}_{\mathrm{B}} \Big/ \sqrt{J \cdot N_{\mathrm{B}}} \quad . \tag{2.47}$$

As for the choice of the number of LVs and the model disgnostics, each PLS part of the JY-PLS structure can be seen as an independent model. Therefore cross-validation and diagnostics can be computed independently, giving for each dataset the same indices described in Section 2.1.1.3 (García-Muñoz, 2004).

Although for simplicity the JY-PLS method description has been limited to two regressor and response variable datasets, the method can be easily extended to consider datasets from multiple sources in a unique modeling framework (as will be shown in Chapter 6). Accordingly, the modified NIPALS algorithm can be generalized to calculate the sets of scrores and loading matrices for the multiple sources of data (*multi-site* JY-PLS). In the original work of García-Muñoz (2004), the technique is also extended to multiple block (MBJY-PLS) and nonlinear versions of the model.

2.1.4 Monitoring charts and control limits

Once a LVM has been calibrated on the available datasets, the model can be used to assess the overall conformance of a new sample $\mathbf{x}^{\mathrm{NEW}}$ to the data used to build the model (i.e. the historical data). This can be done by projecting $\mathbf{x}^{\mathrm{NEW}}$ onto the reduced latent space of the model, in order to calculate the corresponding scores $\hat{\mathbf{t}}^{\mathrm{NEW}}$ $[A \times 1]$:[2]

[2] The superscript $^{\mathrm{NEW}}$ is used throughout this Dissertation to distinguish new samples presented to the model from the ones used for model calibration.
The superscript $^{\wedge}$ is used to indicate that a variable is estimated and belongs to the LVM space.

$$\hat{\mathbf{t}}^{\mathrm{NEW}^{\mathrm{T}}} = \mathbf{x}^{\mathrm{NEW}^{\mathrm{T}}}\mathbf{P} \quad , \tag{2.48}$$

if a PCA model is used, or:

$$\hat{\mathbf{t}}^{\mathrm{NEW}^{\mathrm{T}}} = \mathbf{x}^{\mathrm{NEW}^{\mathrm{T}}}\mathbf{W}^{*} \quad , \tag{2.49}$$

if a PLS model is considered. The scores $\hat{\mathbf{t}}^{\mathrm{NEW}}$ can be used to calculate the Hotelling's T^2 (Eq.(2.19)) of the new sample ($T^2_{\mathbf{x}^{\mathrm{NEW}}}$), which provides a measure of the deviation of the new sample from the average conditions of the data used to build the model. Once the scores have been calculated, sample $\mathbf{x}^{\mathrm{NEW}}$ can be reconstructed from the model for **X**:

$$\hat{\mathbf{x}}^{\mathrm{NEW}} = \mathbf{P}\hat{\mathbf{t}}^{\mathrm{NEW}} \quad , \tag{2.50}$$

which is valid both for a PCA or a PLS model. Furthermore, in the case of the PLS model, a prediction of the response variables can be obtained by reconstructing $\hat{\mathbf{y}}^{\mathrm{NEW}}$ $[M \times 1]$:

$$\hat{\mathbf{y}}^{\mathrm{NEW}} = \mathbf{Q}\hat{\mathbf{t}}^{\mathrm{NEW}} \quad . \tag{2.51}$$

From $\hat{\mathbf{x}}^{\mathrm{NEW}}$ the value of the squared prediction error for $\mathbf{x}^{\mathrm{NEW}}$ ($\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$) can be obtained from Eq.(2.20). This statistic represents the model mismatch for the new incoming sample $\mathbf{x}^{\mathrm{NEW}}$.

The statistics $\hat{\mathbf{t}}^{\mathrm{NEW}}$, $T^2_{\mathbf{x}^{\mathrm{NEW}}}$ and $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$ provide therefore measures of the conformance of $\mathbf{x}^{\mathrm{NEW}}$ to the historical data. Confidence limits can be set for each of them, based on the values they assume for the data in model calibration. In particular, the scores have zero mean, variance equal to their associated eigenvalues and are orthogonal. Assuming that the data used to build the model are independent and identically distributed, scores are normally distributed. Therefore, for the scores on the a-th LV, an univariate confidence limit can be calculated from the critical value of the Student's t-distribution, with $I-1$ degrees of freedom at significance level α:

$$t_{(1-\alpha)\lim}(a) = \pm t_{I-1,\alpha/2} \cdot \sqrt{\lambda_a} \quad , \tag{2.52}$$

Under this assumption, the Hotelling's T^2 can be well-approximated as a Fisher's F-distribution, being it computed from the ratio of approximately normal variables (Eq.(2.19)). Its relevant confidence limit can therefore be estimated as (Mardia *et al.*, 1979):

$$T^2_{(1-\alpha)\lim}(A, I) = \frac{A \cdot (I^2 - 1)}{I \cdot (I - A)} \cdot F_{A,I-A,\alpha} \quad , \tag{2.53}$$

where $F_{A,I-A,\alpha}$ is the critical value of the *F* distribution with *A* and $I-A$ degrees of freedom at significance level α. This determines in the *A*-dimensional score space an ellipsoidal confidence region, whose semi-axes are:

$$sa_a = \sqrt{\lambda_a T^2_{(1-\alpha)\lim}(A,I)} \quad \text{with } a = 1,\ldots,A \quad . \tag{2.54}$$

In particular, to allow a visual representation, confidence ellipses can be determined through Eq.(2.54) for the projections of the scores of data in bi-dimensional planes.

The SPE statistic is a sum of squared errors, which can be assumed to follow a normal distribution. As a consequence, SPE can be approximated as a χ^2-distribution, and its relevant limit calculated as follows:

$$\mathrm{SPE}_{(1-\alpha)\lim} = [\nu/(2\cdot\mu)]\cdot\chi^2_{2\cdot\mu^2/\nu,\alpha} \quad , \tag{2.55}$$

where $\chi^2_{2\cdot m^2/\nu,\alpha}$ is the critical value of the χ^2-distribution with $2\cdot\mu^2/\nu$ degrees of freedom at the significance level α; μ and ν are respectively the mean and the variance of the SPE values of the data used to build the model (Nomikos and MacGregor, 1995b).

On the basis of the computed confidence limits, monitoring charts can be built for the scores, the Hotelling's T^2 and SPE. In particular, when a new sample is available, the mentioned statistics are compared with the relevant confidence limits to judge the similarity and the adherence of $\mathbf{x}^{\mathrm{NEW}}$ to the data used to build the model. Being multivariate indices, charts on T^2 and SPE are more effectively used to this purpose, by observing that:

$$\begin{cases} T^2_{\mathbf{x}^{\mathrm{NEW}}} \le T^2_{(1-\alpha)\lim} \\ SPE_{\mathbf{x}^{\mathrm{NEW}}} \le SPE_{(1-\alpha)\lim} \end{cases} \quad . \tag{2.56}$$

If the conditions in (2.56) are satisfied, $\mathbf{x}^{\mathrm{NEW}}$ is considered in a state of statistical control with a $100(1-\alpha)\%$ probability; otherwise an occurrence of *special cause* is detected. This occurrence may be due to a change in the mean conditions ($T^2_{\mathbf{x}^{\mathrm{NEW}}} > T^2_{(1-\alpha)\lim}$) or in the representativeness of the model ($SPE_{\mathbf{x}^{\mathrm{NEW}}} > SPE_{(1-\alpha)\lim}$) compared to the *common cause* data used to build the model. The procedure in (2.56) is equivalent to test the hypothesis that $\mathbf{x}^{\mathrm{NEW}}$ complies with the calibration (i.e. historical) data according to the T^2 and SPE statistics (Johnson and Wichern, 2007). If a problem is detected, the root cause can be identified by analyzing the relevant contributions, calculated as shown in Eq.(2.19) and Eq.(2.20). Note that the confidence limits calculation procedures can be applied also when using the LVRMs described in Section 2.1.3.

2.2 Latent variable regression model inversion

LVRMs are commonly used to predict a set of response variables $\hat{\mathbf{y}}^{\mathrm{NEW}}$ starting from an available set of regressors $\mathbf{x}^{\mathrm{NEW}}$. However, if a LVRM is available based on historical data and a set $\mathbf{y}^{\mathrm{DES}}$ $[M \times 1]$ of desired response variables is defined, the model can be used also to estimate the set of input variables $\mathbf{x}^{\mathrm{NEW}}$ which provide, according to the model, the desired responses $\mathbf{y}^{\mathrm{DES}}$. This can be achieved through model inversion and can be useful to assist product and process design or process control problems. Assuming that $\mathbf{y}^{\mathrm{DES}}$ is completely defined, the LVRM inversion for a PLS model estimates its projections $\hat{\mathbf{t}}^{\mathrm{DES}}$ onto the latent space of the model (Jaeckle and MacGregor, 1998):

$$\hat{\mathbf{t}}^{\mathrm{DES}} = \left(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y}^{\mathrm{DES}} \quad . \tag{2.57}$$

$\hat{\mathbf{t}}^{\mathrm{DES}}$ can be used in Eq.(2.50) to reconstruct the set of input variables $\hat{\mathbf{x}}^{\mathrm{NEW}}$ corresponding, according to the model, to $\mathbf{y}^{\mathrm{DES}}$ (direct LVRM inversion). In this way, $\hat{\mathbf{x}}^{\mathrm{NEW}}$ follows the same covariance structure of the historical data (Jaeckle and MacGregor, 1998). However, the solution from the LVRM inversion may be not unique, depending on the dimension of the latent spaces of the $\mathbf{X}$ and $\mathbf{Y}$ datasets (i.e. on their statistical rank) and on the number A of LVs used to build the LVRM.

Assuming that $R_{\mathbf{X}}$ is the statistical rank of matrix $\mathbf{X}$, while $R_{\mathbf{Y}}$ is the statistical rank of matrix **Y**, from the cross-validation performed as described in Section 2.1.1.2, it usually results that $A = \max(R_{\mathbf{X}}, R_{\mathbf{Y}})$. Depending on the ranks of the datasets, three different cases may arise in the inversion:

1. $A = R_{\mathbf{Y}}$ $\left(R_{\mathbf{Y}} \geq R_{\mathbf{X}}\right)$: in the most favorable case, i.e. when there is a substantial overlapping between the latent spaces of $\mathbf{X}$ and $\mathbf{Y}$ (Burnham *et al.*, 1999a), all the LVs of the $\mathbf{X}$ space can potentially have an effect on the **Y** space. In this case, the model inversion corresponds to a projection from a high dimensional space ($R_{\mathbf{Y}}$) to a lower dimensional space ($R_{\mathbf{X}}$).
2. $A = R_{\mathbf{X}}$ $\left(R_{\mathbf{X}} > R_{\mathbf{Y}}\right)$: in this case (which is the most common situation) there are some LVs (or their combinations) in the $\mathbf{X}$ latent space that are statistically significant for the description of the systematic variability in $\mathbf{X}$, but do not contribute in explaining the variability of the data in the $\mathbf{Y}$ space. Namely, they account for a part of the $\mathbf{X}$ data variability that is not related to the $\mathbf{Y}$ space (Burnham *et al.*, 1999a). In this case, a projection from a lower ($R_{\mathbf{Y}}$) to a higher ($R_{\mathbf{X}}$) dimensional space is required.
3. $A = R_{\mathbf{X}} = R_{\mathbf{Y}}$ but $\mathrm{rank}\left(\left[\mathbf{X}\,\mathbf{Y}\right]\right) > A$: in this case, even if the rank of the matrices is equal, the rank $R_{\mathbf{XY}}$ of matrix $\left[\mathbf{X}\,\mathbf{Y}\right]$ is greater, meaning that there are $R_{\mathbf{XY}} - A$ latent dimensions which do not overlap between the $\mathbf{X}$ and $\mathbf{Y}$ latent spaces. The situation is therefore similar to the one described in the previous point.

In the first case, direct model inversion in Eq.(2.57) can be applied, giving the least-squares projection onto the model latent space. A unique solution therefore exists. In the second and the third cases, the inversion problem is underdetermined and the set of solutions is infinite. The direct model inversion in Eq.(2.57) provides again the least-squares solution to the problem. However, this solution can be moved along the directions of the latent space not affecting $\mathbf{Y}$ (thus changing $\hat{\mathbf{t}}^{\mathrm{DES}}$), providing *the same* set of desired response variables $\mathbf{y}^{\mathrm{DES}}$. These latent directions form the *null space*, which is a subspace of the model space representing the *locus* of the $\mathbf{X}$ projections with no influence on the quality space (Jaeckle and MacGregor, 1998). The direct inversion solution can therefore be moved along the null space, in order to find the solution that achieves the objectives and satisfies constraints the inversion problem may have (e.g., in product or process design). For this reason, optimization approaches have been proposed to solve LVRM inversion (García-Muñoz *et al.*, 2006 and 2008), which also allow to deal with cases in which the values in $\mathbf{y}^{\mathrm{DES}}$ are not completely specified, but ranges (i.e. inequality constraints) are assigned to some/all the response variables. These approaches will be thoroughly reviewed and discussed in Chapter 4 of this Dissertation. In the following section, insight is provided on the computation of the null space.

2.2.1 Null space computation

The null space (or kernel) of a generic matrix $\mathbf{A}\ [I \times N]$ is defined as the set of vectors $\mathbf{x}$ for which $\mathbf{Ax} = \mathbf{0}$ (Meyer, 2000). As seen above, when in a LVRM $A = R_{\mathrm{X}}$ and $R_{\mathrm{X}} > R_{\mathrm{Y}}$ a null space exists. This means that, when a new sample $\mathbf{x}^{\mathrm{NEW}}$ is presented to the model, the prediction of the response variables $\hat{\mathbf{y}}^{\mathrm{NEW}}$ can be seen as formed by two latent contributions: *i*) a contribution $\mathbf{t}^{\mathrm{NEW}}\ [A \times 1]$ due to the effective scores of $\hat{\mathbf{y}}^{\mathrm{NEW}}$ in the latent space of the model; *ii*) a contribution $\mathbf{t}^{\mathrm{NULL}}\ [A \times 1]$ accounting for the translation of the scores along the null space in order to provide the reconstruction $\hat{\mathbf{x}}^{\mathrm{NEW}}$ at minimum distance from the latent space of the model (i.e. minimum SPE).

$$\hat{\mathbf{y}}^{\mathrm{NEW}} = \mathbf{Q}\left(\mathbf{t}^{\mathrm{NEW}} + \mathbf{t}^{\mathrm{NULL}}\right) \tag{2.58}$$

$$\hat{\mathbf{x}}^{\mathrm{NEW}} = \mathbf{P}\left(\mathbf{t}^{\mathrm{NEW}} + \mathbf{t}^{\mathrm{NULL}}\right) \quad . \tag{2.59}$$

The latent space is such that in Eqs.(2.58)-(2.59) $\mathbf{Qt}^{\mathrm{NULL}} = \mathbf{0}$, while $\mathbf{Pt}^{\mathrm{NULL}} \neq \mathbf{0}$, namely the null space is needed for the model to represent adequately the regressor variables, but it does not contribute in explaining the variability in the response variables. Considering the LVRM parameters, the null space represents therefore the kernel of the loadings $\mathbf{Q}$ matrix. As a consequence, the null space can be computed from the singular value decomposition of matrix $\mathbf{Q}$ (Jaeckle and MacGregor, 2000a):

$$\mathbf{Q} = \mathbf{U_Q S_Q V_Q^{\mathrm{T}}} = \mathbf{U_Q S_Q}\left[\mathbf{G}_1 \vdots \mathbf{G}_2\right]^{\mathrm{T}} \quad , \tag{2.60}$$

where $\mathbf{U}_{\mathbf{Q}}$ is the matrix of the left singular vectors of $\mathbf{Q}$, $\mathbf{S}_{\mathbf{Q}}$ is the diagonal matrix of the singular values of $\mathbf{Q}$ and $\mathbf{V}_{\mathbf{Q}} = [\mathbf{G}_1 \vdots \mathbf{G}_2]$ is the matrix of the right singular vectors of $\mathbf{Q}$. In particular, the right singular vectors corresponding to the vanishing (zeros) singular values of $\mathbf{Q}$ spans the null space of $\mathbf{Q}$. These are included in the columns of matrix $\mathbf{G}_2$ $(A \times (A - R_{\mathbf{Y}}))$, which therefore defines the null space of the model. $\mathbf{t}^{\mathrm{NULL}}$ can therefore be moved arbitrarily along it, without affecting $\hat{\mathbf{y}}^{\mathrm{NEW}}$:

$$\mathbf{t}^{\mathrm{NULL}^{\mathrm{T}}} = \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{G}_2^{\mathrm{T}} \quad . \tag{2.61}$$

In Eq.(2.61), $\boldsymbol{\gamma}$ is a $[(A - R_{\mathbf{Y}}) \times 1]$ vector arbitrary in magnitude and direction. In LVRM inversion, the null space for a desired response variable set $\mathbf{y}^{\mathrm{DES}}$ is computed by imposing that the direct inversion solution projections $\hat{\mathbf{t}}^{\mathrm{DES}}$ belong to the null space of matrix $\mathbf{Q}$. The regressor sets belonging to the null space are then reconstructed according to Eq.(2.59), where $\mathbf{t}^{\mathrm{NEW}} = \hat{\mathbf{t}}^{\mathrm{DES}}$, while $\mathbf{t}^{\mathrm{NULL}}$ is calculated from Eq.(2.61). This ensures that $\hat{\mathbf{x}}^{\mathrm{NEW}}$ adheres to the historical data covariance structure, but it is not ensured that it lies in the range of the historical data, nor that possible constraints assigned to the regressors are satisfied. The above-mentioned optimization approaches are needed to address these issues and will be described in Chapter 4.

Chapter 3

Latent variable models to support process understanding[*]

This Chapter shows how latent variable models (LVMs) can be used as tools to gain process understanding in the development of new pharmaceutical processes. In particular, the use of continuous manufacturing systems is examined. A general procedure is proposed to deal with data referred to different raw materials and different units along the production line. An industrial continuous tablet manufacturing process is used as a test bed for the analysis. It is shown how LVM parameters can be interpreted to identify and to rank the main driving forces acting on the system, providing a starting point to guide a comprehensive and science-based quality risk assessment and the definition of a robust control strategy.

3.1 Introduction

As discussed in Chapter 1, process understanding is an essential step for the implementation of QbD in both pharmaceutical development and manufacturing. Any product and process design and control activity cannot be carried out without the full identification and understanding of the driving forces acting on the system, and of the critical sources of variability that can affect the product and process quality. Under this perspective, the interest towards modeling as a tool to integrate experience-based knowledge in quality risk assessment acivities on pharmaceutical processes has increased, as demonstrated by the several contributions recently appeared in the literature and reviewed in Chapter 1 (Section 1.3 and Section 1.4).

On a parallel side, as a part of the path towards the use of innovative manufacturing systems, in the recent years the interest of the pharmaceutical industry has been focused on the transition of the production from batch to continuous processes, due to the advantages that continuous operations have compared to the batch ones (Plumb, 2005; Schaber *et al.*, 2011), such as reduction of the manufacturing time, maximization of the product yield, and

[*] Tomba, E., M. De Martin, P. Facco, J. Robertson, S. Zomer, F. Bezzo and M. Barolo. General approach to aid the development of continuous pharmaceutical processes using multivariate statistical modeling – An industrial case study. *Int. J. Pharm.*, in press. DOI: 10.1016/j.ijpharm.2013.01.018.

minimization of wastes and energy consumption. Furthermore, continuous plants are usually smaller than batch plants on equal productivity, easier to scale (thus simplifying technology transfer activities) and offer indisputable advantages from a control and safety point of view (Leuenberger, 2001).

One of the main advantages of continuous manufacturing is that it allows to link different processing units into a single manufacturing line, so that the transformations from the raw materials to the final products can occur without interruption. Due to the nature of the available unit operations, different configurations are feasible within the process stream (Boukouvala *et al.*, 2012). A proper selection of the optimal process operating conditions requires the in-depth understanding of the main driving forces acting on each processing unit and on the whole system to obtain a product of acceptable and reproducible quality. The different nature of the raw materials used and the complexity in their characterization, as well as the lack of detailed physical models (or of parameters therein) describing each processing unit, have hindered the development of deterministic models to describe pharmaceutical processes. For this reason, multivariate statistical techniques, and in particular latent variable models (LVMs), can provide a very useful tool to improve process understanding, by identifying the most important mechanisms acting on a manufacturing system and affecting the product attributes.

As pointed out in Chapter 1 (Section 1.4), LVMs are particularly useful in analyzing product and process developmental data, and the interest towards multivariate statistical modeling for QbD has recently grown especially with respect to the identification of the design space of pharmaceutical processes (MacGregor and Bruwer, 2008). In most of the case studies presented in Chapter 1 (Sections 1.4.1 and 1.4.2), the design space is identified by carrying out some designed experiments and a multivariate analysis of the resulting data (Huang *et al.*, 2009; Lourenço *et al.*, 2012; Zacour *et al.*, 2012b). However, development environments are often characterized by the presence of limited datasets, which additionally might be sparse and unstructured or sub-optimally designed in reflection of the product/process development history. When developing a continuous manufacturing process that comprises different processing units in the same manufacturing line, wherein raw materials from different sources can be processed, additional issues arise on how data from each single unit can be analyzed jointly with data from the other units to obtain information on the entire process, and on how different raw materials and different processing conditions within each unit impact on the intermediate and final product quality. This is essential in the definition of the design space for a continuous process, which should be preferably thought as a whole, instead of defining it as the combination of the design spaces of the single unit operations (Chapter 1, Section 1.2.2).

It has been pointed out recently (Gernaey *et al.*, 2012) that extraction of knowledge from available data still represents a bottleneck, and that future research should focus exactly on this issue to improve pharmaceutical manufacturing. Under this premise, in this Chapter a general procedure is proposed for the application of LV methods to support the development of new continuous pharmaceutical processes in the presence of limited experimental data. The objective is providing a framework to improve product/process understanding in the development of a continuous manufacturing process, which represents an essential step toward the definition of the process design space and process control strategy (Chapter 1, Section 1.2.3). Furthermore, it provides a science-based and quantitative tool to perform a robust quality risk assessment, which allows to integrate the modeling perspective with the personnel experience on the process. The intent is not to propose any new multivariate statistical method to analyze an available set of data. Rather, the aim is to show that, with a systematized use of existing methods, even limited data collected under non designed conditions can be turned into knowledge, improving the understanding of the process under development. This may provide a contribution to attenuate the reluctance that traditionally accompanies the introduction of innovative methods in pharmaceutical production processes.

The proposed procedure is applied to an industrial case study concerning the development of a continuous process for the manufacturing of paracetamol tablets. The study aims to provide an answer to such questions as: Which are the main driving forces acting on the single unit operations and on the whole manufacturing process? Can they be ranked in order of importance? Does the origin of raw materials affect the quality of the final product? Which are the main critical-to-quality variables? How are the resulting product properties related to the process settings? The answers are the primary material to enable conducting a thorough risk assessment, defining complementary experimentation and laying down the foundations for the definition of a robust control strategy, possibly inclusive of a design space.

3.2 A general procedure to use latent variable models in the development of continuous processes

Despite being identified as black-box models, LV model structures are transparent and straightforward to interpret, and can be easily understood from a mechanistic point of view. The model parameters (i.e. the loadings) highlight the variables that most contribute to explain the systematic variability in the data and order them according to their importance. Thus, interpreting the parameters of the model from a first-principles perspective can be useful to identify the main driving forces acting on the system and to enable a deep understanding of the process and of the factors that affect the operation (García-Muñoz and Settell, 2009). Note, however, that some *a priori* engineering knowledge of the system is always required in order to obtain physically sound conclusions from an LV modeling exercise.

Since LV techniques are specifically designed to analyze data wherein significant correlation exists also when a limited amount of experimental samples is available, they appear to be particularly suitable to support product and process development in pharmaceutical companies, which traditionally are not "data-rich" organizations. Additionally, when data are collected from different sources, due to the presence of different processing units on the same manufacturing line, a procedure to fuse them is highly desirable, and LV models can be useful tools to face this issue.

In pharmaceutical development, experiments are usually carried out to study the influence of different parameters (e.g., raw materials, process parameters) on the process and on the product quality. These experiments are sometimes not designed in a systematic way (e.g., through design-of-experiments techniques; Montgomery, 2005a), because an extended experimental campaign may be infeasible for economic reasons and limited time horizons, especially when high number of factors have to be tested (e.g. input materials in a product formulation). In other cases, the experiments may be incomplete or not all the interesting variables may be measured, thus causing the presence of missing data in the available experimental databases. A systematic procedure to support the development of continuous processes in the pharmaceutical industry based on LV models can be thought as a sequence of three main activities (Figure 3.1):

1.dataset organization;

2.exploratory data analysis;

3.comprehensive data analysis.

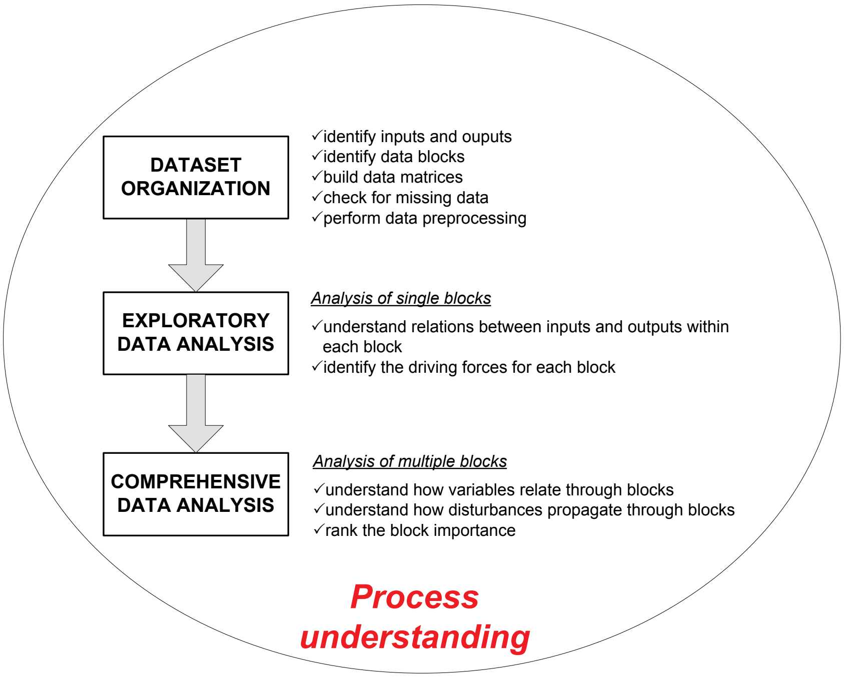

DATASET ORGANIZATION
✓identify inputs and ouputs
✓identify data blocks
✓build data matrices
✓check for missing data
✓perform data preprocessing
EXPLORATORY DATA ANALYSIS
Analysis of single blocks
✓understand relations between inputs and outputs within each block
✓identify the driving forces for each block
COMPREHENSIVE DATA ANALYSIS
Analysis of multiple blocks
✓understand how variables relate through blocks
✓understand how disturbances propagate through blocks
✓rank the block importance
Process understanding

Figure 3.1. *Schematic of the general procedure to support the implementation of a QbD approach in the development of pharmaceutical continuous processes through LV models.*

The rationale behind the proposed procedure is that each processing unit (also called "block" in the following) defining the continuous manufacturing process, and possibly each stream connecting the individual blocks, should be analyzed individually first. To this purpose, the process is first decomposed into blocks and related connecting streams. Then, knowledge is extracted from block/stream data in the form of similarities between available input materials, correlations between variables measured within a block, similarities between manufactured samples, and the like. After sufficient knowledge has been gained on each individual unit, the overall process is reassembled by merging and analyzing the available data altogether. This makes it easier to understand how the variability due to the input materials propagates along the manufacturing line and combines with the selected operating conditions. This in turn allows to understand if and how the selected operating conditions can compensate the input materials variability, and what their impact on the final product is. The information so obtained can be used to possibly define complementary experimentation to gather further understanding that can be combined with prior knowledge as part of the quality risk assessment to define a control strategy with (or without) the development of a formal design space.

3.2.1 Dataset organization

In the dataset organization step, the main operations are the identification of the processing steps, the reorganization of the available data in matrices, and data preprocessing. The main idea behind this step is to arrange the available data in a way that matches the process flowsheet as closely as possible.

After identifying the input variables (e.g. raw material properties, manipulated process variables) and the output variables (e.g. intermediate and final product properties, measured process variables), the available process measurements should be organized in different matrices (i.e., blocks), according to the unit operation they refer to. Likewise, data concerning the properties of the tested input materials and of the intermediate or final products should be organized in different matrices as well. It is preferable to divide input materials according to their specific type (e.g. API and excipient data should be collected in different matrices, as they may undergo different testing). This division should consider also the chemical and physical differences among the used materials.

Finally, all the data preprocessing activities that may be needed prior to performing the subsequent statistical analyses should be performed. These may involve pre-treatment or filtering actions (e.g., scaling and mean centering materials or process data, smoothing spectral data; Eriksson *et al.*, 2006), as well as the selection of a proper algorithm to deal with missing data (Walczak and Massart, 2001a and 2001b; López-Negrete de la Fuente *et al.*, 2010).

While dataset organization may be quite time demanding (particularly the first time around for a given continuous process/product), it is crucial for the entire data analysis procedure. Poor data organization will make the modeling exercise more difficult and, most importantly, can make the interpretation of results cumbersome.

3.2.2 Exploratory data analysis

The second step of the procedure presented in Figure 3.1 involves the exploratory analysis of the data matrices created in the first step. Even if exploratory analysis may include several different types of analysis, in this study the analysis of each single data block is considered. This analysis is intended to identify the most important variables describing the systematic variability in the data for each unit operation, the main correlations among them, and the similarities between samples processed in different experimental runs. PCA can be used as a valid modeling methodology to this end. In the interpretation of the results, it is fundamental to highlight the distinction between process variables that can be modified/manipulated (e.g., some process operating conditions, or some raw material properties such as the particle size distribution, PSD) and those that can be only measured but not manipulated arbitrarily. This distinction is vital particularly if, as a part of the control strategy definition, a design space of the process is pursued, as this should be determined in terms of manipulated variables only (these data will be referred to as *inputs*). From this point of view, it is important to underline that multivariate statistical techniques can only highlight correlations (and not causality) among variables. However, the knowledge of correlations is functional to understanding cause-effect relations that are useful for the definition of a control strategy or of a design space.

Performing an exploratory analysis on the data collected from each single unit operation helps to identify the driving forces acting on each process step. This can be achieved by understanding how the inputs act on the operation of the unit and by identifying the most important variables that should be monitored during the operation to check whether or not the process is under control. Therefore, the results from this interpretation exercise can be feasibly used to support quality risk assessment activities in the identification of the critical process parameters (CPPs) and possibly of the critical-to-quality attributes (CQAs) of the raw materials and of the product (Chapter 1, Section 1.2.1).

3.2.3 Comprehensive data analysis

The third step of the proposed procedure concerns a comprehensive analysis of the available data. This analysis differentiates from the one carried out in the second step because it is thought as a multiblock analysis. Namely, the aim is to study how variables in different blocks relate and interact, in order to analyze how downstream units, or intermediate and final product properties, are affected by different raw materials properties or different settings in the upstream units. The results from this type of analysis allow to identify the most critical blocks in the manufacturing line, as well as the most critical parameters/variables within each block. This kind of results can be obtained by performing multiblock PCA or by relating the different block datasets through regression models such as multiblock PLS (Westerhuis *et al.*, 1998). Interpreting the parameters of multiblock models helps to find out correlations among variables of different blocks, which can then be interpreted from first principles, once the distinction between input and output (i.e. regressor/response) block variables is defined. Therefore, the comprehensive analysis may be particularly useful for risk assessment, in order to define the most critical unit operations for the intermediate or final product properties. It also allows to build a single model for the whole manufacturing line, thus giving the chance to have a valid starting point for the definition of both the design space and the control strategy for the whole process, rather than defining the design spaces for each unit operation.

3.3 Case study and available data

The proposed procedure was applied to support an industrial project concerning the development of a continuous line for the manufacturing of paracetamol tablets, in which a continuous granulator is used to mix the raw materials and enlarge their particle size for subsequent tableting. Note that, for ease of presentation, how the arrangement of the available dataset was carried out (Section 3.2.1.) is reported in this section, although it could be considered a *result* of the proposed data analysis procedure (Figure 3.1).

3.3.1 The tabletting process

Figure 3.2 shows a block flow diagram of the continuous process under investigation. Four main operating steps are included:

1. granulation, carried out in a 16 mm Thermo-Fisher continuous twin screw granulator. The inlet material is fed to the granulator through a K-Tron-Soder T20 with core and coarse spiral screws. The powder mixtures are fed to the granulator via a gravity-drop feed funnel, while the granulating liquid (purified water) is added using a Jasco twin piston pump;
2. drying, performed in an Aeromatic Strea -1. The granules are dried to a water content lower than 2 wt% (where wt% = water mass/wet granule mass × 100) and an outlet temperature of 35°C. The inlet temperature is in all cases 60°C;
3. milling, performed in a Quadro CoMill 197 with a 0.55"R screen with round beater arm;
4. compaction, carried out in a compaction simulator.

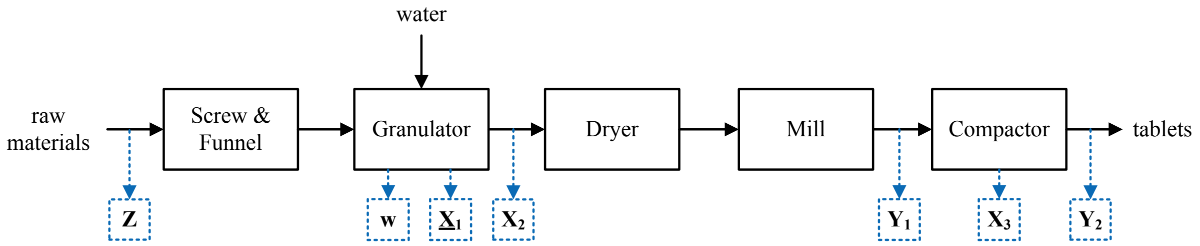


Figure 3.2. *Block flow diagram of the paracetamol tablet manufacturing line in which a continuous granulator is employed. The matrices in which the available data have been organized are indicated within dotted squares (the symbols used to denote the matrices are summarized in Table 3.1 and explained in subsections 3.3.2 to 3.3.6).*

The available data were obtained from non-designed experiments performed at an early stage of process development, by processing input materials with different characteristics under a set of different process operating conditions. Some of the process operating parameters were varied as well during the experiments to study the interaction between the raw material properties and the process operating conditions, and to gain understanding on how to operate the process in order to compensate for possible variability in the raw material properties.

Measurements were taken on the input materials, the process, the granules out of the granulation step and the manufactured tablets. To better clarify the nature of the available data, in Figure 3.2 the matrices in which the available data were arranged following the indications of Section 3.2.1 have been appended to the operations or streams which data refer to. Table 3.1 provides a compact summary of the variables included into each matrix of Figure 3.2. Additional information on the available data, their organization and the considered variables is reported in the subsections to follow.

Table 3.1. *Summary of the matrices in which the available data have been organized with the relevant included variables.*

Matrix name	Dimension	Variables included
$\mathbf{Z}$	[5×11]	Input materials characteristics
$\mathbf{w}$	[13×1]	Granulation water content
$\underline{\mathbf{X}}_1$	[12×9×245]	Granulator online measurements
$\mathbf{X}_2$	[13×4]	Particle size distribution of granules out of the granulator
$\mathbf{Y}_1$	[13×8]	Properties of granules out of the mill
$\mathbf{X}_3$	[201×2]	Compactor operating parameters
$\mathbf{Y}_2$	[201×19]	Compaction process measurements and tablet properties

3.3.2 Input materials characteristics

Data for five different input materials (M1 to M5) were available. The differences between the materials were due to the active pharmaceutical ingredient (API) particle size reduction routes, the techniques used for API isolation/drying, and the point in which formulation (API blending with excipients) occurred. Note that, whatever the input combination, the overall input formulation to the wet granulator was the same in all experiments. Two alternatives were considered for each input material pre-processing:

- wet-milling vs. microfluidization, for the API particle size reduction routes (this alternative will be described by a variable named *Size reduction route* in the following);
- agitated filter dryer vs. centrifuge/conical dryer, for the API isolation mean (described by variable *Isolation mean*);
- formulating at the point of isolation as opposed to adding the excipients post isolation, for the formulation point (described by variable *Formulation*). Namely, a material will be indicated as "+ excipients" if it was formulated at the isolation point, while it will be indicated as "API alone" if it was formulated post isolation. The latter occurred by blending the API with the excipients in a 15 litre Pharmatec IBC bin rotated at 17.5 rpm for 15 min.

Being these three variables categorical, they were included in the datasets as binary variables (0; 1) to distinguish among the two possible alternatives available for each of them. Table 3.2 shows the resulting categorization for all available input materials. The adopted categorization provides a key to the interpretation of results in the subsequent analysis. For example, a *Size reduction route* resulting "high" means a wet-milled API, whereas a *Formulation* resulting "high" stands for a material formulated at the isolation point ("+ excipient").

Table 3.2. *Adopted categorization for the available API materials.*

Material	**Size reduction route**
	wet-milled = 1
	microfluidized = 0
		Isolation mean
		agitated filter drier =1
		centrifugal conical drier = 0
			Formulation
			+ excipient = 1
			API alone = 0

M1	0	1	1
M2	1	0	1
M3	0	0	1
M4	1	0	0
M5	0	0	0

Each of the available five different input materials was further characterized by measuring the bulk density (to assess the material flow properties) and the particle size distribution (PSD). For each case, 100 cc cylinders were used with a VANKEL bulk density apparatus. Densities, both aerated (*ρ aerated raw*) and tapped (*ρ tapped raw*), were measured and reported together with the material Hausner ratio (*Hausner ratio raw*). The PSD measurements were obtained using a Sympatec (HELOS/GRADIS set up). The 10^{th}, 50^{th} and 90^{th} percentiles (*x10 raw*, *x50 raw*, *x90 raw*) of the distribution were reported for the analysis together with the distribution span (*span raw*).

Overall, ten variables were therefore available to characterize the five available input materials. As shown schematically in Figure 3.2, these variables (together with one additional variable that will be discussed in the next subsection) were collected in matrix **Z** [5×11]. Note that not all of the measurements had actually been carried out for all materials. Accordingly, missing data are present in the **Z** dataset.

3.3.3 Granulator parameters and online measurements

All the granulation experiments were carried out by keeping the following machine parameters constant: powder feed rate (2 kg/h), screw speed (200 rpm), screw set-up, barrel set-up, and barrel temperature setpoint (20 °C). To study the effect of the granulating conditions on the granules and on the final products (tablets), the water feed rate was varied between three levels, in order to manufacture wet granules containing 15, 17.5 and 20 wt% of water. A summary of the granulator experimental runs for the different input materials is reported in Table 3.3, where the values of the water content and of the feed factor at charge are reported for each lot of material processed (the five input materials resulted in 13 processed lots). Note that, due to insufficient feedstock, not all the three levels of water were tested for all input materials.

The feed factor at charge in Table 3.3 represents the capacity of the granulator at 100% screw speed and is related to the input material density. Since the powder feed rate was kept constant in all the experiments, the differences in the feed factor are mainly due to the differences in the input material density. For this reason, in this study the feed factor at charge (named *feed factor*) was considered as a condition of the input material (rather than a granulator condition), hence included in matrix **Z**. On the contrary, since only the water amount was varied across all experiments while keeping all the other granulation parameters constant, the water content values reported in Table 3.3 were collected in a separate vector **w** (see Figure 3.2).

Table 3.3. *Water content and feed factor at charge for each granulated lot.*

Lot no.	Water content [wt%]	Feed factor at charge [kg/h]
1	15.0	13.38
2	17.5	13.38
3	15.0	16.03
4	17.5	16.03
5	20.0	16.03
6	15.0	12.8
7	17.5	12.8
8	20.0	12.8
9	15.0	14.12
10	17.5	14.12
11	17.5	14.12
12	15.0	13.89
13	17.5	13.89

During each experiment, several variables were measured online on the granulator with a 1 s sampling interval:

- motor torque, measured from motor drive (indicated in the following as *TorqueNM*);
- percentage of maximum torque, measured from motor drive (*TorquePV*);
- motor torque, measured from the transducer (*TorqueTransd*);
- temperatures in different zones of the granulator (T_7, T_8, T_9 and T_{10});
- motor speed (*SpdAV*);
- feedrate of the powder feeder (*FeederRate*).

The steady state granulation data were collected in the three-way array $\underline{\mathbf{X}}_1$ $[12\times 9\times 245]$ (Figure 3.2). This array includes the trajectories of the 9 above-mentioned variables measured online for 12 lots (Lot 10 and Lot 11 of Table 3.3 are actually the same from the granulation point of view, but they differ in the milling operation). The third dimension of the matrix $[245]$ is the time length representing steady-state conditions, and corresponds to the shortest length registered for the steady states operations among all lots processed.

3.3.4 Granulator output data

For each experimental run, some samples of granules obtained from the wet granulation process were sized (on-line) during the process using an in-house particle imaging measurement system (PIMS). The PSD of each sample was characterized in terms of 10th, 50th and 90th percentile and distribution span (*L10 PIMS*, *L50 PIMS*, *L90 PIMS* and *span PIMS*).

Since several samples were collected for each experiment, the mean value of each variable across the collected samples was included in the corresponding $\mathbf{X}_2$ [13×4] matrix (Figure 3.2). Note that, since not all the processed lots were characterized through the PIMS, some data are missing in $\mathbf{X}_2$.

3.3.5 Mill output data

Granules out of the granulation step were dried and milled. All experiments had been carried out with the same mill settings. After milling, the output materials were characterized in the same way as the input materials. Therefore, aerated and tapped densities (*ρ aerated*, *ρ tapped*) and Hausner ratio (*Hausner ratio*) were measured to determine the flow properties of the output material.

These variables were included in matrix $\mathbf{Y}_1$ $[13\times 8]$, together with the measurements of the 10th, 50th and 90th percentiles and the span of the PSD obtained with Sympatec (*x10*, *x50*, *x90*, *span*). Note that, although one of the granulated lots (lot no.11 in Table 3.3) was not milled, it was nevertheless characterized as the lots that underwent milling. For this reason, in matrix $\mathbf{Y}_1$ thirteen lots are considered.

3.3.6 Compaction data

Five milled lots (lots 2, 4, 7, 10 and 13), all processed at the intermediate water level (17.5 wt%), were compacted in tablets to give the final product. For each lot, ~40 tablets were manufactured by varying the minimum punch tip separation distance and the fill depth. A total of 201 tablets were analyzed and minimum punch tip separation distance and the fill depth measurements were collected in matrix $\mathbf{X}_3$ [201×2].

Table 3.4. *Variables assigned, measured and calculated for the compaction step (matrices $\mathbf{X}_3$ and $\mathbf{Y}_2$ in Figure 3.2).*

Variable #	Variable name	Description	Type
1	minimum punch tip separation distance	operating parameter	assigned
-	duration of profile	operating parameter	assigned
2	Fill depth	operating parameter	assigned
3	Tablet weight	tablet property	measured
4	tablet thickness	tablet property	measured
5	tablet diameter	tablet property	measured
6	Hardness	tablet property	measured
7	Density	tablet property	calculated
8	Relative density	tablet property	calculated
9	Porosity	tablet property	calculated
10	Total energy		calculated
11	Recovered energy		calculated
12	Irrecoverable energy		calculated
13	Plasticity ratio	tablet property	calculated
14	Max Upper Punch Force	compactor variable	measured
15	Max Lower Punch Force	compactor variable	measured
16	Max Ejection Force	compactor variable	measured
17	Max Upper Punch Stress	compactor variable	calculated
18	Max Lower Punch Stress	compactor variable	calculated
19	Max Ejection Stress	compactor variable	calculated
20	Stress Transmission Ratio	compactor variable	calculated
21	Tensile strength	tablet property	calculated

Tablets were characterized by measuring some physical and mechanical properties, and by measuring or calculating some of the compactor variables. These data have been included in the $\mathbf{Y}_2$ [201×19] matrix of compaction process responses. Since not all the variables had been measured for all the tablets, some missing data are present in the dataset. The list of the compression step variables is reported in Table 3.4 together with their description and the distinction in assigned, measured or calculated.

3.4 Results and discussion

3.4.1 Dataset organization

The first step of the proposed procedure (Figure 3.1) involves the dataset organization. These operations have been reported in Section 3.3 for the process under investigation. Additionally, note that all matrix data have been mean-centered and scaled to unit variance prior to perform the analysis. In the case of the multiblock analysis (Section 3.4.3), each block has also been preprocessed as suggested by Westerhuis *et al.* (1998).

3.4.2 Exploratory data analysis

The exploratory data analysis was carried out on each single block reported in Figure 3.2, in order to identify the main driving forces acting *within* each unit operation and to find possible similarities among samples obtained under different experimental conditions. In the following subsections, the objectives and the results of performing an exploratory analysis on the available datasets are reported and discussed.

3.4.2.1 Analysis of input materials data

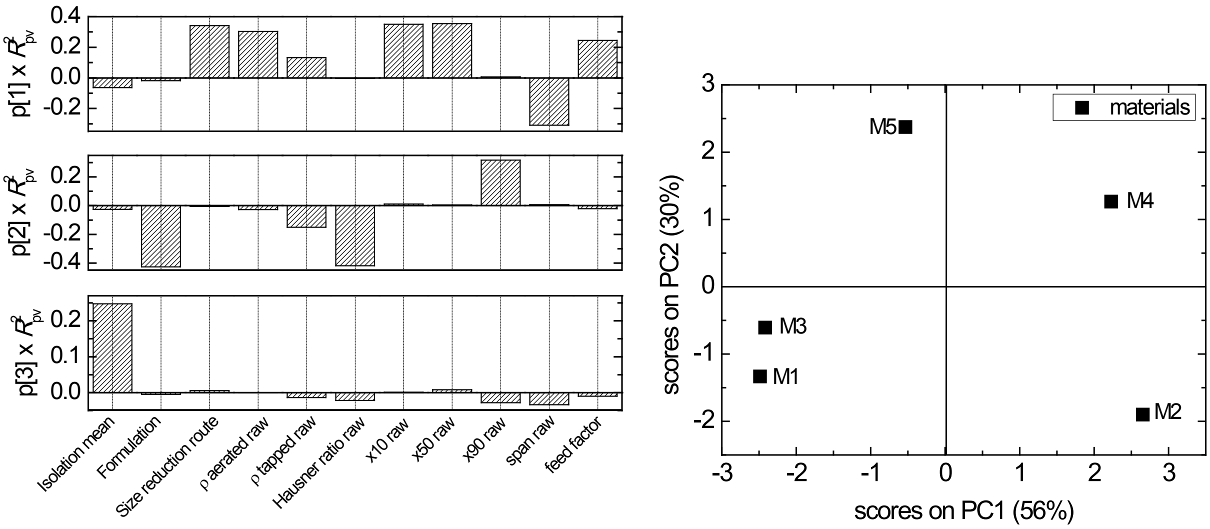
A PCA model was used to analyze the data in matrix **Z**. Table 3.5 reports a summary of the model diagnostics, namely the eigenvalues, the explained variance for PC (R^2) and the cumulative explained variance per PC (R^2_{CUM}). The number of PCs used to build the model has been determined with the "eigenvalue-greater-than-one" rule (Mardia *et al.*, 1979). It can be seen that, although only the first two PCs show an eigenvalue greater than 1, also the eigenvalue corresponding to the third PC can be feasibly rounded up to 1. For this reason, the PCA model on **Z** was built on 3 PCs. Since some data in **Z** were missing, the analysis was carried out using the NIPALS algorithm (see Appendix B), which is known to be robust in calculating the model when the percentage of missing data in the dataset is not high.

Table 3.5. *Diagnostics of the PCA model on the input material matrix* **Z**.

PC	**Eigenvalues**	R^2	R^2_{CUM}
1	6.07	56.02	56.02
2	3.24	29.99	86.01
3	0.85	7.87	93.88
4	0.58	5.61	99.50
5	3e-3	0.36	99.85

In Figure 3.3a the loadings of the PCA model are reported as bar plots, whereas in Figure 3.3b the diagram of the scores on the first 2 PCs is plotted. Note that the loadings in Figure 3.3a have been weighted according to the variance explained per variable by each PC of the model

(R^2_{pv}). Thus, the loadings corresponding to those original variables that are better described by the model have a larger weight, which improves the interpretability of the model (García-Muñoz and Settell, 2009). This weighting operation will be repeated in all the loading diagrams presented in this study. Details on the interpretation of the loadings and scores plots are reported in.



(a) (b)

Figure 3.3. *(a) Loading bar plots of the PCA model on matrix* **Z**. *(b) Score plot on the first 2 PCs of the PCA model on matrix* **Z**.

From the analysis of the top plot of Figure 3.3a, it can be seen that the first PC (accounting for ~56% of the total variability of the data in **Z**) mainly describes differences due to the *Size reduction route* (wet-milled vs. microfluidized) among the processed materials; these differences are accompanied by differences in the PSD and density (especially aerated, *ρ aerated raw*) of the materials. This can be understood by noticing that the bar corresponding to the size reduction route variable is on the same side of the plot (positive correlation), and with similar magnitude, compared to the bars of *x10 raw*, *x50 raw* and *ρ aerated raw*, whereas the bar corresponding to *span raw* is on the opposite side (i.e. negative correlation). Therefore, an API size-reduced by wet-milling (high *Size reduction route*) is characterized by larger particles on average (larger *x10 raw* and *x50 raw*) if compared to the microfluidized one, where the particle distribution is narrower (lower *span raw*). In general, it can be concluded that wet milled input materials have also larger aerated densities (*ρ aerated raw*), which result in higher *feed factor* to the granulator.

The middle plot of Figure 3.3a indicates that the second source of variability (~30%) for the input materials data is mainly related to the point in which the API is formulated (*Formulation*). Materials that have been formulated post isolation ("API alone") are characterized by lower Hausner ratio (and lower *ρ tapped raw*), whereas the corresponding PSDs have larger tails compared to the "+ excipients" materials, being the variable indicating the 90th percentile (*x90 raw*) of the distribution inversely related to the variable indicating the formulation point.

The bottom plot of Figure 3.3a indicates that the third PC mainly describes the different API isolation mean. However, the *Isolation mean* seems rather unrelated with all the other variables included in **Z**, meaning that it has a lower impact on the variability of the input material characteristics compared to the other API pre-treatments. Furthermore, it accounts for only ~8% of the total variability in input materials.

The score diagram of Figure 3.3b reflects the loading structure of the model and indicates the similarities between the available input materials (Wold *et al.*, 2001). From the considerations driven by the top plot of Figure 3.3a, one would expect that PC1 separates wet-milled materials (i.e. materials M2 and M4; see Table 3.2) from microfluidized ones (i.e. M1, M3 and M5). In fact, in Figure 3.3b M2 and M4 project on the right of the diagram, and are separated from M1, M3 and M5, which are located on the left of the diagram. Therefore, the separation indeed occurs along PC1. On the other hand, "+ excipients" materials (M1, M3 and M2; diagram bottom) can be distinguished from "API alone" materials along PC2 (top of the diagram), as anticipated by the middle plot of Figure 3.3a. Note that materials M1 and M3, which are both microfluidized and "+ excipients" (see Table 3.2), indeed project very close to each other in Figure 3.3b.

These results demonstrate that LV modeling provides a very useful method to assess the acceptability of new materials (Duchesne and MacGregor, 2004). According to the model, a new material can be accepted for granulation as long as it falls inside the range of the historically accepted materials, even if it had been subject to different API pre-treatments (for example, if neither wet-milling nor microfluidization were used to isolate API). This procedure could be integrated as part of the risk assessment on input materials and could also be considered as a preliminary step toward the definition of a model representing their design space.

3.4.2.2 Analysis of granulator process data

The data collected online from the granulator and included in matrix $\underline{\mathbf{X}}_1$ were analyzed through PCA with the following aims:

- understanding the relations between the variables monitored during the granulation process;
- understanding if and how the differences in the input materials affect the granulation;
- understanding the role of the water amount and its effect on the granulation process;
- finding potential similarities between the different lots that had been tested.

To allow the analysis with PCA, the three-way array $\underline{\mathbf{X}}_1$ was transformed into a two-way matrix $\mathbf{X}_1$ $[(12 \cdot 245) \times 9]$ through a variable-wise unfolding operation (Nomikos and MacGregor, 1994), which in this case is the most appropriate, as the interest is in the analysis of the steady state-part of the granulation for each lot.

The diagnostics of the PCA model built on $\mathbf{X}_1$ are reported in Table 3.6. It results that 5 PCs are enough to build the PCA model (which captures ~93% of the total variability of the $\mathbf{X}_1$ data), with the first two PCs explaining ~62% of the total variability.

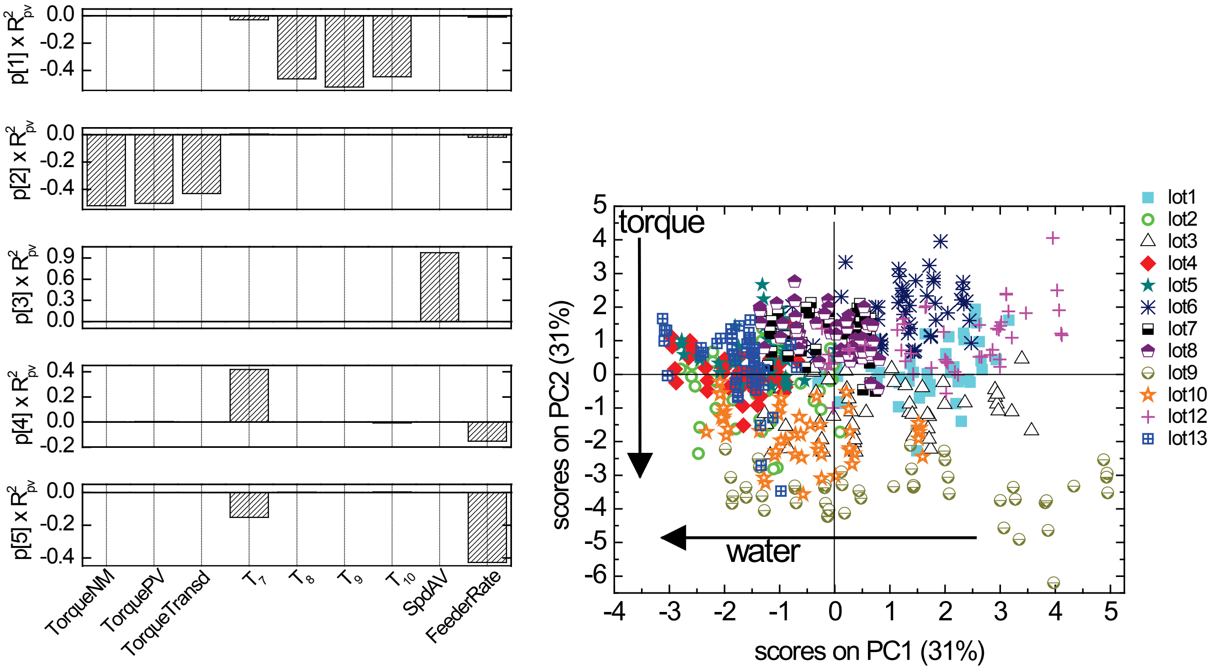
Table 3.6. *Diagnostics of the PCA model on the granulation online data in* $\mathbf{X}_1$.

PC	Eigenvalues	R^2	R^2_{CUM}
1	2.79	31.03	31.03
2	2.77	30.74	61.77
3	1.00	11.14	72.91
4	0.92	10.20	83.12
5	0.89	9.84	92.96
6	0.28	3.07	96.03

Figure 3.4a reports the loadings on the five PCs considered. It can be clearly seen that the granulation process is driven by two main factors of similar importance. The first one (top plot), represented by PC1 (which describes ~31% of the variability of the data), is represented by the temperatures measured along the granulator (T_8, T_9 and T_{10}, which are all correlated on PC1), except the granulator inlet temperature (T_7; this can be explained by the fact that this temperature sensor is located very close to the granulator feed point, and therefore this temperature is much more related to the feed temperature than to the granulation process).

The second important factor (second plot) is concerned with the motor torque measurements, which are all correlated on PC2 and unrelated to the temperature measurements (the temperature measurements bars have negligible widths in the second plot). Also PC2 explains ~31% of the variability of the data, meaning that the two phenomena have a similar importance in the process. This information is useful both from a process understanding and from a quality risk assessment point of view, as it indicates that these variables convey two independent driving forces that can both have a significant impact on the process, and should be monitored to keep the granulator operation under control. Although three additional PCs are significant to describe the variability in the granulator data, their importance is much lower if compared to PC1 and PC2 (~10% of explained variance each).

A combined analysis of the loading plots with the score diagram of Figure 3.4b can give some insights for a deeper physical interpretation of these results. Figure 3.4b shows that most of the samples corresponding to the same lot (i.e., having the same symbol and color in the figure) are located in the same region of the score space (with few exceptions). In general, lots are separated along PC1 because of the differences in the temperatures measured during the granulation, as highlighted by the top plot of Figure 3.4a. These temperature differences can be associated to the different amount of water employed during the granulation. In fact, consider (for example) lots 1, 3, 6, 9 and 12: they all fall mainly on the region of the score diagram with positive PC1; an analysis of the relevant datasets showed that all these lots were processed with a low amount of water (15 wt%). On the contrary, lots whose samples have negative PC1 were usually processed with medium/high amount of water (17.5/20 wt%). It can be concluded that the amount of water used is positively correlated with the temperatures measured online on the granulator.



(a) (b)

Figure 3.4. *(a) Loading bar plots of the PCA model on matrix* $\mathbf{X}_1$. *(b) Score plot on the first 2 PCs of the PCA model on matrix* $\mathbf{X}_1$ *(the different processed lots are indicated by different symbols/colors).*

Comparing Figure 3.4b with the information drawn from the second loading plot of Figure 3.4a, it can be noticed that samples corresponding to lots with higher motor torque mostly project onto the region of the score space with negative PC2 (for example lots 3, 9, 10). Analysis of the relevant databases showed that these lots were all wet-milled, hence characterized by narrow PSD with high mean, and higher density. These material characteristics probably increase the stresses on the granulator screw, leading to higher torque measurements. On the contrary, microfluidized materials (such as lots 1, 6, 7, 8, 12, 13) are characterized by lower torque values (scores with positive PC2), probably because of their smaller PSD.

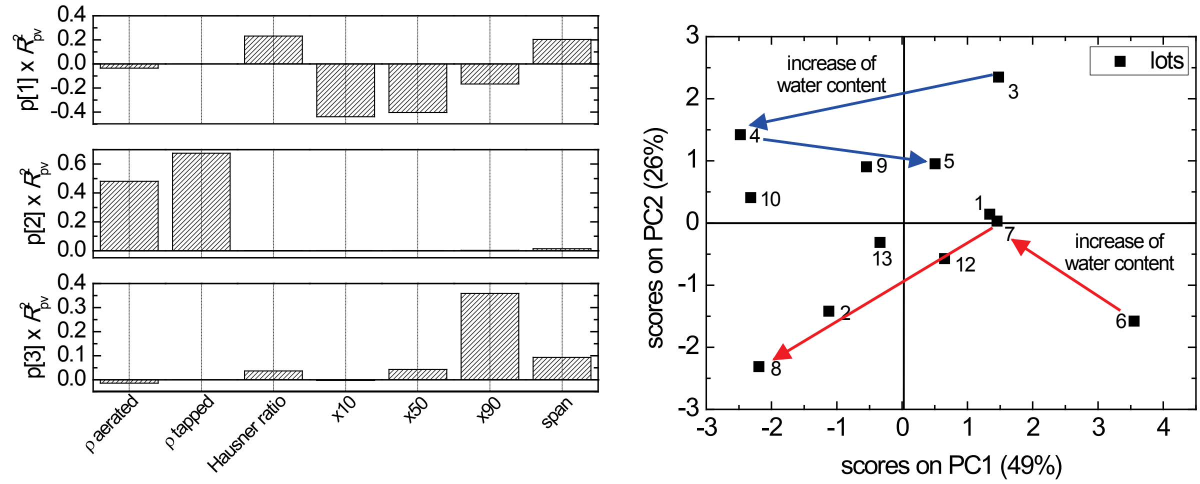
This analysis confirms that the characteristics of the input materials do affect the granulation process variables, and different materials can be feasibly distinguished also based on the score and loading plots.

3.4.2.3 Analysis of mill output data

A PCA model was designed on the matrix of the measured properties of the milled granules ($\mathbf{Y}_1$). The aim of this analysis was to evaluate if the differences in the input materials can be recognized also after milling or if the granulation and milling operations can “filter out” the raw materials differences. If this were the case, the process might be resilient (robust) enough to inherently compensate for input materials variations, hence the design space would enlarge (a very desirable occurrence). The PCA model on $\mathbf{Y}_1$ was built with 3 PCs (total variance captured: ~89%), as can be seen from the model diagnostics in Table 3.7.

Table 3.7. *Diagnostics of the PCA model on the mill output data in* $\mathbf{Y}_1$.

PC	Eigenvalues	R^2	R^2_{CUM}
1	3.40	48.58	48.58
2	1.79	25.62	74.21
3	1.03	14.74	88.95
4	0.61	8.72	97.67
5	0.15	2.09	99.76



(a) (b)

Figure 3.5. *(a) Loading bar plots of the PCA model on matrix* $\mathbf{Y}_1$*. (b) Score plot on the first 2 PCs of the PCA model on matrix* $\mathbf{Y}_1$*. The arrows show the path followed by two different materials processed at increasing levels of granulator water.*

The analysis of the loading plots (Figure 3.5a) shows that most part (~49%) of the variability in $\mathbf{Y}_1$ is due to the differences in the PSD of the milled granules. These differences in turn explain the differences between lots along PC1 on the score plot (Figure 3.5b), where lots resulting in larger granules (e.g., lots 4, 10, 2 and 8) are projected on the left. Since these lots had been granulated with intermediate to high amounts of water (17.5 and 20 wt%), it can be concluded that the granule size increases as the granulator water content increases (as expected). This is confirmed by the analysis of lots 6, 7, 8, which all refer to the same input material (M3): the processed material moves from the far right to the far left of the score plot as the water content is increased (red arrows in Figure 3.5b).

It should however be mentioned that the distinction between the effects of the intermediate and high water level is not entirely clear for all lots (see for example lots 3, 4, and 5, all coming from input material M2 and granulated at increasing water levels; blue arrows in

Figure 3.5b). This may be due to the combination of different processing operations prior to milling, which can partly mask the relationships between water and PSD. Also note that M2 is a wet-milled material, whereas M3 is microfluidized. Therefore, it seems that wet-milled materials are less sensitive to the effect of the highest water amount used for the granulation. However, this conclusion should be checked by further experimentation.

It also appears that the "main footprint" of the source material (i.e., whether microfluidized of wet-milled) is still visible after milling. In fact, note that lots 3, 4, 5, 9 and 10 project onto the region of the score plot with positive PC2, which means that they all result in granules with densities larger than the average. However, all these lots come from wet-milled materials. Therefore, even if the granulation step (water content) can change the location of a lot in the score plot, resulting in similar final PSD for different starting input materials, the "memory" of the milled granule origin is not entirely lost even after milling. This type of information could be utilized as part of the quality risk assessment on the product to demonstrate that an additional control to the input material is necessary rather than controlling the process only.

3.4.2.4 Analysis of compaction data

Compaction data included in matrix $\mathbf{X}_3$ (operating data) and $\mathbf{Y}_2$ (tablet properties) were analyzed in order to understand how the compactor parameters affect the tablet properties and if the differences in the input materials are still visible in the final product. A PCA model was therefore built on the data of matrix $[\mathbf{X}_3\ \mathbf{Y}_2]$, obtained by concatenating matrix $\mathbf{X}_3$ and matrix $\mathbf{Y}_2$ (Figure 3.2). This matrix presents a significant amount of missing data (~ 22%).

Table 3.8 reports the model diagnostics, which indicates that 3 PCs are enough to describe the systematic variability in the data (~91%). It can be seen that the first PC accounts for a very large fraction (~73%) of the total data variability. Since the value of the first eigenvalue is 12.26 (i.e., PC1 represents ~12 original variables), there are a lot of variables in $[\mathbf{X}_3\ \mathbf{Y}_2]$ that are correlated and possibly redundant.

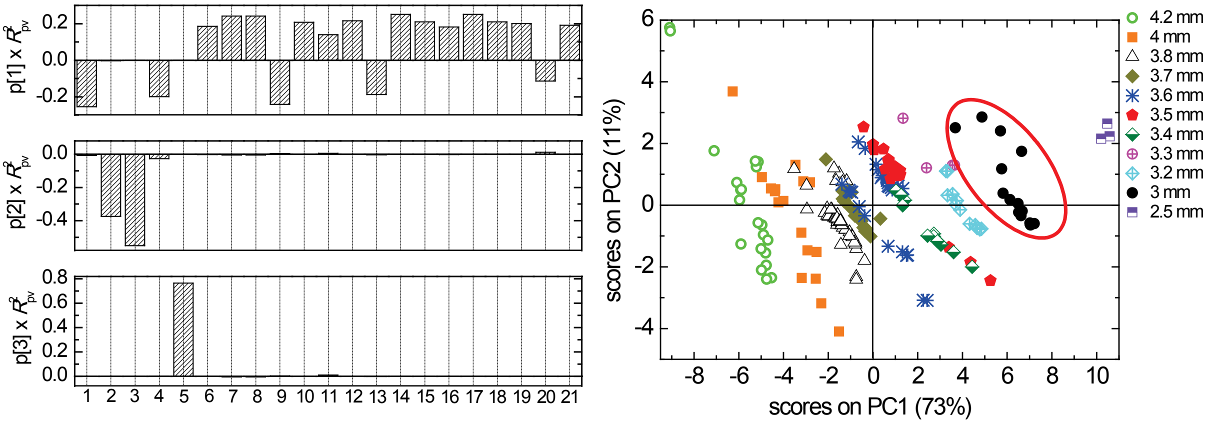
Table 3.8. *Diagnostics of the PCA model on the compression data in* $[\mathbf{X}_3\ \mathbf{Y}_2]$.

PC	Eigenvalues	R^2	R^2_{CUM}
1	12.26	72.96	72.96
2	1.90	11.34	84.29
3	1.13	6.50	90.80
4	0.67	3.85	94.65
5	0.42	2.47	97.13

In Figure 3.6a the loading plots are shown[†]. It results that the correlation structure of tablet data is driven by the compactor operating parameters (i.e. variables [1] and [2], corresponding to minimum punch tip separation distance and fill depth, respectively), with the separation distance dominating (first PC). Furthermore, most of the measured tablet properties data appear strongly correlated, as expected.

The separation distance appears to be:

- correlated to tablet thickness [4], porosity [9], plasticity ratio [13], and to the stress transmission ratio [20] associated with the compactor. This means that higher punch tip distances give higher values of these properties, on average;
- inversely correlated (i.e. anti-correlated) to hardness [6], density [7], relative density [8] and tensile strength [21], and to most of the variables measured on the compactor ([10], [11], [12] and from [14]-[19]). From a practical point of view this means that operating with higher punch tip distances is expected to give tablets that are thicker, more porous and plastic, but at the same time less dense, hard and tensile;
- uncorrelated to (not affecting) tablet weight ([3]) and tablet diameter ([5]), which is obviously expected.



(a) (b)

Figure 3.6. *(a) Loading bar plots of the PCA model on matrix* $\mathbf{Y}_2$*. (b) Score plot on the first 2 PCs of the PCA model on matrix* $\mathbf{Y}_2$*. Symbols refer to different punch tip separation distances. The ellipse indicates the subset of samples for which a dedicated PCA model was also built.*

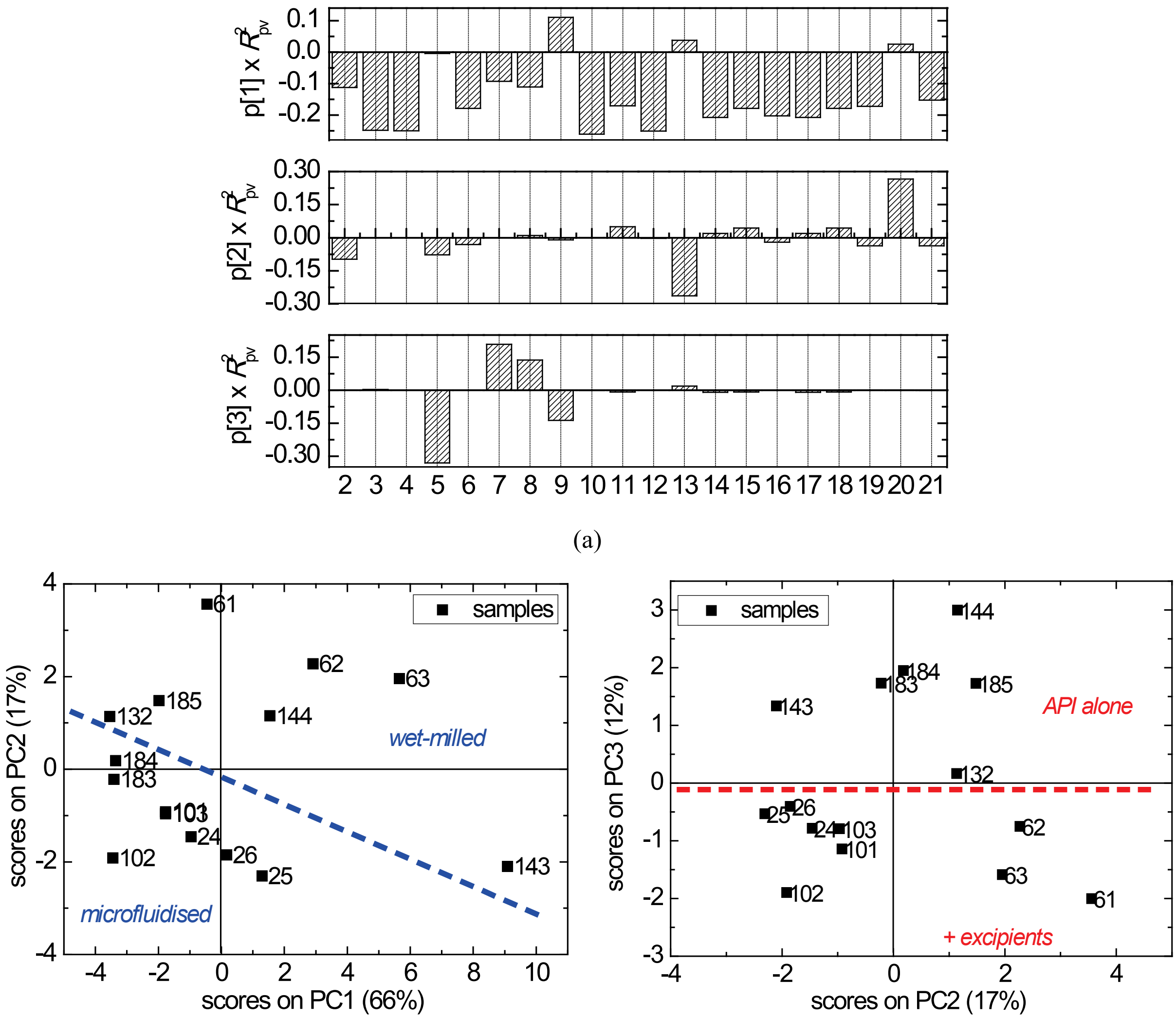
The middle plot of Figure 3.6a shows that the other significant source of variability ([2], fill depth) strongly correlates with tablet weight (higher fill depths give heavier tablets, as expected), but seems not to be correlated with any other tablet property. Tablet diameter [5] is not correlated to any tablet or machine property (Figure 3.6a, bottom plot). The score plot

[†] To improve readability, Figure 3.6a reports numbers instead of the actual names of the variables as bar labels. Table 3.4 can be used to match the reported numbers with the corresponding variables. In the main text, each variable is indicated by the corresponding number *n* within square brackets [*n*].

(Figure 3.6b) shows that PC1 (minimum punch tip separation distance) distinguishes among tablets obtained in different experiments, with samples obtained at larger distances projected far left.

In order to study whether different input materials leave any footprint on the properties of the manufactured tablets, a new PCA model was built using a subset $\mathbf{Y}_2^{\mathrm{SUB}}$ of matrix $\mathbf{Y}_2$ including all samples of different materials that had been processed at the same value of minimum punch tip distance (3 mm), but with different fill depths (black dots within the ellipse in Figure 3.6b).

The results in terms of loading and score diagrams are reported in Figure 3.7 for the 3 PCs used to build the model, which account for ~95% of the total variability in the data. The numbers in the plots indicate the original index of the selected samples in matrix $\mathbf{Y}_2$.



(a)

(b)

(c)

Figure 3.7. *(a) Loading bar plots of the PCA model on matrix* $\mathbf{Y}_2^{\mathrm{SUB}}$. *(a) Plot of the scores on PC1 versus PC2 and (b) of PC2 versus PC3 of the PCA model on matrix* $\mathbf{Y}_2^{\mathrm{SUB}}$. *The numbers in the score plots indicate the original index of the selected samples in matrix* $\mathbf{Y}_2$. *The dashed lines separate the final tablet properties according to types of raw materials they originate from.*

The loadings bar plots in Figure 3.7a better disclose the effect of the fill depth [2] on the measured variables (which was observed in the second plot of Figure 3.6b), showing that the correlations between most of the variables are maintained.

Figure 3.7b shows that samples originating from microfluidized or wet-milled materials project onto different sections of the score plot (with one exception: sample 185 is microfluidized, but projects onto the region of wet-milled materials). By analyzing this score diagram with the corresponding loading bar plots (top and middle plot of Figure 3.7a), it could be concluded that microfluidized materials give harder ([6]) and stronger (more tensile [21]) tablets than wet-milled materials, which reflects the fact that they were processed with higher fill depths ([2]) on average. In the score plot of PC2 versus PC3 (Figure 3.7c), the different lots are separated with respect to the point at which the excipients were added to the API, and this separation occurs along PC3. The analysis of the loadings (bottom plot of Figure 3.7a) suggested that tablets from "API alone" materials have a smaller diameter ([5]) and to be denser ([7], [8]) and less porous ([9]) than the "+ excipients" ones.

Despite the limitations in the dataset due to missing data and the lack of data related to other experimental conditions, the analysis could therefore clarify that input materials coming from different PSD reduction routes and formulated at different points prior to granulation do result in tablets with different properties upon compaction. On the other hand, the API isolation mean does not seem to impact on the final product properties.

3.4.3 Comprehensive data analysis

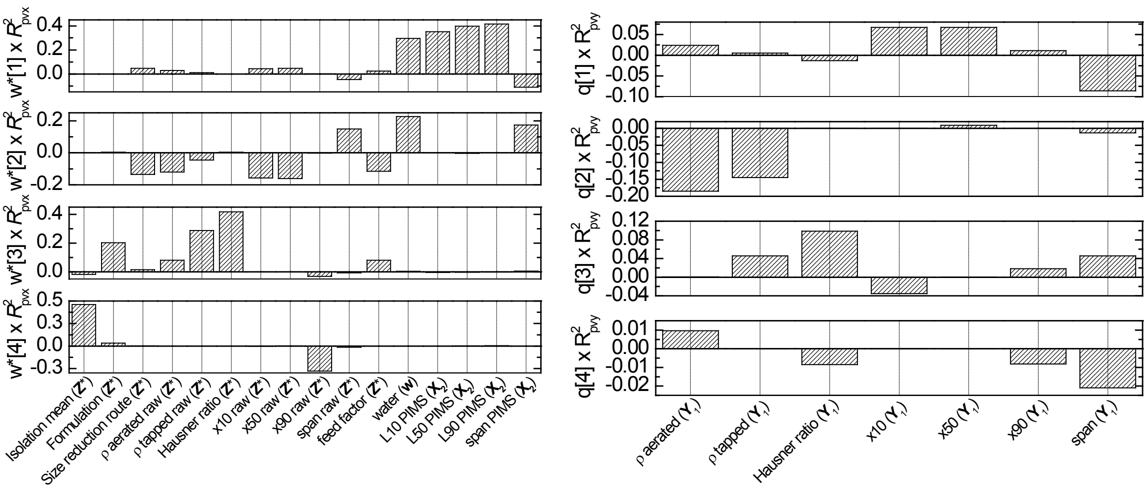
The exploratory analysis on the single datasets (blocks) highlighted correlations among variables *within* each processing unit. The aim of the comprehensive data analysis is instead that of studying the relations between variables pertaining to different blocks and between the blocks themselves. This can be useful for several purposes: understanding how variables relate *through* blocks, understanding which assignable variables matter more in determining the final product properties, understanding how disturbances in materials inputs or to a block propagate to the final product. This can be useful to develop a control strategy for the process, in order to ensure the proper response to possible disturbances entering the system.

Although the final tablet properties that one is most interested at are included in $\mathbf{Y}_2$, the related tablet measurements were available for some lots only. For this reason, a multi-block PLS (MB-PLS) model was built to predict the milled granule properties ($\mathbf{Y}_1$) instead of the tablet properties. Only those variables that could be modified in the experiments (namely, input materials properties and granulator water level) were used as regressors. These were arranged in matrix $\left[\mathbf{Z}^* \; \mathbf{w} \; \mathbf{X}_2\right]$, where $\mathbf{Z}^*$ $[12 \times 10]$ was generated by repeating the rows of matrix $\mathbf{Z}$ (Figure 3.2) for each lot that had been manufactured from the same input material. Note that the granulator variables that had been monitored online (matrix $\underline{\mathbf{X}}_1$) were not used, because (as shown earlier) they can be related to the granulator operating parameters and the input material properties.

Table 3.9 reports the diagnostics of the MB-PLS model in terms of explained variance per LV both for the regressor matrix ($R^2\mathbf{X}$, indicating as **X** the regressor matrix) and for the response matrix ($R^2\mathbf{Y}$, indicating as **Y** the response matrix). The corresponding cumulative values are reported as well ($R^2_{\mathrm{CUM}}\mathbf{X}$ and $R^2_{\mathrm{CUM}}\mathbf{Y}$). Furthermore, the variances explained by the model in cross-validation (Q^2 and Q^2_{CUM}) are reported. It is shown that after the fourth LV the variance explained by the model is no longer significant, both in model building and in cross-validation. For this reason, four LVs were used to build the MB-PLS model.

Table 3.9. *Diagnostics of the MB-PLS model between matrix* $\left[\mathbf{Z}^* \ \mathbf{w} \ \mathbf{X}_2\right]$ *and matrix* $\mathbf{Y}_1$.

LV	$R^2\mathbf{X}$	$R^2_{\mathrm{CUM}}\mathbf{X}$	$R^2\mathbf{Y}$	$R^2_{\mathrm{CUM}}\mathbf{Y}$	Q^2	Q^2_{CUM}
1	49.79	49.79	24.99	24.99	-13.56	-13.56
2	29.88	79.68	17.93	42.93	41.14	27.57
3	10.83	90.51	11.93	54.86	18.48	46.06
4	5.37	95.88	2.87	57.73	4.63	50.69
5	1.56	97.44	5.56	63.29	0.30	50.99
6	1.94	99.38	2.69	65.98	7.53	58.51



(a) (b)

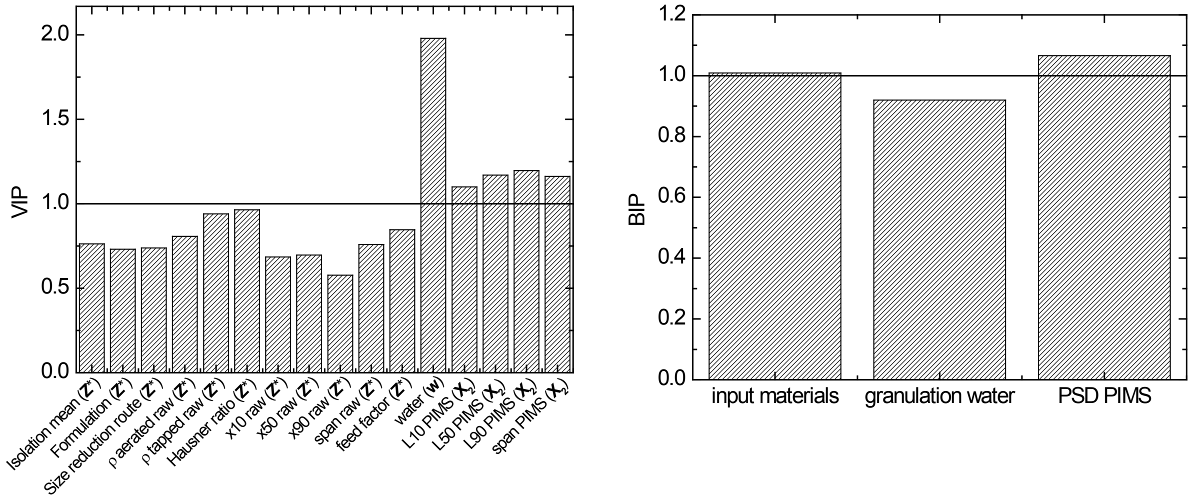
Figure 3.8. *(a) Bar plots of the weights* $\mathbf{W}^*$ *of the MB-PLS model between matrix* $\left[\mathbf{Z}^* \ \mathbf{w} \ \mathbf{X}_2\right]$ *and matrix* $\mathbf{Y}_1$. *(b) Bar plots of the loadings* **Q** *of the MB-PLS model between matrix* $\left[\mathbf{Z}^* \ \mathbf{w} \ \mathbf{X}_2\right]$ *and matrix* $\mathbf{Y}_1$.

To understand the intra-block and inter-block relations, the loadings of the MB-PLS model can be analyzed. In particular, the bar plots of the weights $\mathbf{W}^*$ of matrix $\left[\mathbf{Z}^* \ \mathbf{w} \ \mathbf{X}_2\right]$, weighted on the variance explained by each LV per regressor variable, are reported in Figure 3.8a, whereas the bar plots of the loadings **Q** of the response matrix $\mathbf{Y}_1$, weighted on the variance explained by each LV per response variable, are reported in Figure 3.8b.

The first LV of the model mainly describes the correlation among the water levels and the variables in $\mathbf{X}_2$, i.e. the PSD of the granules out of the granulator. These appear to be positively related (only) with the variables of the PSD of the granules out of the mill, which is an expected occurrence: higher water levels give larger and narrower PSDs out of the granulator, which in turn gives larger and narrower PSDs out of the mill (being the mill settings constant). LV2 shows that less dense granules out of the mill are obtained with higher water levels and with microfluidized materials. LV3 is mainly affected by the formulation point: "+ excipients" materials (which appear to be denser) result also in denser granules out of the mill.

The added value of using an MB-PLS model is that it can be used to *predict* the responses from the input data. The prediction will be mainly affected by the variables that have a stronger influence on the responses. As shown in Chapter 2, (Section 2.1.2.1), the importance of the variable *n* in the projection can be measured through the VIP index (Eq.(2.33)). Similarly, to quantify the importance of each block in the projection, the BIP index can be calculated (Eq.(2.38)).

Figure 3.9 reports the VIP and BIP indices for the variables and blocks involved in the MB-PLS model. Recall that a threshold equal to 1 is usually applied to decide whether a variable or a block is important or not in the prediction of the response variables. In this case, several variables have a VIP-index next to the threshold.



(a) (b)

Figure 3.9. *(a) VIP and (b) BIP indexes of the variables and of the blocks respectively used to build the MB-PLS model between matrix* $[\mathbf{Z}^* \ \mathbf{w} \ \mathbf{X}_2]$ *and matrix* $\mathbf{Y}_1$.

Figure 3.9a shows that the most important variables for the prediction of the milled granule properties are the water level and the variables describing the PSD out of the granulator. However, the correlation between the variables within $\mathbf{X}_2$, and between these variables and the PSD of the milled granules, strongly affects the model (as expected since the mill settings are constant). Figure 3.9b indicates that although the PSD block is the most important in determining the milled granule properties, the contribution of the input materials properties is also significant.

In general, these results further confirm the driving forces that were identified in the exploratory data analysis, and give a quantitative measure on which are the blocks with the most significant contribution in explaining the variability of the process data, providing a valid support to guide risk assessment.

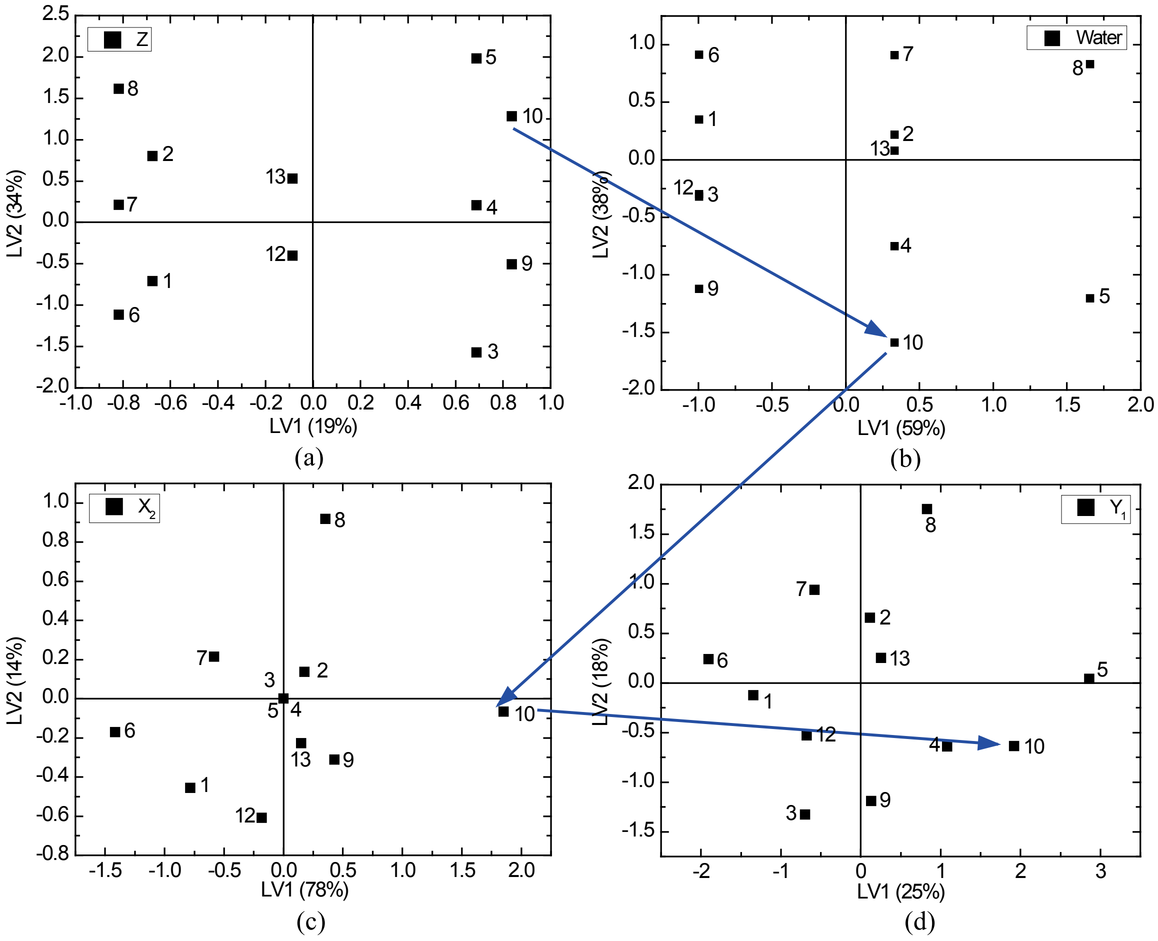


Figure 3.10. *(a) Plot of the scores on LV1 versus LV2 for the matrix Z* in the MB-PLS. (b) Plot of the scores on LV1 versus LV2 for the water levels w in the MB-PLS. (c) Plot of the scores on LV1 versus LV2 for the matrix X_2 in the MB-PLS. (d) Plot of the scores on LV1 versus LV2 for the matrix Y_1 in the MB-PLS.*

Finally, Figure 3.10 shows a very useful information that can be obtained by using an MB-PLS model. As already noted in Chapter 2 (Section 2.1.3.1), since the MB-PLS builds a model for each block involved in the manufacturing line (MacGregor *et al.*, 1994), a score diagram can be identified for each block. Figure 3.10 shows the score diagrams on the first 2 LVs of each block of the model. These score diagrams, which indeed reflect the correlation structure highlighted by the loadings of Figure 3.8, provide an useful tool to assess the performance of the entire manufacturing process, by identifying "paths" along which the material being processed moves along the manufacturing line. These paths identify the directions followed by the process if a particular material and a particular water level is used in the manufacturing. Furthermore, they especially give information on which the expected characteristics for the milled granules are.

Let us take lot 10 as an example. It originates from a “wet-milled” and “API-alone” input material, and can be identified in Figure 3.10a (score plot of matrix $\mathbf{Z}^{*}$). After processing in the granulator with a water level of 17.5 wt%, its projection onto the water level score plot moves to the point indicated by the arrow in Figure 3.10b. The resulting granules are characterized by a large PSD on average, as can be seen by the score diagram of Figure 3.10c, where the projection of lot 10 is within the region of large PSDs out of the granulator (recall that according to the loadings in Figure 3.8a, high scores on LV1 mean larger PSD). Finally, the resulting granules out of the mill are characterized by a large PSD and high density values compared to the other materials, and these properties are projected onto the corresponding region of the score plot in Figure 3.10d. It is interesting to note that in Figure 3.10d lot 10 is close to the projections of lot 4 and lot 5. The common feature between these three lots is that they all originate from a wet-milled API, reinforcing the finding that the particle size reduction route is the most important variable in explaining the variability among the different lots. This representation of the process on the score diagrams as in Figure 3.10 can provide an useful tool for the development of both the design space and the control strategy for the whole manufacturing line.

3.5 Conclusions

In this Chapter, a general strategy to apply multivariate statistical techniques to support the development of continuous processes has been presented. In particular, LVMs have been shown to be very useful tools to extract information from development datasets, which are sometimes sub-optimally structured and sparse, in reflection of the fact that knowledge is generated in a cyclical and incremental manner, which in turns leads to the availability of heterogeneous datasets. Moreover, it has been shown that these techniques may be effective in supporting the design of continuous manufacturing lines, in which data from different unit operations are collected and need to be analyzed jointly.

The proposed strategy aimed at formalizing the application of LV modeling in order to get a systematic support tool to gain process understanding from the available development data. The procedure is based on three main steps. The first step deals with data management, where the data are organized in distinct matrices corresponding to the units (or blocks) of the process. In the second step, an exploratory analysis is carried out on the data of each single block, in order to identify the driving forces acting on each unit operation and to find redundancies and correlation between the measured variables. In this step, the focus is to understand how the design variables act on the process, and if they can potentially have an impact on the subsequent operations and on the intermediate and final product properties. Finally in the third step, a comprehensive analysis on all the available data is performed, with the aim of discovering/confirming relations between the variables of different blocks and the impact of each block on the downstream blocks and on the final product properties.

The proposed framework has been successfully applied to an industrial case study concerning the development of a continuous paracetamol tablet manufacturing line. The aim was to understand how different input material preprocessing and formulation routes, and how different process settings (granulator water levels, compactor settings) impacted on the downstream processing and on the final product properties. Using the proposed procedure, it was shown how the parameters of the LVMs can be interpreted from first principles, allowing to identify the main driving forces acting on the system and to rank them according to their importance. In particular, it was found that the route chosen to reduce the size of the API particle prior to granulation, the point at which the API is formulated and the amount of water used in the granulation steps were the three most important driving forces acting on the process. The different API particle size reduction route (wet-milling or microfluidization) and the point in which the API was formulated (at the isolation point or post isolation) could also be distinguished by analyzing the intermediate (milled granules) and the final (tablets) product property data, meaning that the process parameters (granulation water and compactor settings) can reduce only partially the differences due to the input materials, at least within the domain of available experimental data.

Furthermore, it was shown that multi-block modeling tools can help in identifying which are the most critical units in the process and the most critical variables within them. These tools have also been demonstrated to be useful in identifying paths along which the whole continuous multi-unit process can move, depending on the selected process settings.

The outlined procedure can be used in the earlier stage of a product development framework to help define the input materials/process settings to explore in the next experimentation cycle. In a later stage, it can be used to integrate additional prior knowledge as part of a more thorough quality risk assessment to provide the rationale for defining a robust control strategy. Finally, if sufficiently large and pertinently structured datasets exist on the finalized process, the same procedure could be used to integrate (where feasible) the control strategy with a latent variables-based design space on individual or combined continuous unit operations.

Chapter 4

Latent variable model inversion to support the design of new products and processes[*]

In this Chapter, latent variable regression models (LVRMs) are proposed as tools to assist the design of new products and processes through model inversion. A general framework for LVRM inversion is presented in which different scenarios to invert the model are identified. It is shown that the design problem may have infinite solutions, generating the so-called *null space*, which is demonstrated to share many common features with the *design space* defined by the regulatory Agencies. The proposed framework is tested on an industrial particle engineering problem involving high-shear wet granulation. A discussion is provided on the effect of uncertainties in the reconstruction of the design solution. Finally, strategies are described to exploit the historical data covariance structure in the selection of new desirable product properties most suitable for LVRM inversion.

The Chapter is organized as follows. In the first section, a thorough review of the applications of LVRM inversion is provided. The second section presents the framework, and the different LVRM inversion scenarios are discussed. In the third section, scenarios are tested on the above-mentioned particle engineering problem, considering three different case studies. In the fourth section, details are provided on the use of the historical data to reconstruct new product target profiles. Finally, conclusions and further issues are discussed.

4.1 Introduction

Model-based product and process design requires a mathematical abstraction that represents the complex network of interactions between input materials, processing conditions and

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

Tomba, E., S. García-Muñoz, P. Facco, F. Bezzo and M. Barolo (2012). A general framework for latent variable model inversion to support product and process design. *Computer Aided Chemical Engineering 30*, (I.D.L. Bogle and M. Fairweather, Eds.), Elsevier, Amsterdam (The Netherlands), p.512-516.

Tomba, E., P. Facco, F. Bezzo and S. García-Muñoz (2012). Exploiting historical databases to design the target quality profile for a new product. Submitted to *Ind. Eng. Chem. Res.*.

desired product properties. As stated in Chapter 1, a deterministic model to describe these interactions is always desirable, as it describes the behavior of a system from first principles, giving a transparent representation of the physical phenomena acting upon the system. Several examples which consider the interactions between processes and raw materials for the prediction of the final product quality have appeared in the literature (Hatzantonis *et al.*, 1998; ter Horst *et al.*, 2006). These models typically require a large amount of resources to be developed and may not be a viable solution if there is a lack of detailed understanding to build a model that explains product performance metrics. For this reason, empirical models based on data are often built.

In development activities, a large amount of experiments are typically required to independently excite all the driving forces guiding the relations between input material properties, process parameters and product performances, due for example to the large number of candidate materials and permutations to consider for the formulation of a product (e.g. in the pharmaceutical industry). If data on historically developed products or processes were available, useful information could be drawn from these databases, to support the design and the optimization of new products or processes and/ or to guide the experimentation from the first steps of the development. This would accelerate the development cycle and avoid expensive exploratory experimentation, replacing this with targeted optimal experiments.

Historical data analysis for product/process design was first reported by Moteki and Arai (1986), who used PCA and theoretical models to analyze historical data from a LDPE process and infer process conditions for new grades of products. Other authors proposed the use of expert system tools (like fuzzy logic or artificial neural networks) to model the relations between input variables and product properties and use them to estimate the inputs corresponding to new product characteristics (Borosy, 1999; Sebzalli and Wang, 2001). However, although these methodologies can provide a prediction of the response of interest, they lack of transparency in relating large amount of data and are not easily understandable.

As seen in Chapter 2, LVRMs are tools specifically designed to analyze large datasets of highly correlated data, reducing the vast information included in them in few meaningful LVs, which identify the underlying driving forces relating input data with system outputs. The driving forces identified from the available historical data (represented by the LVs) can therefore be used to support new product and process development activities. From a LVRM point of view, the design of a new product can be seen as the estimation of the best model inputs $\mathbf{x}^{\mathrm{NEW}}$ $[N \times 1]$, where N is the number of considered input variables, which correspond to the desired model outputs $\mathbf{y}^{\mathrm{DES}}$ $[M \times 1]$, where M is the number of considered product quality properties. The values in $\mathbf{x}^{\mathrm{NEW}}$ can therefore be obtained through an operation of model inversion (Chapter 2, Section 2.2).

4.1.1 Latent variable model inversion background

The use of LVRMs to support process design was introduced by Jaeckle and MacGregor (1998, 2000a). In their studies, the authors showed how the inversion of LVRMs built on the available historical development data could be used to estimate a window of conditions at which a process should operate in order to yield a product with desired quality characteristics. In particular, they proposed a framework for the inversion of empirical models in which the new process conditions $\mathbf{x}^{\mathrm{NEW}}$ were estimated from the desired product quality $\mathbf{y}^{\mathrm{DES}}$, through a projection matrix $\mathbf{M}^{\mathrm{T}}$, which depended on the modeling technique used. The solution obtained through this projection matrix represents therefore the analytical inversion of the model.

In the above studies, it was acknowledged that the solution of the LVRM inversion could not be unique due to possible differences in the matrix ranks (as described in Chapter 2, Section 2.2). The multiple solutions arising from the inversion form a subspace of the model space called *null space* (Chapter 2, Section 2.2.1). The selection of the best solution among the multiple ones belonging to the null space required the introduction of appropriate constraints, which the analytical model inversion described in the works of Jaeckle and MacGregor did not allow. For these reasons, in order to obtain a solution physically sound and in the range of the historical data used to build the model, an optimization problem with the appropriate constraints had to be solved. Depending on the case under study, different formulations of the optimization problem have been proposed. A summary of the references to studies in which LVRM inversion has been used is reported in Table 4.1.

Lakshminarayanan *et al.* (2000) proposed to use hard constraints (HC) on the distance of the solution from the origin of the model space (represented by the Hotelling's T^2 statistic) and on the model mismatch in representing the solution (represented by the SPE for $\mathbf{x}^{\mathrm{NEW}}$), to invert an LVRM in which the relation between **X** and **Y** was modeled with genetic programming to account for possible nonlinearities. The objective was to minimize the difference between the desired product properties and those predicted by the model, by forcing the solution T^2 and SPE to lie inside the confidence limits calculated from the data used to build the model. Hwang *et al.* (2004) used a multi-block PLS model inversion to determine the optimal environmental factors to ensure a desired level of cellular function in the development of tissue-engineered devices. In their model inversion procedure they introduced a cost function into the objective function to find the optimal solution, in order to identify the most cost effective combination of environmental factors respecting the correlation structure given by the model.

García-Muñoz *et al.* (2006) extensively investigated the concept of null space and proposed an optimization framework to invert PLS models, with the aim of estimating batch operating policies (the time varying profiles for the manipulated variables) to reach a desired output product quality.

Table 4.1. *Summary of the references to studies which used LVRM inversion. References are listed in a chronological order.*

Reference	Application	Main contribution
Jaeckle and MacGregor (1998)	Process design for new LDPE grades	Direct LVRM inversion/Null space concept
Jaeckle and MacGregor (2000a)	Industrial batch polymerization process condition design	Direct LVRM inversion/Null space concept
Lakshminarayanan *et al.* (2000)	Design of a rubber compound formulation	PCA-based regression model inversion. SC on solution scores and HC on the solution T^2 and SPE.
Hwang *et al.* (2004)	Environmental factors for desired levels of cellular function in tissue engineering.	Multiblock PLS model inversion through cost-based optimization.
Yacoub and MacGregor (2004)	Design and optimization of the operating conditions of an industrial over-molding injection process.	Nonlinear PLS model inversion. HC on the solution T^2, SPE and on $\mathbf{x}^{\mathrm{NEW}}$ elements. SC on $\mathbf{y}^{\mathrm{DES}}$ elements and on quality variables variance.
Flores-Cerillo and MacGregor (2004)	Trajectory tracking in batch polymerization processes.	Dynamic PCA model inversion for set point trajectory tracking. SC on control action. HC on scores and on process and manipulated variables.
Flores-Cerillo and MacGregor (2005)	Control of industrial batch polymerization processes.	Score space optimization for manipulated variable estimation. SC on control action, variable set points and T^2. HC on control action.
Garcia-Muñoz *et al.* (2006)	Optimization of the operating conditions of an industrial batch pulp digester.	Optimization framework for PLS model inversion in the presence of a null space. SC on solution T^2. SC and HC on $\mathbf{y}^{\mathrm{DES}}$ elements.
Muteki *et al.* (2006)	Optimal selection of materials for the development of new polymer blends.	Mixture PLS model inversion. SC on cost and number of materials in the mixture. HC on solution T^2 and SPE. Mixture and logical constraints for material selection.
Muteki and MacGregor (2007)	Sequential design of mixture experiments for new product development.	Mixture PLS model inversion. Inversion problem as in Muteki *et al.* (2006).
Muteki and MacGregor (2008)	Optimal purchasing of raw materials for product design.	Mixture PLS model inversion. Inversion problem as in Muteki *et al.* (2006).
Garcia-Muñoz *et al.* (2008)	Optimization of batch operating policies in an industrial semi-batch polymerization process.	Two steps optimization for PLS model inversion. SC on solution SPE and on batch length/material consumption.
Garcia-Muñoz (2009)	Process operating conditions scale-up.	Two steps optimization for JY-PLS inversion. SC on solution SPE. HC on $\mathbf{x}^{\mathrm{NEW}}$ elements.
Garcia-Muñoz *et al.* (2010)	Feed-forward controller for mid-course correction in a wet granulation process.	PLS model inversion for process control. SC and HC on solution SPE. SC on $\mathbf{y}^{\mathrm{DES}}$ elements. HC on $\mathbf{x}^{\mathrm{NEW}}$ elements.
Yacoub and MacGregor (2011b)	Robust modeling and optimization of an industrial membrane manufacturing process.	Nonlinear PLS model inversion for robust process development. HC on the T^2, SPE and $\mathbf{x}^{\mathrm{NEW}}$ elements. SC on $\mathbf{y}^{\mathrm{DES}}$ elements and on sensitivities of product quality to disturbances.
Yacoub *et al.* (2011a)	Robust modeling and optimization of a tablet manufacturing line.	Nonlinear multi-block PLS model inversion for robust process development. Inversion problem as in Yacoub and MacGregor (2011).
Liu *et al.* (2011b)	Scale-up of a pharmaceutical roller compaction process.	JY-PLS model inversion. SC and HC on solution T^2 and SPE. SC on $\mathbf{y}^{\mathrm{DES}}$ elements. HC on $\mathbf{x}^{\mathrm{NEW}}$ elements.
Muteki *et al.* (2011)	De-risking scale-up of high–shear wet granulation.	PLS model inversion. SC and HC on $\mathbf{y}^{\mathrm{DES}}$ elements. HC on T^2, SPE and $\mathbf{x}^{\mathrm{NEW}}$ elements.
Liu *et al.* (2011a)	Modeling and optimization of a tablet manufacturing line.	Multiblock PLS model inversion. SC and HC on solution T^2 and SPE. SC on $\mathbf{y}^{\mathrm{DES}}$ elements. HC on $\mathbf{x}^{\mathrm{NEW}}$ elements.

The proposed solution strategy could deal with both equality and inequality constraints for the variables in $\mathbf{y}^{\mathrm{DES}}$, and considered a soft rather than a hard constraint on the Hotelling's T^2, in order to avoid to anchor the solution at the given value of the hard constraint. On the contrary, no constraints or conditions were considered for the regressor space. In a later study, García-Muñoz *et al.* (2008) modified the problem in order to include in the framework also possible constraints in the regressor space ($\mathbf{x}^{\mathrm{NEW}}$); the optimization problem was then split in two steps: in the first step the model was inverted to find the optimal latent space projections for the required product quality $\mathbf{y}^{\mathrm{DES}}$; in the second step, the optimal set of regressor variables $\mathbf{x}^{\mathrm{NEW}}$ was found, by matching the projections of $\mathbf{y}^{\mathrm{DES}}$ calculated in the first step with the projections of $\mathbf{x}^{\mathrm{NEW}}$ in the latent space. In order to find a solution of the inversion problem satisfying the possible constraints in the regressor space (i.e. on $\mathbf{x}^{\mathrm{NEW}}$ variables), the optimization procedure might be forced to extrapolate, namely to find a solution that did not belong to the model space. This was obtained by considering a soft constraint (SC) in the optimization problem for the SPE of $\mathbf{x}^{\mathrm{NEW}}$. The same procedure in two steps was used by García-Muñoz (2009) to invert a JY-PLS model (García-Muñoz *et al.*, 2005) in order to estimate the process conditions in a plant, assuming that the same raw materials as in a reference plant were used, obtaining the same product properties $\mathbf{y}^{\mathrm{DES}}$.

Yacoub and MacGregor (2004) extended LVRM inversion to nonlinear models, with the aim of optimizing the final product quality and compensating for the uncontrolled sources of variability (e.g., raw materials and environmental factors). To minimize product variability and increase robustness, they also proposed to use a soft constraint on the variance of the product properties. The product robustness problem was further refined including in the objective function of the LVRM inversion the sensitivities of the product quality with respect to specified disturbances. The effectiveness of the proposed methods was assessed to optimize a tablet manufacturing line (2011a) and a membrane manufacturing process (2011b). In this approach it was assumed that one could measure but not control the disturbances entering the system, and the model inversion was applied to achieve the desired product quality, properly modifying the process parameter settings.

LVRM inversion was also proposed in process control for trajectory tracking (Flores-Cerrillo and MacGregor, 2005) and manipulation (Flores-Cerrillo and MacGregor, 2004), to ensure the quality of the operation in a batch process. In these studies the problem of estimating at the decision point the future control actions to implement was completely solved in the space of the LVs, thus not considering constraints on the mismatch of the model in fitting the previous measured and manipulated variables. This ensured a more conservative approach in the calculation of the control action. Garcia-Muñoz *et al.* (2010) performed a similar exercise in which they applied a feed-forward controller to perform mid-course corrections in a wet granulation process, but imposing also soft and hard constraints to the SPE of the regressors in $\mathbf{x}^{\mathrm{NEW}}$, which included fixed variables (for example the raw material properties or process variables measured until the control action decision point) and the manipulated variable values to calculate, allowing slight model extrapolations.

LVRM inversion was also used for product development to estimate new product formulations (Muteki *et al.*, 2006). In this case, the LVRM was built to relate the properties of the raw materials, weighted according to the their fraction inside the formulation, with the processing conditions and the final product properties. The objective of the inversion in the cited case study was to select the best materials, their fractions in the formulation and the process settings to ensure a product of desired properties. The model inversion optimization problem was properly modified to account for the material selection part (logical constraints) and the mixture constraints. Moreover, in the objective function, soft constraints to minimize the cost for the material and the number of materials used in the formulation were considered.

The same type of strategy was proposed to evaluate raw material purchasing for product manufacturing (Muteki and MacGregor, 2008) and to guide experiments through a sequential procedure of model inversion, result verification and model updating to accelerate the development of new products (Muteki and MacGregor, 2007). In these inversion problems hard constraints on the T^2 and SPE statistics were considered.

As described in Chapter 1, the interest in model-based product and process design is recently increased in the pharmaceutical industry in support of QbD activities. Pharmaceutical scientists are seeking to increasingly apply computational tools using models of multiple natures to design robust and reproducible products and processes. Some applications which involve LVRM inversion have therefore been proposed also in the pharmaceutical industry, as seen in Section 1.4.1.3 of this Dissertation (García-Muñoz, 2009; Liu *et al.*, 2011a; Yacoub *et al.*, 2011a; Liu *et al.*, 2011b; Muteki *et al.*, 2011).

However, although all the above-mentioned studies used LVRM inversion to solve different kinds of problems, the objective function being minimized has often been tailored to the specific case study. In the following sections, a general framework to perform LVRM inversion is proposed, which includes several possible different cases which one may encounter in a product/process design exercise. The framework provides the most appropriate objective function and sets of constraints for each specific scenario the user may encounter, given any combination of constraints in both the quality and the regressor spaces.

4.2 A general framework for latent variable model inversion

Let us consider an LVRM, such as PLS, built between a dataset $\mathbf{X}\ [I \times N]$ of I input conditions in which N variables (e.g., process parameters, raw material properties) were measured (the regressor space) and a dataset $\mathbf{Y}\ [I \times M]$ of I products for which M product properties were measured (the response/quality space). The objective of model inversion is that of using the model to estimate a set of new input conditions $\mathbf{x}^{\mathrm{NEW}}$ corresponding to a desired set of response variables $\mathbf{y}^{\mathrm{DES}}$. As it was shown in Chapter 2 (Section 2.2), depending on the effective latent dimensionality (i.e. the rank) of the matrices involved in the model (and on the number A of LVs used to build the LVRM), the model inversion problem may have infinite solutions, all lying in the model null space. Furthermore, in a design problem the target properties defining the product quality in $\mathbf{y}^{\mathrm{DES}}$ may not be completely assigned but possibly allowed to vary inside acceptance ranges (defined through inequalities). At the same time, some input variables may not be adjustable or allowed to vary only in specific ranges (e.g., the raw material properties), thus representing further constraints for the design problem. For all these reasons, to find the optimal design solution through LVRM inversion a constrained optimization problem has to be solved, whose formulation depends on the problem constraints.

4.2.1 Model inversion problem formulation

In general, LVRM inversion can be summarized by the following steps (Figure 4.1):

1.Build the LVRM between the preprocessed $\mathbf{X}$ and $\mathbf{Y}$ matrices.

2.Determine the desired product specifications $\mathbf{y}^{\mathrm{DES}}$ in terms of assigned values (*equality constraints*), one or two-sided constraints (*inequality constraints*), and physical bounds[††].

3.Determine the necessary constraints for the solution $\mathbf{x}^{\mathrm{NEW}}$ (if any), in terms of equality (assigned values), inequality constraints and physical bounds, so that the solution found is of practical relevance[†].

4.If $\mathbf{y}^{\mathrm{DES}}$ is completely specified (all equality constraints), verify that the LVRM is valid for $\mathbf{y}^{\mathrm{DES}}$ by comparing $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ with the SPE of the historical samples or the relevant historical confidence limit $\mathrm{SPE}_{\mathbf{Y},95\%\,\mathrm{lim}}$ (if meaningful). If $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} > \mathrm{SPE}_{\mathbf{Y},95\%\,\mathrm{lim}}$ it is not recommended to perform model inversion.

5.Invert LVRM solving the appropriate inversion problem.

6.Show the results in terms of estimated input conditions $\mathbf{x}^{\mathrm{NEW}}$, corresponding predicted quality $\hat{\mathbf{y}}^{\mathrm{NEW}}$, Hotelling's T^2 and squared prediction error for the solution $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$.

As can be seen from the above-mentioned steps and from Figure 4.1, in case $\mathbf{y}^{\mathrm{DES}}$ is completely defined (all equality constraints on $\mathbf{y}^{\mathrm{DES}}$) an LVRM can be inverted to estimate $\mathbf{x}^{\mathrm{NEW}}$ only if the model is valid for $\mathbf{y}^{\mathrm{DES}}$, as the relations described by a LVRM are valid only in the space defined by the LVs. One way to assess it is to project $\mathbf{y}^{\mathrm{DES}}$ onto the latent space of $\mathbf{Y}$ and verifying that the value of $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ is at least under the (say) 95% confidence limit calculated from the historical samples (García-Muñoz *et al.*, 2006). Note that this practice provides a robust indication of the closeness of $\mathbf{y}^{\mathrm{DES}}$ to the latent space depending on the

[††] Note that physical bounds represent the variable domain in the optimization procedure. Differently, inequality constraints represent the regions inside which the properties (either quality or regressor) are desired to fall, and are then subsets of the physical bounds.

nature of the historical data in **Y**. As a matter of fact, the 95% confidence limits that can be calculated from the historical samples can be meaningless in the case a limited number of **Y** samples is available (as often happens in development environments). In these situations it would be more informative to compare $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ with the values of the SPE of the available historical samples, rather than with the relevant confidence limit.

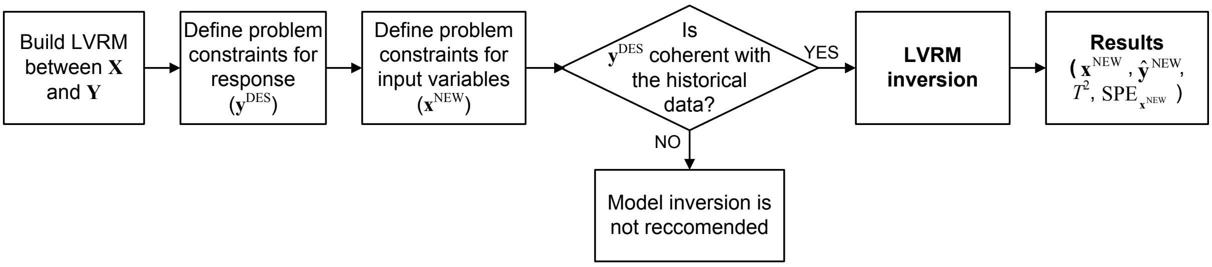

Build LVRM between $\mathbf{X}$ and $\mathbf{Y}$
Define problem constraints for response ($\mathbf{y}^{\mathrm{DES}}$)
Define problem constraints for input variables ($\mathbf{x}^{\mathrm{NEW}}$)
Is $\mathbf{y}^{\mathrm{DES}}$ coherent with the historical data?
YES
NO
LVRM inversion
Results ($\mathbf{x}^{\mathrm{NEW}}$, $\hat{\mathbf{y}}^{\mathrm{NEW}}$, T^2, $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$)
Model inversion is not reccomended

Figure 4.1. *Schematic of the LVRM inversion steps.*

Although $\mathbf{y}^{\mathrm{DES}}$ may be coherent with the historical data in **Y**, if the model mismatch in representing $\mathbf{y}^{\mathrm{DES}}$ is significantly different from zero, the uncertainties propagate in the inversion, thus increasing the uncertainties in the estimation of $\mathbf{x}^{\mathrm{NEW}}$. Managing $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ in the inversion problem would require to consider the prediction uncertainty, which is sample-dependent and formed by different contributions as the uncertainties due to the measurement system, to the lack of fit of the model and to the sample bias, all of which are not easily manageable. For these reasons, in the following, $\mathbf{y}^{\mathrm{DES}}$ is assumed to belong to the space of the quality of the historical samples in **Y** ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} = 0$). Namely $\mathbf{y}^{\mathrm{DES}}$ is projected and reconstructed on the model for **Y**, acknowledging that the handling of the uncertainties on the **Y** space in the LVRM inversion problem still constitutes an open research area. Further details on the reconstruction of $\mathbf{y}^{\mathrm{DES}}$ will be provided in Section 4.4.1.

By analyzing the steps in Figure 4.1 it is clear that, when performing model inversion, three different type of constraints for the problem can be distinguished:

- Model constraints, i.e. the underlying model has to be satisfied.
- Statistical constraints, i.e. the solution should preferable lie within the region established by the historical data used to build the model. This region could be represented by statistical limits on the Hotelling's T^2 and on $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$.
- variable constraints, namely equality constraints or inequality constraints for the product properties in $\mathbf{y}^{\mathrm{DES}}$ or the input variables in $\mathbf{x}^{\mathrm{NEW}}$ which the solution has to obey.

The LVRM defines the first type of constraints. The second type of constraints can be different depending on the problem under study and on the specified variable constraints. In general they can be soft constraints (i.e. constraints included indirectly within the objective function) or hard constraints. The third type of constraints are defined by the user depending on the problem.

As noted by García-Muñoz *et al.* (2010), the statistical constraints (namely the bounds that are imposed on the Hotelling's T^2 and on $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$), are implicitly imposing range boundaries to the variables in the estimated solution set $\mathbf{x}^{\mathrm{NEW}}$ and for the predicted quality $\hat{\mathbf{y}}^{\mathrm{NEW}}$. For this reason it is expected that the limits of the Hotelling's T^2 and of the $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$ ensure that the estimated solution variables respect the bounds of the historical sample variables. Nevertheless, it is useful to consider physical bounds of variables (when needed) in the formulation of the inversion problem, since physical bounds are linear constraints, which can be more effective in aiding the optimization routine to find a solution compared to the statistical constraints, which are nonlinear (Biegler, 2010).

In general, whether constraints on variables exist or not makes the inversion formulation problem different. In particular, in Figure 4.2 a general framework for LVRM inversion is proposed. A first classification between different model inversion problems depends on having or not constraints on the regressor vector $\mathbf{x}^{\mathrm{NEW}}$. Depending on this, different objective functions and problem constraints can be found according to whether all the values in $\mathbf{y}^{\mathrm{DES}}$ are specified or not.

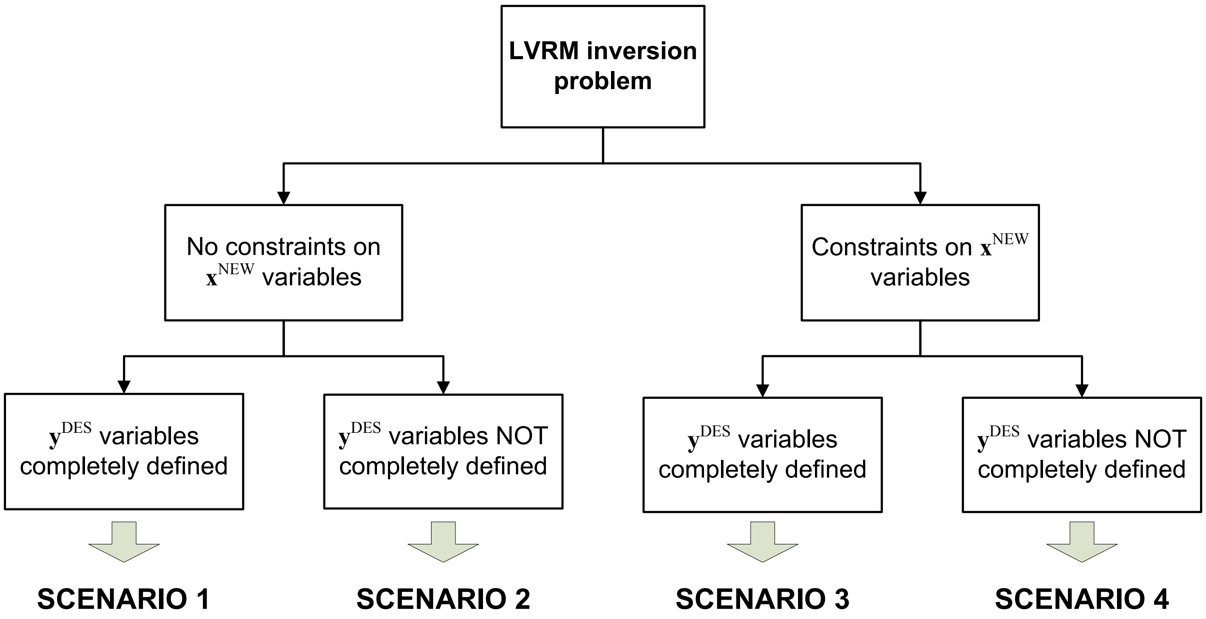

LVRM inversion problem
No constraints on $\mathbf{x}^{\mathrm{NEW}}$ variables
Constraints on $\mathbf{x}^{\mathrm{NEW}}$ variables
$\mathbf{y}^{\mathrm{DES}}$ variables completely defined
$\mathbf{y}^{\mathrm{DES}}$ variables NOT completely defined
$\mathbf{y}^{\mathrm{DES}}$ variables completely defined
$\mathbf{y}^{\mathrm{DES}}$ variables NOT completely defined
SCENARIO 1
SCENARIO 2
SCENARIO 3
SCENARIO 4

Figure 4.2. *General framework for LVRM inversion.*

4.2.1.1 Unconstrained regressors

In the case no constraints are considered for the solution $\mathbf{x}^{\mathrm{NEW}}$, and $\mathbf{y}^{\mathrm{DES}}$ is completely defined (Scenario 1 in Figure 4.2), the direct LVRM inversion described in Chapter 2 (Eq.(2.57)) can be applied. The direct inversion of the model provides the score vector $\hat{\mathbf{t}}^{\mathrm{DES}}$ $[A\times 1]$ corresponding to the desired product quality vector $\mathbf{y}^{\mathrm{DES}}$, from which the input variable vector $\hat{\mathbf{x}}^{\mathrm{NEW}}$ can be reconstructed (Jaeckle and MacGregor, 1998):

$$\hat{\mathbf{x}}^{\mathrm{NEW}} = \mathbf{P}\hat{\mathbf{t}}^{\mathrm{DES}} \quad . \tag{4.1}$$

In Eq.(4.1), $\hat{\mathbf{x}}^{\mathrm{NEW}}$ belongs to the model space and has the same covariance structure of the historical data used to build the LVRM.

In case where the elements in $\mathbf{y}^{\mathrm{DES}}$ are not completely defined because some elements lack an equality constraint, or if an inequality constraint is assigned for that element (Scenario 2 in Figure 4.2), the model inversion problem formulation is the following:

$$\begin{aligned}
&\min_{\mathbf{t}}\left(\hat{\mathbf{y}}^{\mathrm{NEW}}-\mathbf{y}^{\mathrm{DES}}\right)^{\mathrm{T}}\boldsymbol{\Gamma}\left(\hat{\mathbf{y}}^{\mathrm{NEW}}-\mathbf{y}^{\mathrm{DES}}\right)+g_1\cdot\left(\sum_{a=1}^{A}\frac{t_a^2}{s_a^2}\right)\\
&\text{subject to}\\
&\hat{\mathbf{y}}^{\mathrm{NEW}}=\mathbf{Qt} \qquad\qquad\qquad\qquad ,\\
&\hat{\mathbf{x}}^{\mathrm{NEW}}=\mathbf{Pt}\\
&\hat{y}_j^{\mathrm{NEW}}\le b_j\\
&lb_k^y\le\hat{y}_k^{\mathrm{NEW}}\le ub_k^y \qquad lb_l^x\le\hat{x}_l^{\mathrm{NEW}}\le ub_l^x
\end{aligned} \tag{4.2}$$

where $\mathbf{t}$ is the vector of the decision variables, composed by A t_a scores, s_a^2 is the variance of the a-th column of matrix $\mathbf{T}$, $\hat{\mathbf{y}}^{\mathrm{NEW}}$ is the quality variable vector corresponding to the solution $\hat{\mathbf{x}}^{\mathrm{NEW}}$, b_j is the inequality constraint specified for the j-th element of $\hat{\mathbf{y}}^{\mathrm{NEW}}$ ($\hat{y}_j^{\mathrm{NEW}}$); lb_k^y and ub_k^y are respectively the lower and upper physical bounds for the k-th element of $\hat{\mathbf{y}}^{\mathrm{NEW}}$ ($\hat{y}_k^{\mathrm{NEW}}$), while lb_l^x and ub_l^x are the lower and upper physical bounds for the l-th element of $\hat{\mathbf{x}}^{\mathrm{NEW}}$ ($\hat{x}_l^{\mathrm{NEW}}$).

$\boldsymbol{\Gamma}$ is a matrix whose diagonal elements determine how much weight is given to meet the possible specified equality constraints $\mathbf{y}^{\mathrm{DES}}$ in the solution. More weight could be given to those variables which are more important for the specific application under study. Alternatively, the fractions $R^2_{\mathrm{pv},y}$ of the sum of squares of each property explained by the model for the historical samples could be used as weights (Eq.(2.16)). A value of zero is assigned to the weights for those variables for which equality constraints are not specified (García-Muñoz *et al.*, 2006).

As can be seen, the objective function in Eq.(4.2) minimizes the sum of the weighted squared difference between the desired product properties in $\mathbf{y}^{\mathrm{DES}}$ and those predicted by the model included in $\hat{\mathbf{y}}^{\mathrm{NEW}}$ and of the Hotelling's T^2, represented by the second term of the objective function (soft constraint). The Hotelling's T^2 term is weighted according to the weight g_1 to balance the importance of the two terms in the objective function. For this reason, one could vary the weight g_1 assigning more importance to the model representativeness or to the closeness to the historical knowledge. A good choice for g_1 is represented by the reciprocal of the 95% confidence limit for T^2 ($T^2_{95\%\,\mathrm{lim}}$), in order to keep the second term of the objective function below 1 (if possible). Note that the soft constraint on T^2 is included in order to find a solution lying as close as possible to the historical available data when multiple solutions exist (i.e. in the case of inequality constraints for $\hat{\mathbf{y}}^{\mathrm{NEW}}$ or, in alternative, when a null space exists). In general, when $\mathbf{y}^{\mathrm{DES}}$ is completely defined, the analytical model inversion in Eq.(4.1) gives the best possible solution even when a null space is present, if the projections $\hat{\mathbf{t}}^{\mathrm{DES}}$ of $\mathbf{y}^{\mathrm{DES}}$ in the latent space of the model are inside the design space given by the historical data. If not, it would be preferable to use the formulation in Eq.(4.2) instead of the direct model inversion, thus exploiting the soft constraint on T^2 to move the solution along the null space. Moreover, the optimization framework has to be preferred to the direct model inversion if the calculated solution $\hat{\mathbf{x}}^{\mathrm{NEW}}$ does not respect the physical boundaries (lb_l^x and ub_l^x).

4.2.1.2 Constrained regressors

If too many constraints are specified for the input variables in $\mathbf{x}^{\mathrm{NEW}}$, the model inversion solution may be forced to move away from the model plane ($\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}} > 0$). The model inversion problem can be formulated in such a way as to take this occurrence in account, by including a soft constraint for $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$, namely the mismatch of the model in representing $\mathbf{x}^{\mathrm{NEW}}$ (García-Muñoz *et al.*, 2008). Differently from the previous scenarios, the solution will lie outside the model space, although only slightly, as long as $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$ is lower than a specified threshold (which can be represented by the historical confidence limit $\mathrm{SPE}_{\mathbf{X},95\%\,\mathrm{lim}}$). However, instead of considering a two-step optimization problem as proposed by Garcia-Muñoz *et al.* (2008), the inversion problem can be solved in a single step. The problem formulation changes depending on having the desired quality $\mathbf{y}^{\mathrm{DES}}$ completely specified or not. In the former case (Scenario 3 in Figure 4.2), by exploiting the direct model inversion, the inversion problem can be written as:

$$
\begin{aligned}
&\min_{\mathbf{t}} \left(\mathbf{t}-\hat{\mathbf{t}}^{\mathrm{DES}}\right)^{\mathrm{T}} \boldsymbol{\Sigma} \left(\mathbf{t}-\hat{\mathbf{t}}^{\mathrm{DES}}\right) + g_2 \cdot \mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}} \\
&s.t. \\
&\hat{\mathbf{t}}^{\mathrm{DES}} = \left(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y}^{\mathrm{DES}} \\
&\hat{\mathbf{y}}^{\mathrm{NEW}} = \mathbf{Qt} \\
&\hat{\mathbf{x}}^{\mathrm{NEW}} = \mathbf{Pt} \\
&\mathbf{t} = \mathbf{W}^{*\mathrm{T}}\mathbf{x}^{\mathrm{NEW}} \\
&\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}} = \left(\hat{\mathbf{x}}^{\mathrm{NEW}}-\mathbf{x}^{\mathrm{NEW}}\right)^{\mathrm{T}}\left(\hat{\mathbf{x}}^{\mathrm{NEW}}-\mathbf{x}^{\mathrm{NEW}}\right) \le g_3 \cdot \mathrm{SPE}_{\mathbf{X},95\%\,\mathrm{lim}} \\
&x_r^{\mathrm{NEW}} = c_r \\
&x_f^{\mathrm{NEW}} \le d_f \\
&lb_k^y \le \hat{y}_k^{\mathrm{NEW}} \le ub_k^y \qquad lb_l^x \le x_l^{\mathrm{NEW}} \le ub_l^x
\end{aligned}
\quad , \qquad (4.3)
$$

being $\mathbf{\Sigma}$ the covariance matrix of the LV scores $\mathbf{T}$ with s_a^2 in the main diagonal (García-Muñoz *et al.*, 2010), c_r the equality constraints for the *r*-th element of $\mathbf{x}^{\mathrm{NEW}}$, d_f the inequality constraint for the *f*-th element of $\mathbf{x}^{\mathrm{NEW}}$, g_2 a parameter weighting the importance of the soft constraint for $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$ in the objective function, while the other symbols have the same meaning as described above. As in Eq.(4.2), a good choice for g_2 is represented by the reciprocal of the 95% confidence limit for the SPE ($\mathrm{SPE}_{\mathbf{X},95\%\,\mathrm{lim}}$), calculated from the historical data in $\mathbf{X}$ used to build the model. On the basis of the experience in applying this method, a reasonable way to limit the model mismatch is to apply a weight $g_3 < 1$ to decrease the value for the obtained $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$.

Note that in the problem of Eq.(4.3) both a soft and a hard constraint are assigned for the model mismatch $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$. The soft constraint is needed to avoid the solution to have an SPE value anchored to the value set by the hard constraint. At the same time, the use of the soft constraint only could still yield a solution with an unacceptable SPE, i.e. higher than $\mathrm{SPE}_{\mathbf{X},95\%\,\mathrm{lim}}$. For this reason, both the soft and the hard constraint on $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$ are needed (García-Muñoz *et al.*, 2010).

In the case not all the desired quality variables in $\mathbf{y}^{\mathrm{DES}}$ are defined (Scenario 4 in Figure 4.2), because some elements are not specified or inequality constraints are given for these, the formulation of the inversion problem is presented in Eq.(4.4) and represents the most complex scenario for this framework.

$$
\begin{aligned}
&\min_{\mathbf{x}^{\mathrm{NEW}}} \left(\hat{\mathbf{y}}^{\mathrm{NEW}} - \mathbf{y}^{\mathrm{DES}}\right)^{\mathrm{T}} \mathbf{\Gamma} \left(\hat{\mathbf{y}}^{\mathrm{NEW}} - \mathbf{y}^{\mathrm{DES}}\right) + g_1 \cdot \left(\sum_{a=1}^{A} \frac{t_a^2}{s_a^2}\right) + g_2 \cdot \mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}} \\
&s.t. \\
&\hat{\mathbf{y}}^{\mathrm{NEW}} = \mathbf{Qt} \\
&\hat{\mathbf{x}}^{\mathrm{NEW}} = \mathbf{Pt} \\
&\mathbf{t} = \mathbf{W}^{*\mathrm{T}} \mathbf{x}^{\mathrm{NEW}} \\
&\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}} = \left(\hat{\mathbf{x}}^{\mathrm{NEW}} - \mathbf{x}^{\mathrm{NEW}}\right)^{\mathrm{T}} \left(\hat{\mathbf{x}}^{\mathrm{NEW}} - \mathbf{x}^{\mathrm{NEW}}\right) \leq g_3 \cdot \mathrm{SPE}_{\mathbf{X},95\%\,\mathrm{lim}} \\
&\hat{y}_j^{\mathrm{NEW}} \leq b_j \\
&x_r^{\mathrm{NEW}} = c_r \\
&x_f^{\mathrm{NEW}} \leq d_f \\
&lb_k^y \leq \hat{y}_k^{\mathrm{NEW}} \leq ub_k^y \qquad lb_l^x \leq x_l^{\mathrm{NEW}} \leq ub_l^x
\end{aligned}
\quad , \qquad (4.4)
$$

with the same meaning for the notation as in the previous scenarios. Note that in this optimization problem, the decision (optimization) variables are included in vector $\mathbf{x}^{\mathrm{NEW}}$, differently from Eq.(4.2) and Eq.(4.3) where the optimization was performed on the score vector $\mathbf{t}$.

The described framework for LVRM inversion has been implemented in Matlab® (the MathWorks Inc., Natick, MA) using an *in-house* developed multivariate analysis toolbox (*phi v1.7*) while, as it concerns to the optimization part, it has been solved in GAMS (GAMS Development Corporation, Washington DC), with an *in-house* developed interface between them.

In the following sections results are presented on the application of the proposed procedures for a real case study from the pharmaceutical industry, concerning a particle engineering problem when a high-shear wet granulation manufacturing route is necessary to obtain the desired product.

4.3 Case studies: latent variable model inversion for particle engineering

In this case study, the proposed general framework is applied to a particle engineering problem for the design of a product under the assumption that it is manufactured using a high-shear wet granulation process. In this example the experimental data reported in the work of Vemavarapu *et al.* (2009) have been used. In the original work, the authors studied the influence of the raw material properties on the performance of a wet granulated product. Each raw material was characterized and processed at fixed process conditions, which provides the necessary information in order to study the effect of the input properties on the final product. The objective of this exercise is the calculation of the optimal values for the properties of the raw materials, to obtain a product of desired quality through wet granulation. Results for three different model inversion exercises are presented.

In the first case study, all the desired properties for the product ($\mathbf{y}^{\mathrm{DES}}$) are fixed by the user (all equality constraints), while no constraints are given for the input material properties in $\mathbf{x}^{\mathrm{NEW}}$ (Scenario 1). Thus, the values of all the variables in $\mathbf{x}^{\mathrm{NEW}}$ are estimated through the inversion of the model. In the second case study, the objective is to design the particle size distribution (PSD) and the surface area for the raw material in input to the wet granulation process, assuming that the other raw material properties are fixed and not adjustable. As in the first case study, the product quality variables in $\mathbf{y}^{\mathrm{DES}}$ are completely defined (all equality constraints; Scenario 3). In the third case study, the objective is the same as in the second case study, but the product quality variables in $\mathbf{y}^{\mathrm{DES}}$ are not completely defined, as upper or lower limits (inequality constraints) are given for some of them (Scenario 4).

4.3.1 Available data and preliminary analysis

Data are collected in two datasets: a dataset $\mathbf{X}$ [25×7] of input material properties, including 25 different materials and 7 measured variables, and a dataset $\mathbf{Y}$ [25×7] of product properties, including the 25 products corresponding to the different input materials with 7 different quality variables measured. The materials were selected in order to cover a large design space with the variables of interest, but at the same time ensuring workable granulations at the chosen process conditions. The measured properties of the input materials and the variables measured for the product characterization are listed in Table 4.2. The reader is encouraged to refer to the original work of Vemavarapu *et al.* (2009) for more information about these variables and the rationale for selecting them.

Table 4.2. *Measured properties for the input materials (***X***) and for the obtained granules (***Y***).*

X	**Y**
1. H_2O solubility (mg/mL)	1. LOD (%)
2. Contact Angle (°)	2. % Oversize
3. H_2O holding capacity (wt % gain)	3. ΔFlodex (mm)
4. D[3,2] (μm)	4. ΔCompactability (kPa/MPa)
5. D90/D10	5. D[3,2] (μm)
6. Surface Area (m^2/g)	6. D90/D10
7. Pore Volume (cm^3/g)	7. Growth Ratio

A preliminary analysis on the available data was performed to understand the statistical rank of the datasets, namely the number of LVs needed to describe the variability in the data. The most useful way to assess it is to perform a PCA on the $\mathbf{X}$ and on the $\mathbf{Y}$ matrices. A summary of the primary diagnostics of the PCA models for $\mathbf{X}$ and for $\mathbf{Y}$ are reported in Table 4.3 in terms of eigenvalues (Eig**X** and Eig**Y**) and explained variance in model design ($R^2\mathbf{X}$ and $R^2\mathbf{Y}$) per latent variable (LV).

Table 4.3. *Diagnostics of the PCA models for the X and Y datasets: eigenvalues and explained variance per latent variable.*

LV	**EigX**	R^2**X**	**EigY**	R^2**Y**
1	3.06	41.10	2.59	37.63
2	1.63	23.32	2.30	30.84
3	1.20	18.00	1.29	18.35
4	0.73	10.72	0.54	7.80
5	0.27	3.63	0.21	2.95
6	0.12	2.09	0.12	1.70
7	0.06	1.01	0.04	0.64

From the results reported in Table 4.3 it can be seen that the eigenvalues for both the data in $\mathbf{X}$ and in $\mathbf{Y}$ are greater than 1 for the first three LVs. Following the eigenvalue-greater-than-one rule (Mardia *et al.*, 1979), three LVs should then be chosen for both PCA models. This seems a robust choice in the case of the $\mathbf{Y}$ dataset, as there is a sharp decrease in the eigenvalue and in $R^2\mathbf{Y}$ between the third and the fourth LV; furthermore the fourth LV eigenvalue is significantly below 1. Differently, in the case of the $\mathbf{X}$ dataset it can be seen from the value of $R^2\mathbf{X}$ that the fourth LV still describes a significant amount of the variability of $\mathbf{X}$, while the subsequent LVs are much less explanatory. For these reasons four LVs seem to be more appropriate to properly describe the systematic variability of the data in $\mathbf{X}$, while three LVs are chosen for the **Y** dataset. This also means that, when a PLS model is fit between $\mathbf{X}$ and $\mathbf{Y}$**,** in the most favorable case (i.e. all LVs in $\mathbf{Y}$ are explained by $\mathbf{X}$; Burnham *et al.*, 1999) there is one LV in the **X** space that has no or little effect on $\mathbf{Y}$ and generates a one-dimensional null space that must be considered in the model inversion exercise.

A PLS model was then built and cross-validated using the $\mathbf{X}$ and $\mathbf{Y}$ datasets. Note that before the design of the PLS model, some of the variables in $\mathbf{X}$ and in $\mathbf{Y}$ were log-transformed, to account for probable nonlinearities, which were highlighted also in the original paper (Vemavarapu *et al.*, 2009). In particular, H_2O solubility, D[3,2], surface area and pore volume among the regressors, and D[3,2] among the quality variables, were log-transformed.

In the design of predictive LVRMs like PLS, the main interest is towards the prediction of the response variables in $\mathbf{Y}$. For this reason, as was shown in Chapter 2 (Section 2.1.2.1), the cross-validation of a regression model is traditionally based on diagnostics depending on samples of $\mathbf{Y}$, like the explained variance in cross validation (Q^2) or the root mean square error of cross-validation (*RMSECV*). Indeed, a LVM captures the covariance structure also for the variables in $\mathbf{X}$. A comprehensive strategy for the selection of the number of LVs needed to design a LVRM should therefore also consider a metric to diagnose the performances of the model in cross-validation using the $\mathbf{X}$ data. This is essential in particular when performing model inversion, where the objective is the estimation of the regressors starting from the response variables, and it must be ensured that the model adequately represents not only the $\mathbf{Y}$ but also the $\mathbf{X}$ space. In this work a new metric (referred to as P^2) is proposed, representing the variance explained by the model in cross-validation for the data in $\mathbf{X}$.

Table 4.4 provides a summary of the diagnostics for the PLS model between $\mathbf{X}$ and $\mathbf{Y}$. In particular, the values of the explained variances for $\mathbf{X}$ and $\mathbf{Y}$ per LV in model building ($R^2\mathbf{X}$ and $R^2\mathbf{Y}$**)** and in cross-validation (Q^2 and P^2) are reported. The corresponding cumulative values ($R^2_{\mathrm{CUM}}\mathbf{X}$ and $R^2_{\mathrm{CUM}}\mathbf{Y}$, Q^2_{CUM} and P^2_{CUM}) are reported as well. Cross-validation was performed with a jackknife approach (Duchesne and MacGregor, 2001).

Table 4.4. *Diagnostics of the PLS model between* **X** *and* **Y**.

LV	$R^2\mathbf{X}$	$R^2_{\mathrm{CUM}}\mathbf{X}$	P^2	P^2_{CUM}	$R^2\mathbf{Y}$	$R^2_{\mathrm{CUM}}\mathbf{Y}$	Q^2	Q^2_{CUM}
1	40.84	40.84	35.71	35.71	32.34	32.35	24.68	24.68
2	18.69	59.54	18.26	53.97	23.34	55.69	24.31	48.99
3	15.99	75.53	15.38	69.35	8.30	63.99	11.72	60.71
4	17.45	92.98	22.03	91.38	1.58	65.57	2.30	63.02
5	2.49	95.47	3.66	95.04	2.77	68.34	2.88	65.90
6	2.93	98.40	2.60	97.64	0.45	68.79	0.10	66.01
7	1.43	99.83	2.17	99.82	0.56	69.34	1.59	67.59

From the results reported in Table 4.4, it can be seen that both $R^2\mathbf{X}$ and P^2 show a significant decrease in the amount of explained variance after the fourth LV. Differently, analyzing the values of $R^2\mathbf{Y}$ and especially of Q^2 it can be seen that the LVs after the third do not explain a significant amount of variance. Even if the analysis of $\mathbf{Y}$ would suggest to use three LVs, four LVs were selected for the PLS model design, considering both the values of $R^2\mathbf{X}$ and P^2 and that the model should adequately describe $\mathbf{X}$ for the model inversion exercise. Indeed, the values reported in Table 4.4 confirm the results of Table 4.3.

4.3.2 Case study 1: product quality completely defined and no constraints on the input variables

In this case study the objective was to design the complete set of chemical and morphological characteristics of the input material ($\mathbf{x}^{\mathrm{NEW}}$), in order to obtain a desired set of product properties ($\mathbf{y}^{\mathrm{DES}}$), assuming a wet granulation process is to be used. It was assumed that all the product properties in $\mathbf{y}^{\mathrm{DES}}$ were fixed by the user while no constraints were considered for the input variable space (Scenario 1). To evaluate the method, the inversion exercise was tested on three completely different product profiles, which were selected from the original datasets in such a way that they spanned a large range of product quality attributes and raw material properties. The considered targets were the granules obtained in the original paper when using Avicel PH200, Isoniazid and maize starch. For each material, the corresponding rows in $\mathbf{X}$ and $\mathbf{Y}$ were extracted from the datasets and the model was rebuilt without the considered target material. The value used for $\mathbf{y}^{\mathrm{DES}}$ was the reconstruction of the real material quality (i.e. the extracted row of $\mathbf{Y}$) through the PCA model built on the historical quality variable space $\mathbf{Y}$. This was done to ensure that the desired quality profile had the same correlation structure as the historical $\mathbf{Y}$. Note that even if no constraints were assigned for the input variable space, some physical bounds were specified for some of the variables in $\hat{\mathbf{x}}^{\mathrm{NEW}}$. In particular:

- The contact angle could vary between 0° and 180°.
- The H_2O holding capacity had to be greater than 0 wt. %.
- D90/D10 had to be greater than 1.

Considering the presence of the null space, the design specifications and the physical bounds, the procedure used for the model inversion is the one described in Eq.(4.2).

Note that the direct validation of any inversion exercise would be the experimental validation of the results obtained *in-silico*. However in this case study it was not possible to perform the experiments to validate the model results. Therefore, the inversion performances were evaluated by comparing the results estimated from the model inversion with the real properties of the considered material. In particular, since the optimization is performed in the model latent space, in Table 4.5 the projections in the LVRM space of the solution obtained through direct inversion ($\hat{\mathbf{t}}^{\mathrm{DES}}$) and through the optimization approach in Eq.(4.2) ($\hat{\mathbf{t}}$) are compared to the projections of the real input material variables ($\hat{\mathbf{t}}^{\mathrm{REAL}}$) for each of the three considered materials. Furthermore, to have a better comparison between the direct inversion and the optimization solution, Table 4.5 reports the values of the Mahalanobis distances (indicated as $\|\cdot\|_M$; Mardia *et al.*, 1979) between $\hat{\mathbf{t}}$ and $\hat{\mathbf{t}}^{\mathrm{REAL}}$, and between $\hat{\mathbf{t}}$ and $\hat{\mathbf{t}}^{\mathrm{DES}}$.

Table 4.5. *Comparison between the optimization solution ($\hat{\mathbf{t}}$), the direct inversion solution ($\hat{\mathbf{t}}^{\mathrm{DES}}$) and the real material property projections ($\hat{\mathbf{t}}^{\mathrm{REAL}}$).*

Material		t_1	t_2	t_3	t_4	$\left\|\hat{\mathbf{t}}-\hat{\mathbf{t}}^{\mathrm{REAL}}\right\|_{\mathrm{M}}$	$\left\|\hat{\mathbf{t}}^{\mathrm{DES}}-\hat{\mathbf{t}}^{\mathrm{REAL}}\right\|_{\mathrm{M}}$
Avicel PH200	$\hat{\mathbf{t}}$	-0.236	1.414	-0.520	0.241	0.487	0.398
	$\hat{\mathbf{t}}^{\mathrm{DES}}$	-0.226	1.678	-0.616	0.985		
	$\hat{\mathbf{t}}^{\mathrm{REAL}}$	-0.210	1.429	-1.125	0.643		
Isoniazid	$\hat{\mathbf{t}}$	-2.894	-1.338	-0.121	-0.576	0.741	4.338
	$\hat{\mathbf{t}}^{\mathrm{DES}}$	-3.001	-2.051	-1.680	-1.740		
	$\hat{\mathbf{t}}^{\mathrm{REAL}}$	-2.481	-1.186	-0.217	-0.347		
Maize Starch	$\hat{\mathbf{t}}$	-0.583	2.593	1.596	0.420	3.023	47.493
	$\hat{\mathbf{t}}^{\mathrm{DES}}$	-0.561	3.579	4.921	-3.818		
	$\hat{\mathbf{t}}^{\mathrm{REAL}}$	-0.506	2.167	1.537	2.105		

As can be seen from Table 4.5, for all the considered materials the results obtained from the optimization ($\hat{\mathbf{t}}$) are very close to the real input material properties projections ($\hat{\mathbf{t}}^{\mathrm{REAL}}$) compared to the direct inversion solution ($\hat{\mathbf{t}}^{\mathrm{DES}}$). In general it can be observed that the difference between the calculated and real scores on the first LVs are lower than the difference between the scores on the last LVs. This is due to the fact that the first LVs explain an higher percentage of the variability of the data, as resulted from Table 4.4, and are then predominant in the calculation of $\hat{\mathbf{y}}^{\mathrm{NEW}}$ and in the minimization of the first term of the objective function in Eq.(4.2).

To better understand the results in Table 4.5, Figure 4.3 reports the projections of the direct inversion solution (○ marker), of the optimized solution (blue □ marker) and of the real input variables (△ marker) in the model score spaces of the first and second LVs and of the second and third LVs respectively. Each plot reports also the scores of the historical samples used to build the model (black dots), the 95% confidence ellipse for the sample scores (dashed black line), and the null space projections on the considered score planes (solid black line).

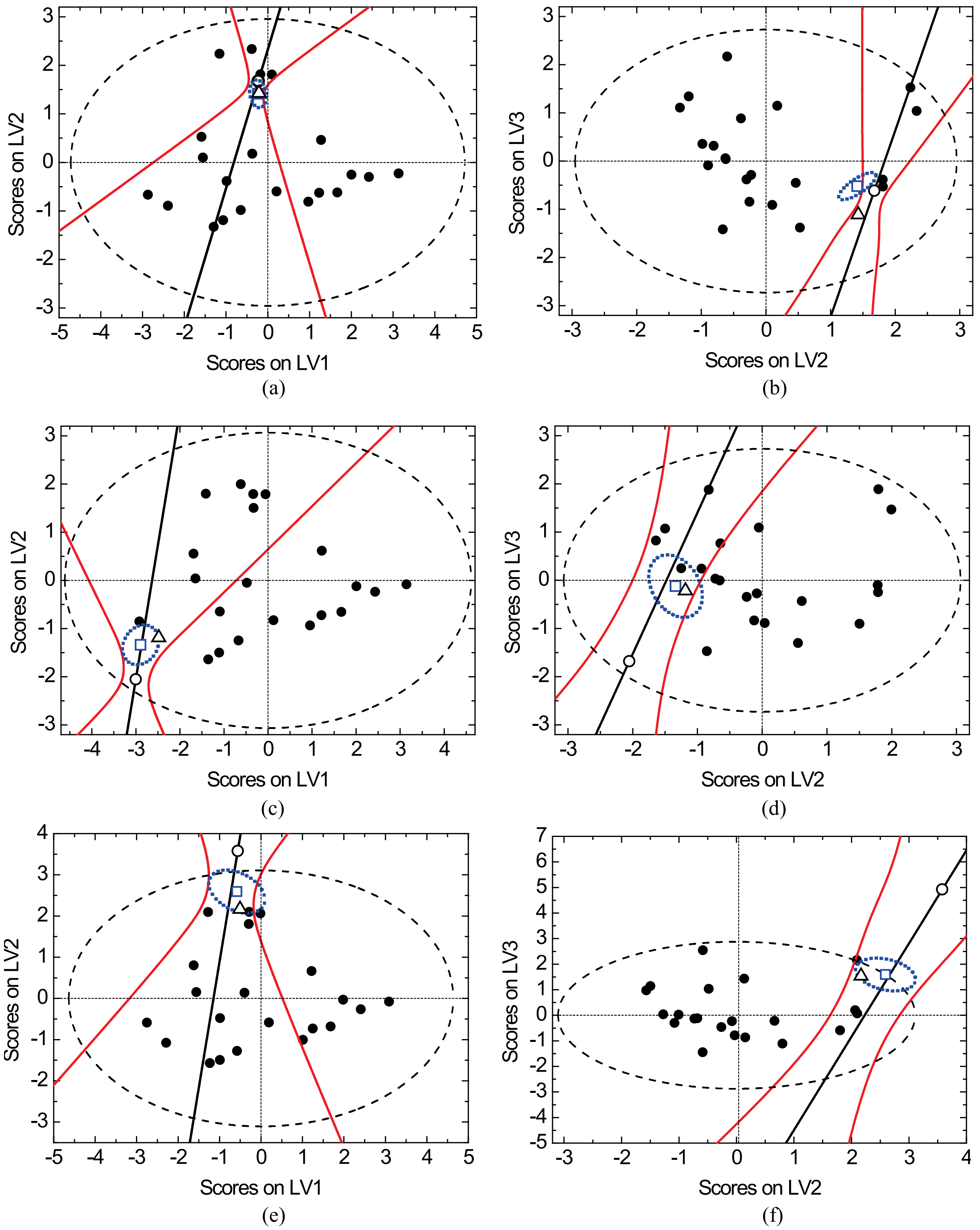


Figure 4.3. *Projections of the LVRM inversion solutions in the model score plots. (a) t_1 vs t_2 for Avicel PH200; (b) t_2 vs t_3 for Avicel PH200; (c) t_1 vs t_2 for Isoniazid; (d) t_2 vs t_3 for Isoniazid; (e) t_1 vs t_2 for maize starch; (f) t_2 vs t_3 for maize starch. In each plot the analytical optimum ($\hat{\mathbf{t}}^{\text{DES}}$, ○), the solution from the optimization in Eq.(4.2) ($\hat{\mathbf{t}}$, □ in blue) and the real input material variable projections ($\hat{\mathbf{t}}^{\text{REAL}}$, △) are reported. The solid black line represents the projection of the null space on the considered planes. The solid red lines represent the uncertainties in the null space calculation, while the dotted blue lines the uncertainty in the optimization solution calculation.*

The calculation of the null space has been performed as described in Chapter 2 (Section 2.2.1). If the rank-deficiency of the matrices were exact, the null space calculation would identify a true multivariate null space. Namely, a point along the null space could be moved without affecting the model estimation for the product quality. However, measurement errors and model mismatch contribute to increase the uncertainties in the estimates of the parameters of the LVRM. As a consequence, these occurrences increase the uncertainties also in the null space estimation. The estimation of the uncertainties is essential to understand the reliability and the quality of the model inversion solution. In this exercise a procedure is proposed and applied to calculate the 95% confidence limits for the estimation of the null space.

The procedure is based on jackknife (Duchesne and MacGregor, 2001) and described in Appendix C. The calculated limits, which are represented by the red lines in the plots of Figure 4.3, represent the variation in the null space calculation due to model uncertainty. Likewise, the uncertainties in the estimation of the inversion solution $\hat{\mathbf{t}}$ were calculated in terms of 95% confidence limits using the same jackknife approach, and are represented as confidence ellipses (dotted blue lines) in the plots of Figure 4.3.

From the analysis of Figure 4.3, it can be observed that the direct model inversion solution $\hat{\mathbf{t}}^{\mathrm{DES}}$ belongs to the null space and that the null space confidence limits are divergent. This is an important occurrence due to the fact that in this case the estimated null space is not properly a pure null space. Namely, it is not properly orthogonal to the $\mathbf{Y}$ space, but it contributes to explain part of the systematic variability in the quality space. This can happen when the null space is a *pseudo-null space*, which is generated from the combination of the single variable (univariate) null spaces (García-Muñoz *et al.*, 2006), or when the model is not representative enough of the historical data (i.e. when the selected LVs explain a limited percentage of the variance of the data or there is a low correlation between the $\mathbf{X}$ and $\mathbf{Y}$ datasets). Thus, the uncertainty in the estimation of the null space is limited when the solution is close to $\hat{\mathbf{t}}^{\mathrm{DES}}$, but is larger for the points of the null space far from $\hat{\mathbf{t}}^{\mathrm{DES}}$, as can be noted from the plots in Figure 4.3. Two important remarks should be emphasized at this point: first, even if the null space is a pseudo-null space and is not completely independent from the $\mathbf{Y}$ space, it still represents the direction of minimum influence on $\mathbf{Y}$; second, since the null space calculation is strongly affected by the model parameter uncertainties related to the model performances in fitting the data, and the uncertainties for the null space increase as the solution is moved away from the direct inversion solution, a model inversion solution should be considered belonging to the null space as long as it falls inside the null space confidence limits, even if this belonging is more uncertain as the solution gets farther from the direct inversion one.

From the analysis of Figure 4.3a and Figure 4.3b it can be seen that in the case of Avicel PH200 the desired (i.e. calculated) material property projections fall inside the historical design space. In this case both the optimization and the direct inversion procedures give an estimation for the input material properties which is very similar to the real ones. This can be appreciated by the Mahalanobis distance values in Table 4.5 and by the fact that the projections of $\hat{\mathbf{t}}$, $\hat{\mathbf{t}}^{\mathrm{DES}}$ and $\hat{\mathbf{t}}^{\mathrm{REAL}}$ are overlapped in the space of the scores on the first two LVs.

Differently, in the case of Isoniazid (Figure 4.3c and Figure 4.3d) and in the case of maize starch (Figure 4.3e and Figure 4.3f) it has been verified that the desired product quality profiles are out of the range of the historically known product quality. Both these cases well describe the role of the null space in the inversion: the optimization procedure moves the direct solution along the null space (or inside its 95% confidence limits) until it finds a compromise solution between the two terms constituting the objective function in Eq.(4.2). In both cases the solution $\hat{\mathbf{t}}$ is very similar to the real input material properties projected onto the latent space compared to the direct inversion solution, as can also be seen from the Mahalanobis distances in Table 4.5.

4.3.2.1 Reducing a null space to practice

The advantage of finding a true multivariate null space in the LVRM inversion is that the solution can move along the null space without affecting the product quality. If a pseudo-null space is found (García-Muñoz *et al.*, 2006), moving the solution along the null-space guarantees the least amount of deviation in the obtained quality set $\hat{\mathbf{y}}^{\mathrm{NEW}}$ versus the target $\mathbf{y}^{\mathrm{DES}}$. Therefore an infinite set of input conditions $\hat{\mathbf{x}}^{\mathrm{NULL}}$ can be obtained from the null space projection points:

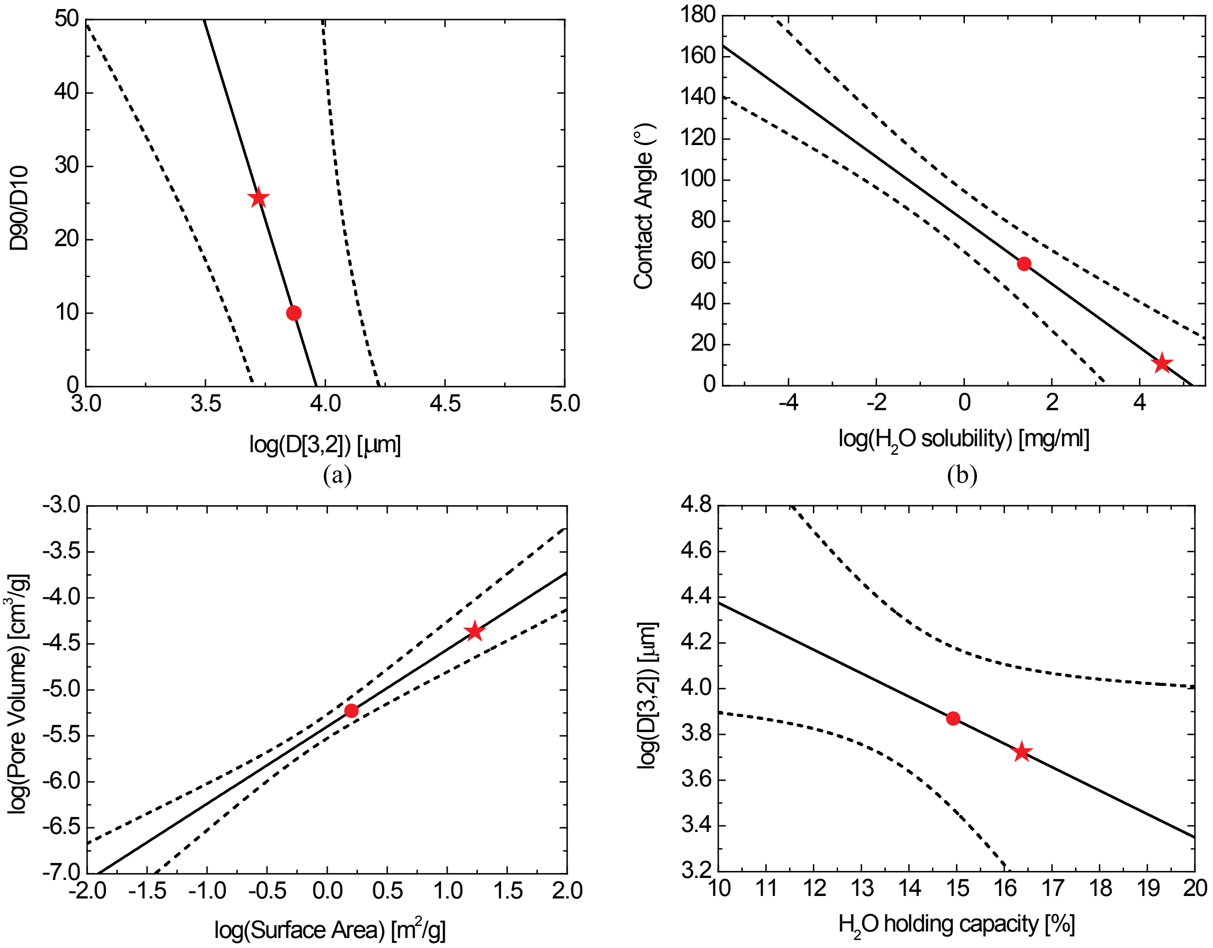
$$\hat{\mathbf{x}}^{\mathrm{NULL}} = \mathbf{P}\mathbf{t}^{\mathrm{NULL}} \quad , \tag{4.5}$$

where $\mathbf{t}^{\mathrm{NULL}}$ is the vector of the scores corresponding to a null space point. All the input conditions $\hat{\mathbf{x}}^{\mathrm{NULL}}$, whose projections belong to the null space, are associated by the correspondence to the same product quality, according to the model, and form a multivariate space of the input variable combinations with no (theoretically) or minimum impact on the product quality.

The null space is a mathematical concept not easy to understand from the design point of view for many reasons. First, it is not defined into the real design space of the input variables but in the reduced space of their projections on the LVs; second it is infinite, while a design space is expected to be finite, since physical variables move in a finite range.

The results along the null space need to be communicated in terms of input variable values to the person or group in charge of implementing the design. For these reasons, there is the need to define a link between the null space and the *design space*, which, following the definition of the regulatory Agencies (Chapter 1, Section 1.2.2), is intended as the space of the input variable combinations that robustly ensure to obtain a defined product in output. This is essential in order for the product/process design personnel to understand what a variation inside the null space means in terms of changes in the input variables.

This link can be established by building the multivariate space of the input variable sets reconstructed from the null space projections according to Eq.(4.5). This would generate an *N*-dimensional space of the combinations of the input variables that (theoretically) correspond to the same assigned product quality.



(a) (b)

(c) (d)

Figure 4.4. *Plots of the input variable combinations belonging to the null space. (a) D90/D10 versus log(D[3,2]). (b) Contact Angle versus log(H_2O solubility). (c) log(D[3,2]) versus log(H_2O solubility). (d) log(Pore Volume) versus log(Surface Area). The solid black lines indicate the points reconstructed from the null space (i.e. the solid black lines in Figure 4.3); the dashed black regions represent the 95% confidence limits. The red dots and the red stars in each diagram represent two different sets of reconstructed variables each one having the same null space projections.*

Since the multivariate space is *N*-dimensional and cannot be represented graphically, in Figure 4.4 the projections of this multivariate space onto bidimensional plots of pairs of the input variables are reported for the Avicel PH200 model inversion exercise. Four of the 21 (in total) diagrams are represented. The diagrams report the projections on the planes of D90/D10 *vs* D[3,2], Contact Angle *vs* H_2O solubility, Pore Volume *vs* Surface Area, and D[3,2] *vs* H_2O

holding capacity. In each plot the solid black line includes the points whose projections belong to the null space (namely to the black line in Figure 4.3a and 4.3b). The different input variable combinations belonging to the null space form the solid black lines in the diagrams of Figure 4.4.

Thus, each point in one of the diagrams of Figure 4.4 corresponds to one appropriate point in each of the other diagrams, according to the combination computed by the model in Eq.(4.5). For example, in each of the diagrams of Figure 4.4 the points highlighted by the red dots correspond to the same set calculated by the same projection on the null space. Analogously, the red stars represent another set of variables reconstructed from the same point of the null space, and here reported to clarify the directionality of the points in the plots. The dashed lines represent instead the 95% confidence limits, which were estimated through the above-mentioned jackknife procedure. Note that the space of the input variable combinations was truncated when some of the variables in the combination resulted out of their physical boundaries or meaningless (e.g. the contact angles greater than 180°).

Finally, note that the relations between the variables highlighted by the diagrams in Figure 4.4 are linear because the model is linear (even if some variables used to build it were nonlinearly transformed). If a nonlinear model had been used the trend would have been nonlinear. This is to say that the shape of the null space depends on the model that was built on the data. Therefore the null space may not exactly represent the design space of the process, but most probably a subset of the design space, which can be helpful as a basis for further experimentation to properly develop it.

4.3.3 Case study 2: product quality completely defined and constraints on the input variables

In this case study it was assumed that only some of the properties of the input material could be modified to obtain a desired product. Specifically, it is assumed that the particle size of the crystals can be modified by milling, and hence a target particle size is needed (D[4,3], D90/D10), which would in turn modify the surface area. All the other variables measured for the input material are assumed to be fixed and not adjustable because of the chemistry of the system. Therefore there are equality constraints both in the input variable and in the quality spaces (Scenario 3). Since all equality constraints are specified for $\mathbf{y}^{\mathrm{DES}}$ and for some elements of $\mathbf{x}^{\mathrm{NEW}}$, the inversion problem is solved exploiting the formulation in Eq.(4.3).

The granules originally obtained with Avicel PH200 are considered as the target product. Since in this case there are no constraints on the T^2 of the solution, it must be checked *a priori* that the desired quality vector is consistent with the covariance in the historical data. The rows corresponding to the target product in **X** and **Y** were then extracted from the datasets and the model rebuilt without the considered material. It is assumed that the row extracted from **Y** corresponding to Avicel PH200 had the desired quality $\mathbf{y}^{\mathrm{DES}}$. More precisely, $\mathbf{y}^{\mathrm{DES}}$ is the reconstruction of the real Avicel PH200 quality set through the PCA model built on the historical quality variable space **Y**, so that it belongs to the quality space. The model was then inverted according to Eq.(4.3) with the specified constraints. Additionally, the constraint D90/D10>1 was considered as done earlier to ensure the soundness of the solution.

In Table 4.6 and Table 4.7 the results from the LVRM inversion are reported in terms of the original variables reconstructed through the PLS model equations. In particular, in Table 4.6 $\mathbf{y}^{\mathrm{DES}}$ is reported together with $\hat{\mathbf{y}}^{\mathrm{NEW}}$, which represents the quality corresponding to the inversion solution $\mathbf{x}^{\mathrm{NEW}}$ through the model. Moreover, the percentage variances explained by the model in cross-validation are reported for each response variable (Q^2_{pv}). Differently, Table 4.7 reports the real values for the input variables $\mathbf{x}^{\mathrm{REAL}}$, the obtained input variable solution $\mathbf{x}^{\mathrm{NEW}}$ and the corresponding value of SPE. For each variable, the value of the percentage variances explained by the model is reported for cross-validation (P^2_{pv}). Note that the values of D[3,2], D90/D10 and surface area calculated through the inversion and the corresponding reference ones are indicated with a # superscript.

Table 4.6. *Desired product quality ($\mathbf{y}^{\mathrm{DES}}$), product quality corresponding to the calculated solution ($\hat{\mathbf{y}}^{\mathrm{NEW}}$), percentage variance explained by the model per response variable in cross-validation (Q^2_{pv}).*

	LOD (%)	Overisize (%)	ΔFlodex (mm)	ΔCompactability (kPa/MPa)	D[3,2] (μm)	D90/D10	Growth ratio
$\mathbf{y}^{\mathrm{DES}}$	2.85	10.85	2.15	-1.37	130	9.3	0.79
$\hat{\mathbf{y}}^{\mathrm{NEW}}$	2.40	15.24	2.78	-0.89	141	10.0	1.8e-8
Q^2_{pv}	0.884	0.452	0.225	0.550	0.756	0.296	0.252

Table 4.7. *Constraints for the input material variables ($\mathbf{x}^{\mathrm{REAL}}$), input conditions calculated through the model inversion ($\mathbf{x}^{\mathrm{NEW}}$), percentage variance explained by the model per input variable in cross-validation (P^2_{pv}).*

	H_2O Solubility (mg/ml)	Contact Angle (°)	H_2O holding capacity (wt. %)	D[3,2]$^{\#}$ (μm)	D90/D10$^{\#}$	Surface Area$^{\#}$ (m^2/g)	Pore Volume (cm^3/g)	$\mathrm{SPE_x}$
$\mathbf{x}^{\mathrm{REAL}}$	0.1	73	10.97	124$^{\#}$	4.6$^{\#}$	1.19$^{\#}$	2.95E-3	0.795
$\mathbf{x}^{\mathrm{NEW}}$	0.1	73	10.97	99$^{\#}$	6.8$^{\#}$	0.98$^{\#}$	2.95E-3	0.696
P^2_{pv}	0.749	0.810	0.963	0.920	0.890	0.885	0.909	-

# indicates an input whose value was *not* assigned as a constraint in the optimization procedure (i.e. a variable *calculated* by model inversion)

As it can be seen from Table 4.6, the optimization procedure finds a solution that is very close to the desired quality values. This can be noticed especially for those variables which are well represented by the model as LOD ($Q^2_{\mathrm{pv}} = 0.884$) and D[3,2] ($Q^2_{\mathrm{pv}} = 0.756$), while for other variables like the growth ratio the estimation is worse. However, note that in the case of the growth ratio the model itself has bad performances if compared to the other variables ($Q^2_{\mathrm{pv}} = 0.252$).

From the analysis of Table 4.7, it can be seen that the values obtained for D[3,2], D90/D10 and for the surface area of the solution are satisfactory if compared to the reference values given by $\mathbf{x}^{\mathrm{REAL}}$. The value of $\mathrm{SPE}_{\mathbf{x}^{\mathrm{NEW}}}$ is slightly smaller than $\mathrm{SPE}_{\mathbf{x}^{\mathrm{REAL}}}$, but they both are under the 95% confidence limit ($\mathrm{SPE}_{\mathbf{X},95\%\,\mathrm{lim}} = 1.218$) that was used as a hard constraint for SPE, with $g_3 = 0.9$ in the relevant constraint in Eq.(4.3).

To confirm the goodness of the LVRM inversion solution and better understand the importance of the terms in the objective function of the problem in Eq.(4.3), the Mahalanobis distances (indicated as $\|\cdot\|_M$) between the scores of the solution obtained though the optimization problem $\hat{\mathbf{t}}$, the real input material variable projections $\hat{\mathbf{t}}^{\mathrm{REAL}}$ and the desired quality projections $\hat{\mathbf{t}}^{\mathrm{DES}}$ are compared in Table 4.8.

Table 4.8. *Mahalanobis distances between the projections of the optimization solution (*$\hat{\mathbf{t}}$*), of the desired quality (*$\hat{\mathbf{t}}^{\mathrm{DES}}$*) and of the real input material variables (*$\hat{\mathbf{t}}^{\mathrm{REAL}}$*).*

$\left\|\hat{\mathbf{t}} - \hat{\mathbf{t}}^{\mathrm{REAL}}\right\|_{\mathrm{M}}$	$\left\|\hat{\mathbf{t}}^{\mathrm{DES}} - \hat{\mathbf{t}}^{\mathrm{REAL}}\right\|_{\mathrm{M}}$
0.045	0.417

As can be seen, the optimization solution ($\hat{\mathbf{t}}$) is much closer to the real input variable projections ($\hat{\mathbf{t}}^{\mathrm{REAL}}$) than is the direct model inversion solution ($\hat{\mathbf{t}}^{\mathrm{DES}}$). This is due to the fact that the first term in the objective function tends to force the solution toward the desired quality projections, while the equality constraints on $\mathbf{x}^{\mathrm{NEW}}$ indirectly force the solution to approach the real input material variable projections $\hat{\mathbf{t}}^{\mathrm{REAL}}$ (which are obviously not known *a priori*). The result is therefore a compromise between these two requirements, with the second term in the objective function additionally aiming at minimizing the SPE of $\mathbf{x}^{\mathrm{NEW}}$. The difference between $\hat{\mathbf{t}}^{\mathrm{DES}}$ and $\hat{\mathbf{t}}^{\mathrm{REAL}}$ not due to the null space forms the model prediction error in the estimation of $\mathbf{y}^{\mathrm{DES}}$.

4.3.4 Case study 3: inequality constraints both on the product quality and on the input variables

In this case study the objective is the same as the one illustrated in Case study 2; however, the variables in $\mathbf{y}^{\mathrm{DES}}$ are not completely specified with equality constraints (Scenario 4). Namely, equality and/or inequality constraints are given for $\mathbf{y}^{\mathrm{DES}}$. In the same way, equality and/or inequality constraints are specified for the input variables in $\mathbf{x}^{\mathrm{NEW}}$. In particular, in this case study the focus is on the design of the PSD (i.e. D[3,2] and D90/D10) and of the surface area of Avicel PH200, in order to obtain a desired product quality $\mathbf{y}^{\mathrm{DES}}$ falling inside an acceptance region defined through inequality constraints. As before, the values of the other variables of $\mathbf{x}^{\text{NEW}}$ are assumed to be fixed due to the chemistry of the system. The following constraints for $\mathbf{y}^{\text{DES}}$ were specified (also reported in Table 4.9):

- LOD greater than 2%;
- Percentage of oversize granules less than 15%;
- ΔFlodex between before and after the granulation greater than 0 mm;
- D[3,2] greater than 140 μm;
- D90/D10 less than 12;
- Growth ratio greater than 3.

No requirements for ΔCompactability were specified, leaving the model to calculate it. All these constraints define a region around the real Avicel PH200 quality variable set, which was used in the previous exercise (i.e. $\mathbf{y}^{\text{DES}}$ in Table 4.6).

Since different types of constraints are involved both in the quality and in the regressor space, the most general inversion procedure problem, formulated in Eq.(4.4), was applied. In Table 4.9 and Table 4.10 the results obtained from the LVRM inversion are shown, together with the constraints enforced on the system. The values of D[3,2], D90/D10 and surface area calculated through the inversion and the corresponding reference ones are indicated with a # superscript in Table 4.10.

Table 4.9. *Desired product quality constraints ($\mathbf{y}^{\text{DES}}$) and product quality corresponding to the calculated solution ($\hat{\mathbf{y}}^{\text{NEW}}$).*

	LOD (%)	Overisize (%)	ΔFlodex (mm)	ΔCompactability (kPa/MPa)	D[3,2] (μm)	D90/D10	Growth ratio
$\mathbf{y}^{\text{DES}}$	≥ 2	≤ 15	> 0	-	≥ 140	≤ 12	≥ 3
$\hat{\mathbf{y}}^{\text{NEW}}$	2.2	15	6.41	-0.46	140	10.4	6.4

Table 4.10. *Constraints for the input material variables ($\mathbf{x}^{\text{REAL}}$) and input conditions calculated through the inversion of the model ($\mathbf{x}^{\text{NEW}}$).*

	H_2O Solubility (mg/ml)	Contact Angle (°)	H_2O holding capacity (wt. %)	D[3,2] (μm)	D90/D10	Surface Area (m^2/g)	Pore Volume (cm^3/g)	SPE$_x$
$\mathbf{x}^{\text{REAL}}$	0.1	73	10.97	124$^{\#}$	4.6$^{\#}$	1.19$^{\#}$	2.95E-3	0.795
$\mathbf{x}^{\text{NEW}}$	0.1	73	10.97	55$^{\#}$	14.4$^{\#}$	0.82$^{\#}$	2.95E-3	0.751

# indicates an input whose value was *not* assigned as a constraint in the optimization procedure (i.e. a variable *calculated* by model inversion)

From Table 4.9 it can be seen that all the constraints enforced on $\mathbf{y}^{\text{DES}}$ are fulfilled by $\hat{\mathbf{y}}^{\text{NEW}}$, which represents the quality corresponding (through the model) to the calculated solution $\mathbf{x}^{\text{NEW}}$. From Table 4.10 it can be seen that the inversion has identified the values of the PSD variables and of the surface area for Avicel PH200 that are needed to obtain a product with the desired quality. It can be noted some difference with the real Avicel PH200 values in

$\mathbf{x}^{\rm REAL}$, even if the values of $\mathrm{SPE}_{\mathbf{x}^{\rm NEW}}$ is similar to that of the SPE of $\mathbf{x}^{\rm REAL}$ ($\mathrm{SPE}_{\mathbf{X},95\%\,\lim} = 1.218$). This is caused by the soft constraint on T^2 in Eq.(4.4), which let the procedure find a solution at a minimum distance from the origin of the model space. The T^2 of the optimization solution is in fact found to be equal to 1.310, while the T^2 of the real input variable projections is equal to 3.368. This provides a conservative approach to the inversion, in order to limit the distance from the historical knowledge and the possibility of extrapolations. However, if the desired quality is found to be inside the historical confidence limits, the soft constraints on T^2 could be relaxed. For example, Table 4.11 reports the Mahalanobis distances (indicated as $\|\cdot\|_M$) between the scores of the solution obtained through the optimization problem $\hat{\mathbf{t}}$, the real input material variable projections $\hat{\mathbf{t}}^{\rm REAL}$ and the real quality projections $\hat{\mathbf{t}}^{\rm DES}$, in the case the soft constraint (SC) on T^2 is considered or not in the Avicel PH200 model inversion problem.

Table 4.11. *Mahalanobis distances between the projections of the optimization solution ($\hat{\mathbf{t}}$), of the real Avicel PH200 quality ($\hat{\mathbf{t}}^{\rm DES}$) and of the real input material variables ($\hat{\mathbf{t}}^{\rm REAL}$) considering or not the soft constraint on T^2 in the problem formulation.*

	$\left\|\hat{\mathbf{t}}-\hat{\mathbf{t}}^{\rm REAL}\right\|_{\rm M}$	$\left\|\hat{\mathbf{t}}-\hat{\mathbf{t}}^{\rm DES}\right\|_{\rm M}$	$\left\|\hat{\mathbf{t}}^{\rm DES}-\hat{\mathbf{t}}^{\rm REAL}\right\|_{\rm M}$
SC on T^2	0.727	0.590	0.417
no SC on T^2	0.259	0.357	0.417

In the first case, it can be seen that in order to limit the distance of the solution from the model space origin, the optimization finds a solution having a distance from the real granule quality ($\left\|\hat{\mathbf{t}}-\hat{\mathbf{t}}^{\rm DES}\right\|_{\rm M}$) and from the real input material variable projections ($\left\|\hat{\mathbf{t}}-\hat{\mathbf{t}}^{\rm REAL}\right\|_{\rm M}$) greater than the distance between them ($\left\|\hat{\mathbf{t}}^{\rm DES}-\hat{\mathbf{t}}^{\rm REAL}\right\|_{\rm M}$). In the second case, if no SC is considered for T^2 it can be seen that the optimization solution is closer to both $\hat{\mathbf{t}}^{\rm REAL}$ and $\hat{\mathbf{t}}^{\rm DES}$ than in the case considering the SC on T^2. Moreover, in this case $\hat{\mathbf{t}}$ is closer than $\hat{\mathbf{t}}^{\rm DES}$ to $\hat{\mathbf{t}}^{\rm REAL}$, as happened in the results of Table 4.8, as the equality constraints on $\mathbf{x}^{\rm NEW}$ pull more the solution toward $\mathbf{x}^{\rm REAL}$.

Finally it must be underlined that the solutions from the LVRM inversion should not be taken as absolute, but the validation through the execution of the suggested experiment is highly recommended, to physically verify the goodness of the procedure, and to better calibrate the model around the design space regions of interest, in order to obtain better estimates from the inversion. The general framework should therefore be applied iteratively: after the first inversion, the results should be validated through the experiments, and the data from the experiments added to the historical database, with which the model should be rebuilt (locally weighting the data) and re-inverted, until an acceptable convergence between the experiments and the LVRM inversion solution is reached.

4.4 Exploiting historical data to design new product quality profiles

As seen in the previous sections, LVRM inversion is built on an optimization framework aiming at identifying a solution (in terms of score vector $\hat{\mathbf{t}}$ or input variables set $\mathbf{x}^{\mathrm{NEW}}$) for which the difference between the desired set of output variables (i.e. product properties) $\mathbf{y}^{\mathrm{DES}}$ and the one estimated through model inversion $\hat{\mathbf{y}}^{\mathrm{NEW}}$ is minimum. As can be seen from the problem formulations in Eqs.(4.2)-(4.4), this can be achieved by setting soft constraints for $\hat{\mathbf{y}}^{\mathrm{NEW}}$ in the formulation of the objective function using *ad-hoc* weights.

Assigning *soft* constraints in the objective function offers the advantage that the equality constraints specified in $\mathbf{y}^{\mathrm{DES}}$ do not necessarily lie in the model sub-space to find a solution to the optimization. Otherwise stated, the optimal solution $\hat{\mathbf{y}}^{\mathrm{NEW}}$ would differ from the desired one in proportion to the orthogonal distance between the assigned values of $\mathbf{y}^{\mathrm{DES}}$ and its projection onto the model hyperplane of the LVs. This may become an issue when the end customer assigns specific values to some of the elements of $\mathbf{y}^{\mathrm{DES}}$ or only slight variations are allowed; in this case, it is necessary to iterate between the customer requirements and a feasible $\mathbf{y}^{\mathrm{DES}}$ complying with the model. Furthermore, there is the inherent need for the user to define weights for each of the terms of the objective function and for each of the elements in $\mathbf{x}^{\mathrm{NEW}}$ and $\mathbf{y}^{\mathrm{DES}}$. Soft constraints add also additional degrees of freedom to the optimizer, thus making the optimization exercise harder.

Mathematically, the alternative is to set *hard* constraints for the elements in $\hat{\mathbf{y}}^{\mathrm{NEW}}$, by forcing them to be equal to the ones in $\mathbf{y}^{\mathrm{DES}}$. The establishment of hard equality constraints reduces the number of iterations in the application of the LVRM inversion procedure to support the design problem, since, once the solution is found, there is no discrepancy between the obtained and the desired values of the response. Moreover, the problem is easier to solve for the optimizer from a numerical point of view. The downside of using hard constraints in the LVRM inversion problem is that the set of constraints given for $\mathbf{y}^{\mathrm{DES}}$ must be coherent with the covariance structure of the original matrices used to build the model. In fact, if hard constraints were used, there is the possibility that the assigned values of $\mathbf{y}^{\mathrm{DES}}$ do not lie on the model hyperplane. In this case, the optimization step may fail due to the possible infeasibility of the hard constraints, since it may be numerically impossible to obtain the desired values for the elements of $\mathbf{y}^{\mathrm{DES}}$ and simultaneously satisfy the covariance structure described by the model.

In this Section, the above-mentioned challenges are addressed by proposing a structured approach to guide the selection of a target attribute profile ($\mathbf{y}^{\mathrm{DES}}$) with the same covariance structure as the matrix of historical response variables in **Y** used to build the model. Given such a vector for $\mathbf{y}^{\mathrm{DES}}$, it is then possible to use hard rather than soft constraints into the optimization formulation for the LVRM inversion, allowing a product developer to achieve the desired values for as many elements of the desired product profile $\mathbf{y}^{\mathrm{DES}}$ as possible.

4.4.1 On the reconstruction of the product target attribute profile

As described in Section 4.2.1, a metric that can be used to quantify the distance of $\mathbf{y}^{\mathrm{DES}}$ from the LVRM space is the squared prediction error ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$). In principle, it would be desired that $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} \approx 0$ so that the hard constraint $y_m^{\mathrm{DES}} = \hat{y}_m^{\mathrm{DES}} = \mathbf{q}_m\mathbf{t}$ for the *m*-th variable specified for $\mathbf{y}^{\mathrm{DES}}$ can be established, where $\mathbf{q}_m$ represents the *m*-th row of the $\mathbf{Q}$ matrix. In order for $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ to be approximately zero, $\hat{\mathbf{y}}^{\mathrm{DES}}$ (namely a reconstruction of $\mathbf{y}^{\mathrm{DES}}$ through the model) should be used instead of $\mathbf{y}^{\mathrm{DES}}$ for the model inversion. Two alternatives can then be considered. In the first case, $\hat{\mathbf{y}}^{\mathrm{DES}}$ could be estimated by directly projecting and reconstructing $\mathbf{y}^{\mathrm{DES}}$ through the model. This strategy was applied in Section 4.4.2 and Section 4.4.3, where the PCA model on **Y** was used to reconstruct $\mathbf{y}^{\mathrm{DES}}$ in order to discard the (possible) uncertainties in the quality variables, which are not handled in the presented model inversion framework.

However, the reconstruction $\hat{\mathbf{y}}^{\mathrm{DES}}$ can be very different from the desired product quality set $\mathbf{y}^{\mathrm{DES}}$, despite being its best reconstruction onto the model space (minimum distance from $\mathbf{y}^{\mathrm{DES}}$). In fact, none of the product quality variables in $\hat{\mathbf{y}}^{\mathrm{DES}}$ will have the same value as that originally specified in $\mathbf{y}^{\mathrm{DES}}$, thus giving a solution $\hat{\mathbf{y}}^{\mathrm{NEW}}$ that may be significantly different from $\mathbf{y}^{\mathrm{DES}}$.

Since the interest is to satisfy the constraints as closely as possible for the desired values of the elements of $\mathbf{y}^{\mathrm{DES}}$, the second alternative is to force some of the elements of $\hat{\mathbf{y}}^{\mathrm{DES}}$ to be equal to those assigned in $\mathbf{y}^{\mathrm{DES}}$, while estimating a proper value for the others (e.g. the conditional mean).

In the following sections two different strategies for the selection of $\hat{\mathbf{y}}^{\mathrm{DES}}$ are proposed. The strategies differ according to the way in which the specifications for the elements in $\mathbf{y}^{\mathrm{DES}}$ (namely, the equality constraints) are managed. In the first approach, one of the elements of $\mathbf{y}^{\mathrm{DES}}$ is assigned at a time, while the other elements are calculated through the model using a direct model inversion approach in order to obtain $\hat{\mathbf{y}}^{\mathrm{DES}}$ belonging to the model space. In the second approach, $\hat{\mathbf{y}}^{\mathrm{DES}}$ is calculated by assigning the largest number of elements of $\mathbf{y}^{\mathrm{DES}}$ still leading to obtain a $\hat{\mathbf{y}}^{\mathrm{DES}}$ within the model space; the number and type of the elements are selected according to an optimal criterion. In both approaches it is assumed that the values for all the *M* elements of $\mathbf{y}^{\mathrm{DES}}$ have been assigned by the user, but the methods can be easily applied even if $L < M$ elements have been specified.

An issue, however, arises on which model to use to project and reconstruct $\hat{\mathbf{y}}^{\mathrm{DES}}$. In general, the covariance structure of the historical data can be optimally described by a PCA model on the historical product dataset **Y** or by the $\mathbf{Q}$ loadings of the PLS model between **X** and **Y** (Chapter 2, Section 2.1.2). Thus $\hat{\mathbf{y}}^{\mathrm{DES}}$ could feasibly be reconstructed either exploiting the PCA or the PLS model loadings. However, the covariance structure described by the PCA model on **Y** could potentially not be as the one described by the $\mathbf{Q}$ loadings of the PLS model, given the different objectives of the two techniques (see Chapter 2).

Given that the objective of the proposed procedures is to allow the use of a hard constraint for $\mathbf{y}^{\mathrm{DES}}$ in the optimization formulation of the LVRM inversion problem, the reconstruction of $\hat{\mathbf{y}}^{\mathrm{DES}}$ has been based on the $\mathbf{Q}$ loadings of the PLS model. The proposed methodologies can however be applied with no modifications to the case in which the PCA loadings are used to reconstruct $\hat{\mathbf{y}}^{\mathrm{DES}}$. Further details on this issue are provided in Appendix C.

4.4.1.1 Theoretical considerations

Let us consider the vector of the desired product profile $\mathbf{y}^{\mathrm{DES}}$ autoscaled according to the mean and the standard deviation of the columns in the historical data of product properties (**Y**) used to build the model. Although the method outlined in the subsequent section (Method 1; Section 4.4.1.2) is proposed to provide an estimate of the $M-1$ free elements of $\mathbf{y}^{\mathrm{DES}}$, given an equality constraint enforced on each *i*-th element y_i^{DES} of $\mathbf{y}^{\mathrm{DES}}$, it is reasonable to think that such an estimate is uniquely defined only if the $M-1$ free elements of $\mathbf{y}^{\mathrm{DES}}$ have a strong correlation with the *i*-th that is being assigned. This situation would imply that the effective rank of **Y** is one (only one LV is necessary to represent **Y**) and hence the reconstruction of $\hat{\mathbf{y}}^{\mathrm{DES}}$ based on one element is reasonable.

However, if the effective rank of **Y** is of higher order, assigning the value of the *i*-th element of $\mathbf{y}^{\mathrm{DES}}$ would create an *induced null space* where multiple values of $\mathbf{t}$ can provide the same predicted value for y_i^{DES} while providing multiple possible values for the *M*-1 free elements of $\mathbf{y}^{\mathrm{DES}}$, depending on the correlation structure of **Y**. This artificial null space will change depending on what element of $\mathbf{y}^{\mathrm{DES}}$ is being assigned and can be explicitly determined by the linear system of equations $\hat{\mathbf{y}}^{\mathrm{DES}} = \mathbf{Qt}$ represented by:

$$\begin{array}{llllll}
\hat{y}^{\mathrm{DES}}(1) = & q_{1,1}t_1 & +q_{1,2}t_2 & +q_{1,3}t_3 & \cdots & +q_{1,A}t_A \\
\hat{y}^{\mathrm{DES}}(2) = & q_{2,1}t_1 & +q_{2,2}t_2 & +q_{2,3}t_3 & \cdots & +q_{2,A}t_A \\
\hat{y}^{\mathrm{DES}}(3) = & q_{3,1}t_1 & +q_{2,2}t_2 & +q_{3,3}t_3 & \cdots & +q_{3,A}t_A \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\hat{y}^{\mathrm{DES}}(M) = & q_{M,1}t_1 & q_{M,2}t_2 & q_{M,3}t_3 & \cdots & q_{M,A}t_A
\end{array} , \tag{4.6}$$

being $q_{m,a}$ in Eq.(4.6) the element on the *i*-th row and *j*-th column of the $\mathbf{Q}$ matrix, and t_a the *a*-th element of the column vector $\mathbf{t}$, with $m = 1,...,M$ and $a = 1,...,A$. For illustrative purposes, consider four scenarios:

a) the variables of **Y** are completely independent (**Y** is full rank) and each variable is represented by one LV;

b) **Y** is full rank, but the variables are explained in groups by different LVs;

c) **Y** is rank deficient and correlation is captured in *A* components, where $A < M$;

d) **Y** is rank deficient and correlation is captured in *A* components, where $A \leq M$, but $A > \mathrm{rank}(\mathbf{Y})$.

The appearance of these induced null spaces is obvious in case *a)*, where the right hand of Eq.(4.6) is reduced to the main diagonal elements (Eq.(4.7)). If an equality constraint is applied to one of the elements of $\mathbf{y}^{\mathrm{DES}}$, the rest of the equations define the available null space in which the user can pick any value of the free scores.

$$\begin{array}{ll} \hat{y}^{\mathrm{DES}}(1) = & q_{1,1}t_1 \\ \hat{y}^{\mathrm{DES}}(2) = & \qquad q_{2,2}t_2 \\ \hat{y}^{\mathrm{DES}}(3) = & \qquad\qquad q_{3,3}t_3 \\ \quad\vdots & \qquad\qquad\qquad \ddots \\ \hat{y}^{\mathrm{DES}}(M) = & \qquad\qquad\qquad\quad q_{M,A}t_A \end{array} . \tag{4.7}$$

Scenario *b)* would imply that there are as many independent directions of variability in **Y** as columns in it; however, each direction of change affects multiple variables. Consider the scenario of a three dimensional space with three LVs such that the representative loadings across all LVs are as in Eq.(4.8). Given the below situation and a constraint placed on the first element of $\mathbf{y}^{\mathrm{DES}}$, one could isolate t_1 from the first row and replace it into the third row to end with a system of two equations with four unknowns ($t_2, t_3, \hat{y}^{\mathrm{DES}}(2)$ and $\hat{y}^{\mathrm{DES}}(3)$). This system represents the two dimensional induced null space where any solution chosen for t_2 or t_3 can be used to estimate t_1 and satisfy the equality condition for the first element of $\mathbf{y}^{\mathrm{DES}}$ while also keeping the vector $\hat{\mathbf{y}}^{\mathrm{DES}}$ in the latent space.

$$\begin{array}{llll} \hat{y}^{\mathrm{DES}}(1) = & q_{1,1}t_1 & & +q_{1,3}t_3 \\ \hat{y}^{\mathrm{DES}}(2) = & & q_{2,2}t_2 & \\ \hat{y}^{\mathrm{DES}}(3) = & q_{3,1}t_1 & & +q_{3,3}t_3 \end{array} . \tag{4.8}$$

In scenario *c)*, the number of LVs is lower than the number of variables in **Y** and not all variables are represented in all latent spaces (e.g., Eq.(4.9)). In such case, a hard constraint enforced on the first element will assign the value of the fourth and will result in a one dimensional induced null space where the user can choose any value of t_1 (which in turn defines the values of the second and third element of $\hat{\mathbf{y}}^{\mathrm{DES}}$).

$$\begin{array}{lll} \hat{y}^{\mathrm{DES}}(1) = & & q_{1,2}t_2 \\ \hat{y}^{\mathrm{DES}}(2) = & q_{2,1}t_1 & \\ \hat{y}^{\mathrm{DES}}(3) = & q_{3,1}t_1 & \\ \hat{y}^{\mathrm{DES}}(4) = & & q_{4,2}t_2 \end{array} . \tag{4.9}$$

Finally, consider scenario *d)*. In this case the number of LVs is lower than the number of variables in **Y**, however it is greater than the effective rank of **Y**. As it was shown in Chapter

2 (Section 2.2), this is a common situation when building a PLS model between two matrices **X** and **Y**, and $\mathrm{rank}(\mathbf{X}) > \mathrm{rank}(\mathbf{Y})$. In these cases, if the PLS model is built with $A = \mathrm{rank}(\mathbf{Y})$ LVs, a null space due to the different matrix rank is generated, which gives additional degrees of freedom in the estimation of the score vector $\mathbf{t}$, in addition to the induced null spaces which can generate in situations similar to those described above. For example, assume the same case as in Eq. (4.9) in which the **Y** space is four-dimensional and $\mathrm{rank}(\mathbf{Y}) = 2$, but consider that three LVs were chosen to build the PLS model, as they were needed to represent adequately the **X** space (Eq.(4.10)). In this case, a hard constraint imposed on the first element will result in the estimation of t_2, but the user can choose any value for t_1 and t_3. The system represents a two dimensional null space, which however is formed by the combination of a one-dimensional induced null space and the PLS null space due to the differences in the ranks of **X** and **Y**.

$$\begin{aligned}
\hat{y}^{\mathrm{DES}}(1) &= && q_{1,2}t_2 && \\
\hat{y}^{\mathrm{DES}}(2) &= q_{2,1}t_1 && && \\
\hat{y}^{\mathrm{DES}}(3) &= q_{3,1}t_1 && && +q_{3,3}t_3 \\
\hat{y}^{\mathrm{DES}}(4) &= && q_{4,2}t_2 && +q_{4,3}t_3
\end{aligned} \quad . \tag{4.10}$$

Note that the case presented in scenario *d)* (Eq.(4.10)) could only occur when $\hat{\mathbf{y}}^{\mathrm{DES}}$ is reconstructed through the PLS **Q** loadings, differently from the situations described in the previous scenarios which are valid also in the cases in which $\hat{\mathbf{y}}^{\mathrm{DES}}$ is reconstructed based on the PCA loadings (Appendix C).

In practice, it is common to have only desirable ranges for some quality attributes of the product while having specific assigned conditions for other quality descriptors. In the following, methodologies are proposed to handle the free elements of $\mathbf{y}^{\mathrm{DES}}$ as missing data for the sake of simplicity and to expedite the decision of a vector of quality properties that will result in the target overall performance for a new product. Other researchers have already presented and studied the behavior of the analytical estimators of missing data methods and have already discussed whether the expected predicted value for the missing elements are close to the unconditional mean, the conditional mean, the least Mahalanobis distance or the least SPE (Nelson *et al.*, 1996; Arteaga and Ferrer, 2002). From the perspective of this application, similar approaches are used here as a shortcut to the construction of potential vectors representing the target quality profile for a product. The discussion on the induced null space is presented due to the obvious reaction towards estimating the majority of the $\hat{\mathbf{y}}^{\mathrm{DES}}$ vectors based on one element (in the worst of the cases, they will be nothing but the unconditional means). In other words, the presented methods offer a simple way to appreciate the tradeoffs of assigning one element of the quality profile versus another one, as discussed in the following sections.

4.4.1.2 Method 1: assigning one quality variable at a time

In the first method, the selected product quality set $\hat{\mathbf{y}}^{\mathrm{DES}}$ is calculated imposing that, for the m-th element of $\hat{\mathbf{y}}^{\mathrm{DES}}$, $\hat{y}^{\mathrm{DES}(m)} = y^{\mathrm{DES}(m)}$, while the values of the other elements in $\hat{\mathbf{y}}^{\mathrm{DES}}$ are assumed to be missing.

Several approaches have been proposed to deal with missing data when using multivariate statistical techniques like PCA or PLS, especially in multivariate statistical process control or process modeling applications (Nelson *et al.*, 1996; Arteaga and Ferrer, 2002). In all these contributions the objective was to use the model to estimate the scores corresponding to a new sample presented to the model characterized by missing measurements in the input data (e.g., in the regressor side, if a PLS model is considered). Differently, in this application, in order to reconstruct the product quality profile, the measurements referred to the response set $\hat{\mathbf{y}}^{\mathrm{DES}}$ are considered as missing, thus using the model inversion to reconstruct them on the basis of the available fixed values. The proposed method exploits the sub-model constituted by the PLS $\mathbf{Q}$ loadings to reconstruct the new product target profile $\hat{\mathbf{y}}^{\mathrm{DES}}$ through a direct inversion of the PLS model (Eq.(2.57)).

For each variable m, the proposed procedure aims at estimating the scores $\hat{\mathbf{t}}^{(m)}$ of $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ on the basis of the m-th element of $\mathbf{y}^{\mathrm{DES}}$ ($y^{\mathrm{DES}(m)}$) which is assigned, by projecting it back to the model plane through a direct inversion of the model:

$$\hat{\mathbf{t}}^{(m)} = \left(\mathbf{Q}^{(m)\,\mathrm{T}}\mathbf{Q}^{(m)}\right)^{-1}\mathbf{Q}^{(m)\mathrm{T}}y^{\mathrm{DES}(m)} \quad , \tag{4.11}$$

where $\mathbf{Q}^{(m)}$ is the sub-matrix of the loadings $\mathbf{Q}$ in which only the $[1\times A]$ row of $\mathbf{Q}$ corresponding to the element assigned in $y^{\mathrm{DES}(m)}$ is considered.

This method is applied to all the M variables specified for $\mathbf{y}^{\mathrm{DES}}$, giving then in output a matrix $\hat{\mathbf{Y}}^{\mathrm{DES}^{\mathrm{T}}}$, whose columns are the M different reconstructions for the product quality $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ obtained assigning in turn each element m. The procedure goes through the following steps:

1. Assign the value of the m-th element of $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ in order for it to be equal to the corresponding element in $\mathbf{y}^{\mathrm{DES}}$ ($\hat{y}^{\mathrm{DES}(m)} = y^{\mathrm{DES}(m)}$), considering the other elements in $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ as missing data.
2. Estimate the score vector $\hat{\mathbf{t}}^{(m)}$ corresponding to $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ through the direct model inversion in Eq.(4.11).
3. Reconstruct $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ from $\hat{\mathbf{t}}^{(m)}$ and the $\mathbf{Q}$ loadings of the PLS model and store it in the matrix $\hat{\mathbf{Y}}^{\mathrm{DES}^{\mathrm{T}}}$:

$$\hat{\mathbf{y}}^{\mathrm{DES}(m)} = \mathbf{Q}\hat{\mathbf{t}}^{(m)} \quad . \tag{4.12}$$

4. Assign the next desired product property, until all the M properties in $\mathbf{y}^{\mathrm{DES}}$ have been considered.

Thus, from the different suggestions for the new product quality set $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$, the user can have an idea of the mutual variation of the variables according to the historical knowledge, and select the combination involving the most interesting product property.

In some cases, reconstructing $\hat{\mathbf{y}}^{\mathrm{DES}(m)}$ through the direct inversion of the model (Eq.(4.11)) may lead to an unfeasible solution clashing against the physical limits some of the product quality variables may have. If that occurs, additional flexibility to the procedure described for Method 1 may be added by substituting the step 2 and 3 of the above procedure (i.e. Eq.(4.11) and Eq.(4.12)) with the optimization problem described in the next section (Eq.(4.13)). That will be better clarified when discussing the case study results in Section 4.4.2.1.

4.4.1.3 Method 2: assigning more than one quality variables

The second proposed method is based on an approach that is somehow dual to Method 1. The method starts from the originally defined product quality vector $\mathbf{y}^{\mathrm{DES}}$, and uses an iterative procedure to progressively find the assigned variables in $\mathbf{y}^{\mathrm{DES}}$ that contribute the most to the $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ value, and to relax the corresponding equality constraints until a new estimated desired product quality $\mathbf{y}^{\mathrm{NEW}}$ is obtained that is as close as possible to the model space. Let us assume that the variables in $\mathbf{y}^{\mathrm{DES}}$ are completely (or for the most part) assigned (all equality constraints) and let us define a (small) threshold ε setting the acceptability limit for the value of $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$. A general schematic of the procedure is reported in Figure 4.5. After setting $\mathbf{y}^{\mathrm{NEW}} = \mathbf{y}^{\mathrm{DES}}$, at each iteration the procedure verifies if the desired set $\mathbf{y}^{\mathrm{NEW}}$ belongs to the model space, by calculating $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ and comparing it to ε. If this is not verified, the contributions to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ are calculated according to Eq.(2.20) (Chapter 2, Section 2.1.1.3). The element in $\mathbf{y}^{\mathrm{NEW}}$ with the highest contribution to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ is selected and relaxed. An optimization problem is then solved to update the design set $\mathbf{y}^{\mathrm{NEW}}$:

$$\begin{aligned}
&\min_{\hat{\mathbf{t}}}\left(\mathbf{y}^{\mathrm{NEW}} - \mathbf{Q}\hat{\mathbf{t}}\right)^{\mathrm{T}} \boldsymbol{\Gamma}\left(\mathbf{y}^{\mathrm{NEW}} - \mathbf{Q}\hat{\mathbf{t}}\right) + g \cdot \sum_{a=1}^{A} \frac{\hat{t}_a^2}{s_a^2} \\
&s.t. \\
&y_m^{\mathrm{NEW}} = y_m^{\mathrm{DES}} \qquad\qquad , \qquad\qquad (4.13) \\
&y_j^{\mathrm{NEW}} < b_j \\
&lb_k < y_k^{\mathrm{NEW}} < ub_k
\end{aligned}$$

where the equality constraint $y_m^{\mathrm{NEW}} = y_m^{\mathrm{DES}}$ is set for all the elements in $\mathbf{y}^{\mathrm{DES}}$, except the relaxed ones. The meaning of the other symbols is the same as in Eq.(4.2).

From the optimization problem, a new set $\mathbf{y}^{\mathrm{NEW}}$ representing the new estimated target quality profile is obtained, which is again assessed against threshold ε. The procedure of progressively relaxing the equality constraints initially specified for the elements of $\mathbf{y}^{\mathrm{DES}}$ is repeated until $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ is found below the given threshold ε. Then, a new product quality set $\mathbf{y}^{\mathrm{NEW}}$ $(=\hat{\mathbf{y}}^{\mathrm{DES}})$ is obtained, which represents the best compromise between the set of the target quality profile initially defined by the user ($\mathbf{y}^{\mathrm{DES}}$) and the model requirements.

Conversely, if after relaxing all the equality constraints in $\mathbf{y}^{\mathrm{DES}}$, $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}} > \varepsilon$ still holds, then the problem is unfeasible and a revision on the constraints specified in Eq.(4.13) should be considered.

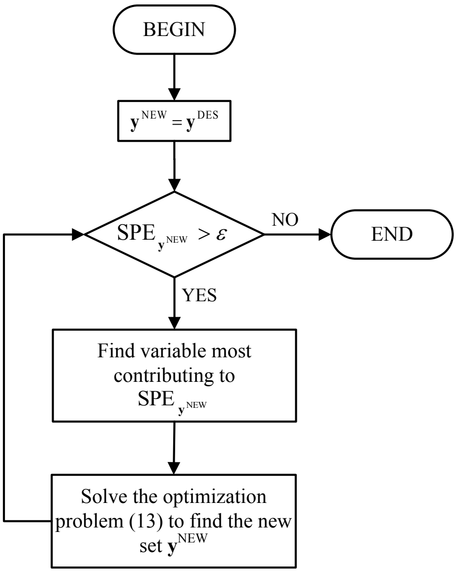

BEGIN
$\mathbf{y}^{\mathrm{NEW}} = \mathbf{y}^{\mathrm{DES}}$
$\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}} > \varepsilon$
NO
END
YES
Find variable most contributing to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$
Solve the optimization problem (13) to find the new set $\mathbf{y}^{\mathrm{NEW}}$

Figure 4.5. *Schematic of the algorithm implemented for Method 2.*

The second term of the objective function in Eq.(4.13) represents the Hotelling's T^2 of the solution. This term is added to the objective function to consider the cases in which an induced null space is present due to the structure of the loadings $\mathbf{Q}$. In these cases, the null space can be exploited to move the solution along it, in order to find a new set $\mathbf{y}^{\mathrm{NEW}}$ that belongs to the model space, but at the same time is inside (or close to) the range of the properties of the historical products (thus avoiding extrapolated solutions). To this end, $g \neq 0$ and reliably $g << 1$ in Eq.(4.13), in order to give more importance in the objective function to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ rather than to the Hotelling's T^2. In the case the null space is due to the differences in the ranks of the **X** and **Y** matrices (namely, the $\mathbf{Q}$ loading matrix is redundant), g can be set to zero, since the Hotelling's T^2 of the solution can be considered in the subsequent PLS inversion problem, for the estimation of the regressors which provide the desired responses $\mathbf{y}^{\mathrm{NEW}}$, by applying one of the previously proposed scenarios in Figure 4.2.

Finally, note that the solution of Method 2 coincides with that of Method 1 (namely, the direct inversion of the model) if an equality constraint for only one element of $\mathbf{y}^{\mathrm{DES}}$ is assigned in Eq.(4.13), $g = 0$ and no inequality constraints or boundaries are present in the inversion problem.

4.4.2 Case study: defining the quality profile for a wet-granulated product

The proposed methodologies are applied to a particle engineering problem to design the quality profile of a wet-granulated product. The case study, the available datasets and the considered PLS model are the same as described in Section 4.3.

For the aims of this study, it is assumed that it is required to manufacture a granulated product with the characteristics $\mathbf{y}^{\mathrm{DES}}$ reported in Table 4.12. Note that these data do not correspond to a real product, but in general they represent an example of the combination of desirable properties for a wet granule.

Table 4.12. *Desired product properties* $\mathbf{y}^{\mathrm{DES}}$ *for a wet-granulated product.*

	LOD (%)	Oversize (%)	ΔFlodex (mm)	ΔCompactability (KPa/MPa)	D[3,2] (μm)	D90/D10	Growth ratio
$\mathbf{y}^{\mathrm{DES}}$	3	0	20	5	400	2.5	8

Before proceeding with any inversion, it must be ensured that the model is able to adequately describe the desired product quality $\mathbf{y}^{\mathrm{DES}}$ in Table 4.12. $\mathbf{y}^{\mathrm{DES}}$ is then projected onto the latent space of the historical samples in **Y**. Figure 4.6 reports the values of the SPE versus the values of the Hotelling's T^2 for the historical samples (black dots), together with the relevant 95% (red short-dashed lines) and 99% (blue dashed lines) confidence limits. The squared dot (□) represents the values of $T^2_{\mathbf{y}^{\mathrm{DES}}}$ and $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$. As it can be seen, even if $T^2_{\mathbf{y}^{\mathrm{DES}}}$ is inside the 95% (= 10.97) confidence limit, meaning that the values specified in $\mathbf{y}^{\mathrm{DES}}$ are not far from the mean of the historical values, $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ is above the 95% relevant limit (= 5.75), meaning that the historical correlation structure is not valid for $\mathbf{y}^{\mathrm{DES}}$. Thus, the model is not really appropriate in representing $\mathbf{y}^{\mathrm{DES}}$ due to the high model mismatch, and it is not recommended to perform the inversion with $\mathbf{y}^{\mathrm{DES}}$.

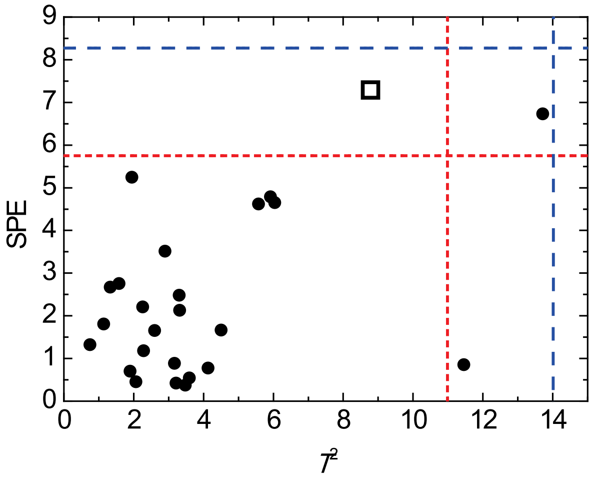


Figure 4.6. *Plot of SPE versus T^2 values for the historical products in* **Y** *(●) and the new desired product quality set* $\mathbf{y}^{\mathrm{DES}}$ *(□). The lines represent respectively the 95% (short-dashed red) and 99% (dashed blue) confidence limits.*

The methods previously described can then be applied to exploit the historical covariance structure of the data in order to give suggestions on possible new product properties sets $\hat{\mathbf{y}}^{\mathrm{DES}}$, which can be feasibly used for the model inversion.

4.4.2.1 Method 1: results

In Table 4.13 the results obtained after applying Method 1 using the proposed direct inversion approach are reported. Each row represents a vector $\hat{\mathbf{y}}^{\mathrm{DES}}$ suggested by the model, obtained fixing one at a time each one of the 7 properties of $\mathbf{y}^{\mathrm{DES}}$ in Table 4.12. The assigned properties for each of the 7 $\hat{\mathbf{y}}^{\mathrm{DES}}$ sets are those bold in brackets in Table 4.13, and coincide with the original values of $\mathbf{y}^{\mathrm{DES}}$ in Table 4.12. In the two last columns of the table, the values of the T^2 and of the SPE are also reported for each calculated set.

Table 4.13. *Method 1. New sets of product properties* $\hat{\mathbf{y}}^{\mathrm{DES}}$ *for a wet-granulated product suggested on the basis of the historical data model. The bold values represent the assigned element in each set (the original values of* $\mathbf{y}^{\mathrm{DES}}$ *are reported in brackets). The proposed approach has been used to calculate the others.* T^2 *and SPE statistics are also given.*

	LOD (%)	**Oversize (%)**	**ΔFlodex (mm)**	**ΔCompactibility (KPa/MPa)**	**D[3,2] (μm)**	**D90/D10**	**Growth ratio**	T^2	SPE
$\hat{\mathbf{y}}^{\mathrm{DES}(1)}$	**3 (3)**	13.5	10.2	-0.58	137	8.7	12.2	0.90	~ 0
$\hat{\mathbf{y}}^{\mathrm{DES}(2)}$	2.6	**0 (0)**	11.2	0.22	68.4	15.0	11.7	1.00	0
$\hat{\mathbf{y}}^{\mathrm{DES}(3)}$	1.4	22.9	**20 (20)**	1.47	147	12.3	30.4	1.33	0
$\hat{\mathbf{y}}^{\mathrm{DES}(4)}$	-2.1	34.7	21.6	**5.00 (5)**	111	22.6	35.4	7.64	~ 0
$\hat{\mathbf{y}}^{\mathrm{DES}(5)}$	1.1	41.1	9.4	0.28	**400 (400)**	4.5	17.5	0.44	~ 0
$\hat{\mathbf{y}}^{\mathrm{DES}(6)}$	1.8	38.2	8.4	-0.49	411	**2.5 (2.5)**	15.1	0.54	~ 0
$\hat{\mathbf{y}}^{\mathrm{DES}(7)}$	1.8	21.1	6.3	0.03	166	9.6	**8.0 (8)**	0.35	~ 0

As it can be seen, for all the suggested new product quality sets $\hat{\mathbf{y}}^{\mathrm{DES}}$, the calculated values for the assigned elements of $\hat{\mathbf{y}}^{\mathrm{DES}}$ are equal to the corresponding equality constraint $\mathbf{y}^{\mathrm{DES}}$ (reported between paranthesis). Moreover, from the values of the SPE, the suggested variable combinations result to be all lying onto the model space.

It is interesting to note that for each of the calculated sets $\hat{\mathbf{y}}^{\mathrm{DES}}$, the covariance structure of the historical samples in $\mathbf{Y}$ is optimally used in the proposed approach to estimate the new sets $\hat{\mathbf{y}}^{\mathrm{DES}}$, even if only one of the 7 variables is constrained in each case. Prior to such discussion, the reader is referred to Table 4.14, where the mean and standard deviation values of the variables in $\mathbf{Y}$ are provided for a clear comparison with the results of Table 4.13, and to Figure 4.7, which reports the loadings $\mathbf{q}$ (i.e. the columns of the $\mathbf{Q}$ matrix) of the PLS model on the four considered LVs (see Table 4.4 for model diagnostics). The loadings have been standardized by $R^2_{\mathrm{pv},y}$ to get a better contrast and to allow for a "cross-component" analysis. The interpretation of these plots goes beyond the aims of this discussion and is reported in

Appendix A. Nonetheless, it is interesting to note that the analysis of these plots allows assessing the main driving forces, which explain the variability in the **Y** data most related to **X**, that are exploited by the proposed methodology for the selection of the new target product quality profiles $\hat{\mathbf{y}}^{\mathrm{DES}}$.

Table 4.14. *Mean and standard deviation (st. dev.) values for the properties of the wet granulated products in the historical database* **Y**.

	LOD (%)	Oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	D[3,2] (μm)	D90/D10	Growth ratio
mean	1.55	24.5	10.6	0.5	181	9.8	15.8
st. dev.	1.41	29.1	9.6	1.9	410	13.1	15.1

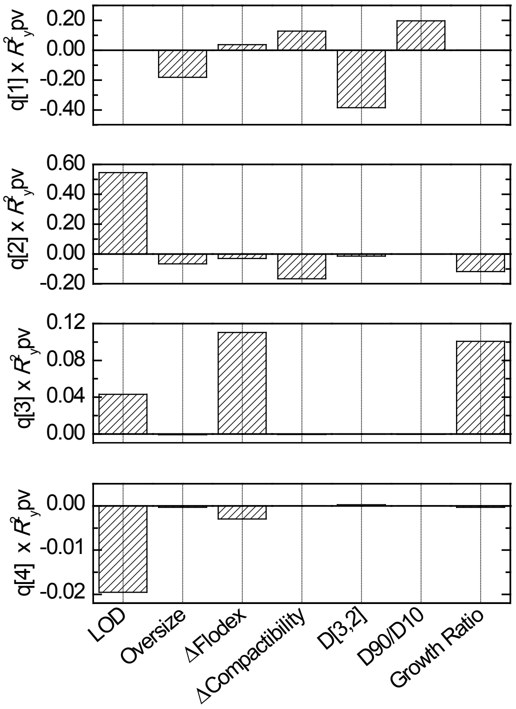


Figure 4.7. *Loadings* **Q** *of the PLS model on the* **X** *and* **Y** *datasets.*

Let us consider the case in which the percentage of oversize granules was assigned to 0 ($\hat{\mathbf{y}}^{\mathrm{DES}(2)}$), which is quite different from the historical mean in Table 4.14. From Figure 4.7, it can be seen that this variable is related to D[3,2] and inversely related to D90/D10 on LV1; furthermore, it is scarcely related to the other LVs. In fact, giving that in $\hat{\mathbf{y}}^{\mathrm{DES}(2)}$ the oversize percentage is lower than the historical mean, the suggested set is characterized by a low value of D[3,2] and a high value of D90/D10 (even if within one standard deviation as from the values in Table 4.14). The other variables are more similar to their unconditional mean (Table 4.14), because they do not show a strong relation with the oversize percentage (Figure 4.7).

This kind of analysis can be repeated for the other sets in Table 4.13. In particular, in the case of $\hat{\mathbf{y}}^{\mathrm{DES}(4)}$, a value of ΔCompactability out of the range of the historical product data was assigned. In this case, it can be observed that in order to obtain such a value of

ΔCompactability, the product needs to exhibit a low and broad PSD, and higher-than-mean oversize percentage and ΔFlodex. Furthermore, a negative (and physically meaningless) value of loss on drying (LOD) is obtained. This can be explained because Method 1 does not allow the inclusion of physical boundaries for the variables and, as a consequence, unfeasible design outputs may be achieved sometimes (especially when extreme values are desired for some other variables). To account for this issue, as anticipated in the end of Section 4.4.1.2, the problem has been solved using the procedure described for Method 1 by substituting the direct model inversion to reconstruct $\hat{\mathbf{y}}^{\mathrm{DES}(4)}$ (Eq.(4.11) and Eq.(4.12)) with the optimization framework formalized in Eq.(4.13), which allows to bound the variables to physically sound values. In particular, the assigned value for ΔCompactability (ΔCompactability = 5) has been set as a hard constraint for $\mathbf{y}^{\mathrm{NEW}}$ in Eq.(4.13).

Using an optimization framework in this case has also another advantage. By analyzing the loading plots in Figure 4.7, it can be seen that ΔCompactability is mainly described by the first two LVs (Figure 4.7), while it does not contribute significantly on the other LVs. This means that by assigning this property, it is possible to isolate the score on LV1 or on LV2, and to replace it for the reconstruction of the other variables (see Eq.(4.8) above). The scores on the other LVs can be selected independently, thus generating an induced null space, which intersects with the existing PLS model null space (Section 4.3.1). The soft constraint on the Hotelling's T^2 proposed in Eq.(4.13) allows for the optimizer to move the solution along this null space towards the origin of the latent space so as to find a solution that satisfies the given constraints in the range of the historical data.

In Table 4.15 the solution obtained applying the procedure described for Method 1 by substituting the direct model inversion with the optimization framework in Eq.(4.13) is shown, when the equality constraint is set for ΔCompactability ($\hat{\mathbf{y}}^{\mathrm{DES}(4)}$). Physical boundaries were specified for LOD and oversize percentage which were assigned to vary between 0 and 100, while D90/D10 and the growth ratio were set to be greater than 1. The solution is presented for the cases in which the soft constraint (SC) on the Hotelling's T^2 is considered ($g = 10^{-4}$) or not ($g = 0$) in the optimization formulation.

Table 4.15. *Method 1. New sets of product properties calculated using the optimization framework in Eq.(4.13) instead of the direct model inversion in the case ΔCompactibility of the granulated product is assigned ($\hat{\mathbf{y}}^{\mathrm{DES}(4)}$), considering or not the soft constraints (SC) on T^2. The bold values in brackets represent the values assigned to ΔCompactability. T^2 and SPE statistics are also given.*

	LOD (%)	**Oversize (%)**	**ΔFlodex (mm)**	**ΔCompactability (KPa/MPa)**	**D[3,2] (μm)**	**D90/D10**	**Growth ratio**	T^2	SPE
No SC on T^2	0	0	12.4	**5 (5)**	11	36.7	7.1	11.8	0
SC on T^2	0	0	28.5	**5 (5)**	22	31.5	37.5	8.4	1.8e-6

As can be seen, both the obtained solutions satisfy the equality constraint on ΔCompactibility but are completely different from $\hat{\mathbf{y}}^{\mathrm{DES}(4)}$ in Table 4.13, in particular with respect to LOD and oversize percentage, which are found to be at their relevant boundaries. Note that the Hotelling's T^2 statistic for the solution obtained without considering the soft constraint on T^2 is above the 95% historical confidence limit represented in Figure 3. Including the soft constraint on the Hotelling's T^2 aids the optimizer to find a solution closer to the origin of the model space (and thus to the historical product profiles), and approximately lying on the model space ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} \approx 10^{-6}$, second row of Table 4.15).

Two important considerations on the solution with the soft constraint on T^2 should be remarked. First, it can be seen that the addition of the soft constraint into the objective function penalizes the minimization of $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$, which is slightly different from zero. This confirms that the induced null space is actually a *pseudo-null space*, i.e. moving the solution along it (while keeping fixed ΔCompactability) does not ensure that the solution belong to the model space, but slight deviations may occur ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} \neq 0$). As can be seen, the decrease in T^2 due to the presence of the soft constraint is limited, as the boundaries specified for some of the variables do not allow moving the solution further towards the origin of the model space.

Second, note that in general the largest differences between the two solutions reported in Table 4.15 are mainly due to the Growth Ratio and to ΔFlodex. As can be seen from the first two plots of Figure 4.7, these variables scarcely affect the first two LVs of the model, which instead are those that better describe ΔCompactability, while they are the most significant on the third LV. This means that they are the most important variables on the induced null space, namely those which undergo the highest variations by moving the solution along it.

4.4.2.2 Method 2:results

In Table 4.16 the results obtained from the application of Method 2 are shown in terms of suggested new product quality sets $\mathbf{y}^{\mathrm{NEW}}$. The procedure has been applied specifying an additional constraint for ΔCompactability, which was asked to be greater than 3 kPa/MPa (which represents a limit condition considering the available historical dataset), while LOD and oversize percentage were assigned to vary between 0 and 100, and D90/D10 and the growth ratio were set to be greater than 1 (physical boundaries). Furthermore, with reference to Section 4.4.1.3, ε was set to 10^{-6} in order to obtain a new product quality set $\mathbf{y}^{\mathrm{NEW}}$ that feasibly lies onto the model space. Results are reported for two cases: in the first case the soft constraint on the Hotelling's T^2 in Eq.(4.13) was not considered (i.e., $g = 0$), while in the second it was included ($g = 10^{-4}$). For both cases in Table 4.16, the values of $\mathbf{y}^{\mathrm{NEW}}$ which according to the procedure can be kept equal to the ones specified in the original desired set $\mathbf{y}^{\mathrm{DES}}$ are indicated with the # superscript. In the last two columns, the values of the T^2 and SPE statistics for $\mathbf{y}^{\mathrm{NEW}}$ are reported, as well.

Table 4.16. *Method 2. New set of product properties for a wet-granulated product suggested on the basis of the historical data model. The values with the # superscript represent the variable values equal to the ones in* $\mathbf{y}^{\mathrm{DES}}$. *$T^2$ and* SPE *statistics are also given.*

	LOD (%)	Oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	D[3,2] (μm)	D90/D10	Growth ratio	T^2	SPE
No SC on T^2	$3^{\#}$	$0^{\#}$	$20^{\#}$	5.4	3	42.7	$8^{\#}$	49.9	0
SC on T^2	0.1	16.1	17.6	3	75	20.5	25.1	2.1	0

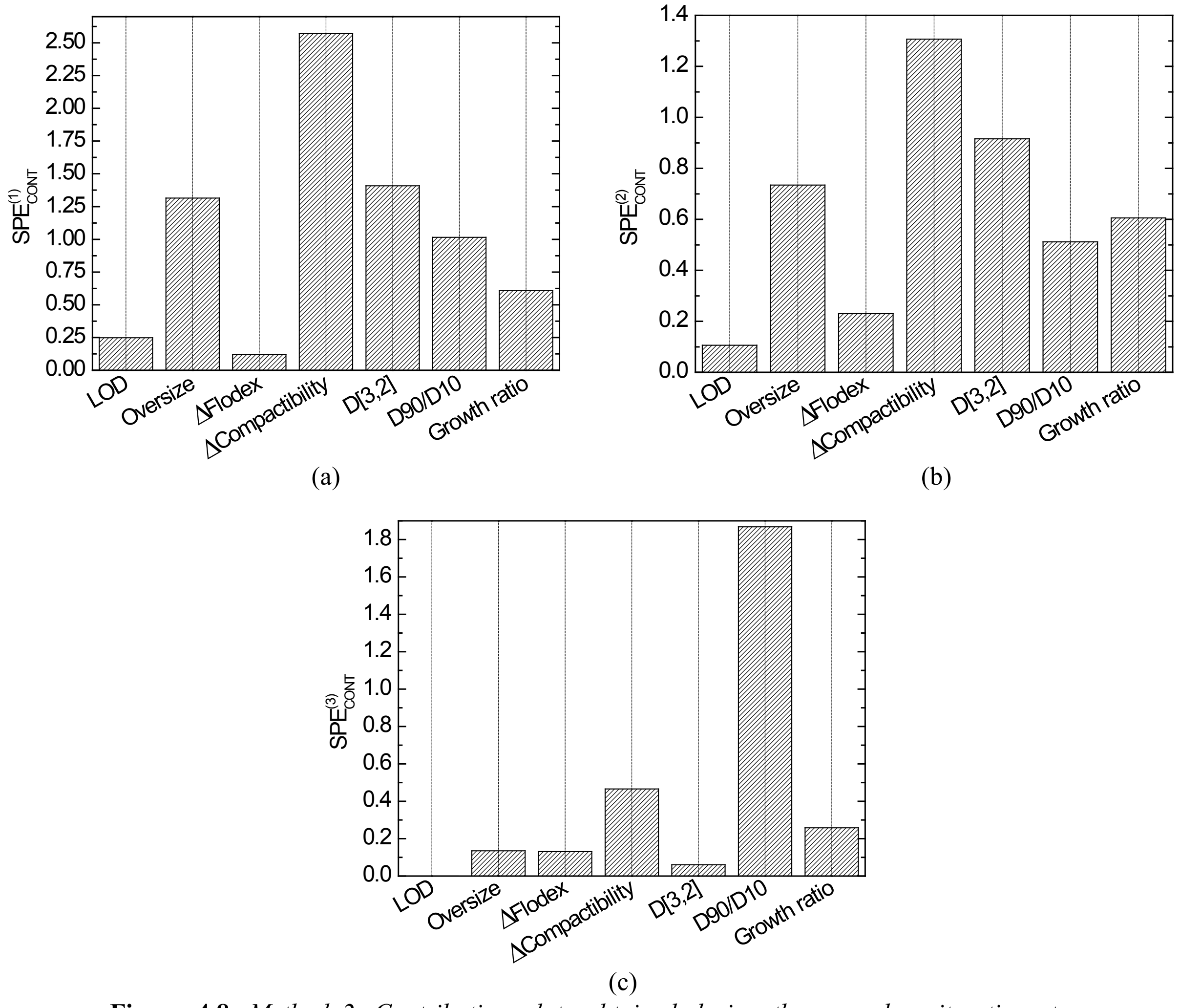
As can be seen from the first row of Table 4.16 (no soft constraint on T^2), four of the seven product properties are maintained equal to the ones specified in $\mathbf{y}^{\mathrm{DES}}$, namely LOD, the percentage of oversize granules, ΔFlodex and the growth ratio, while the solution $\mathbf{y}^{\mathrm{NEW}}$ can be considered lying onto the model space ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}} \approx 0$). This means that the user can at most keep these values equal to the corresponding ones in $\mathbf{y}^{\mathrm{DES}}$ to obtain a set $\mathbf{y}^{\mathrm{NEW}}$ which belongs to the model space. Otherwise stated, to obtain a product with the 4 indicated values assigned, the values of the other properties have to be those reported in the first row of Table 4.16 for the product to adhere to the historical product property covariance structure. Indeed $\mathbf{y}^{\mathrm{NEW}}$ represents the best tradeoff between the original target quality profile $\mathbf{y}^{\mathrm{DES}}$ and the model requirements. Note that the T^2 value (49.9) indicates that an extrapolated solution above the T^2 confidence limits is eventually obtained (Figure 4.6). As stated above (Section 4.4.1.3), this may not be a problem since, when applying LVRM inversion according to the scenarios proposed in Figure 4.2, this allows moving the solution along the null space in order to limit extrapolations in the input variable space, even if the desired product profile is out of the range of the historical products.

The procedure described in Method 2 considering a soft constraint in T^2 was also applied. Note that, as discussed in Section 4.4.1.1, an induced null space (depending on the subset of $\mathbf{y}^{\mathrm{DES}}$ being assigned) may be generated in this case, too. The second row of Table 4.16 shows the solution obtained considering the soft constraint on T^2 in Eq.(4.13). As can be seen, none of the variables keeps its value equal to those specified in $\mathbf{y}^{\mathrm{DES}}$. Moreover, note that inequality constraint for ΔCompactability is active, while the T^2 of the solution is significantly decreased, compared to the previous case (2.1 versus 49.9). In practice, in order to keep the solution onto the model hyperplane by limiting the corresponding Hotelling's T^2 and by satisfying the provided constraint, the procedure has to relax all the equality constraints on $\mathbf{y}^{\mathrm{DES}}$ by estimating a completely new product quality set $\mathbf{y}^{\mathrm{NEW}}$, in which only the inequality constraint on ΔCompactibility is satisfied.

To clarify the iterative procedure on which Method 2 is based, in Table 4.17 the estimations of the product profiles per iteration are reported together with the corresponding values of T^2 and SPE for the first case presented in Table 4.16, in which the soft constraint on T^2 is not considered. In Figure 4.8 the plots of the contributions of each variable to SPE ($\mathrm{SPE}_{\mathrm{CONT}}$) are shown for each iteration. In Table 4.17, the variables for which the corresponding constraint is relaxed at each iteration are underlined, while those which the procedure keeps equal to the ones specified in the original desired set $\mathbf{y}^{\mathrm{DES}}$ are indicated with the # superscript. From the combined analysis of the plots of Figure 4.8 and the results in Table 4.17, a deeper insight on the algorithmic procedure can be obtained.

Table 4.17. *Method 2. Estimations of the product profiles* $\mathbf{y}^{\mathrm{NEW}}$ *obtained at each iteration of the procedure described in Section 4.4.1.3.*

	LOD (%)	Oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	D[3,2] (μm)	D90/D10	Growth ratio	T^2	SPE
$\mathbf{y}^{\mathrm{NEW}(1)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$5^{\#}$	$400^{\#}$	$2.5^{\#}$	$8^{\#}$	8.8	7.3
$\mathbf{y}^{\mathrm{NEW}(2)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$\underline{3}$	$400^{\#}$	$2.5^{\#}$	$8^{\#}$	3.0	4.4
$\mathbf{y}^{\mathrm{NEW}(3)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$\underline{3}$	$\underline{29}$	$2.5^{\#}$	$8^{\#}$	6.7	2.9
$\mathbf{y}^{\mathrm{NEW}(4)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$\underline{5.4}$	$\underline{3}$	$\underline{42.7}$	$8^{\#}$	49.9	0



(a)

(b)

(c)

Figure 4.8. *Method 2. Contribution plots obtained during the procedure iterations to calculate* $\mathbf{y}^{\mathrm{NEW}}$. *(a) 1st iteration; (b) 2nd iteration; (c) 3rd iteration.*

From the projection at the first iteration of $\mathbf{y}^{\mathrm{NEW}(1)} = \mathbf{y}^{\mathrm{DES}}$ onto the $\mathbf{Q}$ loadings of the PLS model, the contributions $\mathrm{SPE}_{\mathrm{CONT}}^{(1)}$ are calculated (Figure 4.8a). It can be observed that ΔCompactability is the property most contributing to SPE. Therefore, the corresponding equality constraint is relaxed, and the optimization problem in Eq.(4.13) is solved to find the new set $\mathbf{y}^{\mathrm{NEW}}$ ($\mathbf{y}^{\mathrm{NEW}(2)}$ in Table 4.17), in which the value of ΔCompactability satisfies the specified inequality constraint for it. It results that $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(2)}} = 4.4$, which is still above the threshold ε. In Figure 4.8b the contributions to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(2)}}$ are reported ($\mathrm{SPE}_{\mathrm{CONT}}^{(2)}$). It can be noted that the highest $\mathrm{SPE}_{\mathrm{CONT}}^{(2)}$ is still due to ΔCompactability. However, the value of ΔCompactability, whose equality constraint has already been relaxed, is kept fixed by the inequality constraint. Given that this inequality constraint cannot be relaxed due to the product requirements, in the next iteration, the optimization problem in Eq.(4.13) is solved relaxing the equality constraints on both ΔCompactability and D[3,2], which is the variable with the second highest contribution to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(2)}}$. The new set $\mathbf{y}^{\mathrm{NEW}(3)}$ (Table 4.17) presents $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(3)}} = 2.9$, still above ε; $\mathrm{SPE}_{\mathrm{CONT}}^{(3)}$ (Figure 4.8c) highlights that the model mismatch is mainly due to D90/D10. Accordingly, the optimization problem is solved again, relaxing the constraint on this variable, too. Finally, solution $\mathbf{y}^{\mathrm{NEW}(4)}$ exhibits $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(4)}} \approx 0$ and thus it represents the optimal solution (first row of Table 4.16).

By analyzing the iterative solution through Figure 4.8 and the values of SPE at each iteration, it can be noted that in this case a solution inside the 95% SPE confidence limit (Figure 4.6) is obtained simply relaxing the equality constraint on ΔCompactibility ($\mathbf{y}^{\mathrm{NEW}(2)}$ in Table 4.17). The procedure could therefore have stopped at the first iteration. However, to allow for the inclusion of the hard constraints for $\hat{\mathbf{y}}^{\mathrm{DES}}$ in the LVRM inversion procedures (namely $\hat{\mathbf{y}}^{\mathrm{DES}} = \mathbf{Qt}$), the methods were asked to find a solution very close to the model space ($\varepsilon \to 0$). Note that in general Method 1 and Method 2 return different information. The first method provides a general perspective on how the assignment of a product specification affects the other variables according to the historical knowledge. The second method provides an optimal solution as a tradeoff between the desired product quality variables and the need to fulfill the relationships between product variables obtained from the historical data and represented by the $\mathbf{Q}$ loadings of the PLS model. However, it must be noted that Method 1 is more susceptible to uncertainty due to the calculation of the inverse of the $\mathbf{Q}^{(i)\mathrm{T}}\mathbf{Q}^{(i)}$ matrix. Depending on the variable which is assigned for $\hat{\mathbf{y}}^{\mathrm{DES}}$ or on the number of LVs selected to build the PLS model, matrix $\mathbf{Q}^{(i)\mathrm{T}}\mathbf{Q}^{(i)}$ may be ill-conditioned due to variable correlation. The ill-conditioning and the presence of noise in the measurements, which masks the effective rank of the $\mathbf{Y}$ space, could result in poor estimations of the solution scores using Eq.(4.11). For this reason, as shown in Section 4.4.2.1, Method 1 can be applied by substituting the direct inversion of the model with an optimization framework, which allows moving the solution along the (induced) null space to find the minimum Mahalanobis distance (i.e. with minimum Hotelling's T^2) solution (García-Muñoz *et al.*, 2006 and 2008).

Finally consider that it may occur that none of the Methods returns a solution if the design of a product with very different properties from the ones in historical dataset is required.

4.5 Conclusions

In this Chapter a general framework to perform LVRM inversion has been proposed, in which the most appropriate problem is solved depending on the objectives and the constraints the user may have in the product/process development activity. The framework identifies 4 different inversion approaches depending on the type of constraints the desired product quality variables have to fulfill, which can be equality and/or inequality constraints. Each approach gives rise to a different constrained optimization problem. Namely, if no constraints exist for the input variables, the optimization is performed in the LVRM space and the solution will lie in that space. If constraints are specified for the input variables as well, the optimization procedure may be forced to find an extrapolated solution (namely a solution that does not belong to the LVRM space) to satisfy the given constraints. The proposed methodology has been successfully applied in a particle engineering problem for the design of the raw material properties in a high-shear wet granulation process, to obtain granules with specified quality characteristics in output. Three design cases, with different problem constraints and objectives, have been presented and discussed.

The null space concept has been further investigated and it has been shown how it can be employed in the definition of the design space of a process under a QbD framework. Since the null space represents a subspace of the reduced LVRM space, there is the need to translate it into ranges for the design variables. Bi-dimensional plots have been used to identify the design variable combinations whose projections belong to the null space.

Some important issues have also been addressed. First, it has been emphasized that the model should not only provide accurate enough predictions of the response, but it should also provide a reliable reconstruction for the regressors (especially for model inversion). To aid the component selection a new metric has been proposed (the P^2 statistic). This metric can be obtained in a similar way the Q^2 is obtained in a cross-validation exercise but removing and reconstructing the elements from the regressor set.

Secondly, the goodness of the results obtained from an LVRM inversion exercise is affected by the goodness of the model in describing the empirical data used to build it. Not every input variable and response datasets can be used for LVRM building for inversion. In general, if the regressor dataset does not contain enough information to describe the variability in the response dataset, the lack of appropriate response variable fitting, which is quantified by the percentage of unexplained variance, will propagate in the inversion, increasing the uncertainties in the estimation of the solution. The robustness of the model inversion results strongly depends on these uncertainties, which affect the model parameter estimation. To account for them, a strategy based on a jackknife approach has been proposed to estimate the confidence limits in the calculation of the null space and of the optimization solution. These provide a metric to understand the reliability of the solution.

Thirdly, it must be noted that under certain circumstances the results from the LVRM inversion may not give acceptable solutions in the first iteration (the obtained quality may not be exactly on target). Under this occurrence, it would be required to refit the model including the results from the experiment carried out based on the first attempt model inversion solution, and then re-run the inversion. By iteratively repeating this procedure (modeling, inversion, experimentation), the solution obtained by inversion will converge to the desired product target profile.

Finally, note that the basic assumption of the implemented LVRM inversion procedures is that the model is able to represent the desired quality for the product to be designed, to an acceptable level of uncertainty. In particular, in the case studies presented for LVRM inversion, it has been assumed that the desired quality adhered to the historical product covariance structure (i.e. $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} = 0$). This could be achieved by reconstructing $\mathbf{y}^{\mathrm{DES}}$ according to the PCA model on **Y** or on the basis of the **Q** loadings of the PLS model. The different way of reconstruction changes the constraints for $\mathbf{y}^{\mathrm{DES}}$ in the optimization problems from soft to hard. However, this reconstruction does not ensure that the desired output variables $\mathbf{y}^{\mathrm{DES}}$ are obtained. A methodology has therefore been proposed to exploit the covariance structure of the historical data, in order to guide the selection of a target attribute profile ($\mathbf{y}^{\mathrm{DES}}$), as a tradeoff between the model structure and the desired product characteristics.

Two different procedures have been presented to this purpose. In the first one, elements of $\mathbf{y}^{\mathrm{DES}}$ are assigned one at a time, and an approach based on model inversion of the model is used to estimate the other variables. In the second procedure, an algorithm is used to iteratively select and relax the constraints of the variables that are found to be most responsible for the model mismatch. An optimization problem is solved to calculate their values according to the loadings of the **Y** space. It has been demonstrated how in both these methods an induced null space may be generated depending on the assigned variables, in which different solutions, all satisfying the given constraints, can be found.

The proposed approaches have shown their effectiveness in assessing the feasibility of a new product, and in suggesting a reliable and physically sound product quality profile for the considered wet granulation case study. These methods allow to not set any constraints in the LVRM inversion problem on the mismatch in representing the desired quality (i.e. on $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$) and to reduce the number of iterations in the application of the LVRM inversion procedure. In particular, the inclusion in the framework of constraints on the model mismatch for $\mathbf{y}^{\mathrm{DES}}$ would imply to estimate and decompose the error in its different contributions (since these are typically well understood), namely the model mismatch, the measurement uncertainties and any sample bias. These contributions to the quality variable uncertainties are usually sample-dependent and not easily identifiable. For these reasons, the way to handle them in the model inversion problem still remains an open issue for future research.

Chapter 5

Latent variable model inversion for *in silico* product formulation[*]

In this Chapter, a methodology based on latent variable regression model (LVRM) inversion is proposed to aid the design of the formulation for pharmaceutical products, namely the selection of the best excipient types and amounts to mix with a given active pharmaceutical ingredient (API). The general framework presented in Chapter 4 is here extended to consider also constraints on the excipient selection and to account for different objectives the formulation problem may have (e.g., API dose maximization). The procedure is tested on an industrial case study to design the formulation for a proprietary API. Results obtained *in silico* are validated experimentally, demonstrating the effectiveness of the proposed methodology. A user-friendly interface has been developed to allow formulators to apply the proposed methodology, by assigning the desired objectives and constraints through dropdown menus.

5.1 Introduction

In the development of a drug product, the first important decision to take is the choice of the formulation, namely the selection of the appropriate excipients that are to be mixed with an active pharmaceutical ingredient (API) into the final drug product. This choice is driven by constraints mainly related to the safety, the efficacy but also the processability of the drug product (Hamad *et al.*, 2010).

In general, if a model describing the mixture properties from the properties of the excipients and the APIs involved in the formulation were available, it could be used to aid the selection of the best materials that ensure to obtain a mixture of desired properties. To this end, the model itself can act as a constraint to an optimization problem aiming at maximizing/minimizing an objective function, which may represent a product performance index or a cost function quantifying the difference between model predictions and desired product quality. The optimization will then give as outputs the estimates of the best raw

[*] Tomba E., M. Barolo and S. García-Muñoz (2013). *In-silico* product formulation design through latent variable model inversion. *In preparation*.

material types and amounts to be used in order to achieve the desired mixture (or blend) properties (Smith and Ierapetritou, 2010).

The use of deterministic models to describe these relationships would always be desirable, as deterministic models give a transparent representation of the physical phenomena acting on the system, explaining them from first principles. In the field of mixture modeling, some examples have been reported for the semiconductor industry (Kumar, 2003) and for polymeric blend design (Bernardo *et al.*, 1996). The development of such models requires a detailed knowledge and understanding of the interactions between the materials entering the formulation and of their implications for the product properties. This may be burdensome to achieve in pharmaceutical product design, due to the variety of products and raw materials of different physical/chemical characteristics that may enter in a drug product formulation (e.g., APIs, fillers, binders, disintegrants, lubricants), which can be difficult to manage in a deterministic modeling framework. For these reasons, formulators have often resorted to building models based on data obtained either from targeted experiments or from historical manufacturing databases.

Mixture design of experiments (DoE) and response surface modeling techniques (Montgomery, 2005a) have been among the first systematic multivariate approaches used to assist the design of the formulations of pharmaceutical products (Campisi *et al.*, 1998). Due to the nature of the pharmaceutical formulations, the use of DoE techniques in the early stages of pharmaceutical development often requires performing a large number of experiments, as a large number of candidate materials and permutations have to be considered. To address this issue, multivariate design of experiments has been proposed (Wold *et al.*, 1986). This approach combines DoE techniques with the multivariate analysis of databases of the available raw materials (e.g., through PCA), to select the most suitable for the experimental design. As an example, Gabrielsson *et al.* (2003) used a multivariate design to evaluate in a systematic way a large number of candidate excipients to include in a formulation, based on their similarity assessed through a PCA of their characterization data. Latent variable regression models (LVRMs) based on PLS were then built on the data obtained from the experiments and used to test new formulations in order to obtain a product of desired disintegration time and crushing strength (Gabrielsson *et al.*, 2004).

Indeed, one of the important characteristics a data-based model should have to support product formulation design, is the ability in accurately predicting the mixture properties based on the raw materials data. Many efforts have been produced by researchers in finding methods able to give models with good prediction performances, often regardless of the model structure. For example, several contributions on the use of black box models to support the design of pharmaceutical formulations have been proposed (Rowe and Roberts, 1998). Learning techniques like neural networks (Agatonovic-Kustrin and Beresford, 2000; Takayama *et al.*, 2003; Sun *et al.*, 2003), neuro-fuzzy logic (Abraham *et al.*, 2007; Landín *et* *al.*, 2009), genetic programming (Barmpalexis *et al.*, 2011) and expert systems (Shao *et al.*, 2007) have been applied to build models based on historical or DoE formulation data, and then used in an optimization framework to suggest the experiments to be performed for the design of the product formulation. Despite these tools are very attractive and the results are promising, most black box models suffer from the lack of transparency in understanding the mechanisms behind predictions. Model design is essentially based on the accuracy of predictions rather than on the optimization of the model complexity. As a consequence, black box models parameters are often difficult to interpret and may not provide a scientific understanding of the system being modeled. This conflicts with the QbD requirements, for which the ability to predict has to reflect a high degree of process understanding (FDA, 2004c).

Conversely, multivariate statistical methods like LVRMs combine the accuracy in prediction to a scientifically-sound understanding and interpretation of both model building procedure and model parameters, as was also shown in Chapter 3. In most of the contributions that used LVRMs to model mixture data, the model is however limited to describe the relationships between a matrix of regressors, including only the fractions of the raw materials in the tested formulations, and a response matrix, including the properties of the mixture. This type of analysis does not properly include in the model the physical/chemical properties of all the raw materials that could possibly be included in the final formulation. Muteki and MacGregor (2007b) addressed this issue by proposing an LV method called L-shaped PLS (LPLS), which includes in a unique LVRM framework the database of the physical/chemical characterization of all the available raw materials, the database on the fractions of each raw material in the historical formulations, and the properties of the obtained mixture. This modeling approach was then combined with LVM inversion techniques to successfully support the development of a new blend of rubbers (Muteki *et al.*, 2006).

Recently, García-Muñoz and Polizzi (2012) recognized that, despite its effectiveness, the approach proposed by Muteki and MacGregor (2007b) did not consider those situations in which the desired product is obtained by mixing materials of different nature or which underwent different characterization procedures, as typical for example of APIs and excipients in pharmaceutical formulations. To address this issue, the authors proposed a new approach in which they considered in a unique modeling framework datasets referring to materials of different nature (e.g., APIs and excipients) with the historical formulation and the properties data of the obtained mixtures. This method, called weighted-scores PLS (WSPLS; García-Muñoz and Polizzi, 2012), was applied for the prediction of the particle, powder and compact mechanical properties of pharmaceutical blends, starting from the raw materials properties and amounts, without resorting to extensive experimentation (Polizzi and García-Muñoz, 2011).

In this Chapter, the LVRM strategy proposed by Polizzi and García-Muñoz (2011) (called blend prediction model) is used to perform *in-silico* a new drug product formulation design. The strategy is based on an optimization framework for LVRM inversion (based on the one proposed in Chapter 4; Figure 4.2) which, starting from a given API, returns as output the set of excipient type and amounts most suitable to reach a blend of desired powder, flow and mechanical properties. Since the formulation design problem may have different objectives and constraints (e.g., the maximization of the API dosage, the minimization of the tablet weight, the choice of excipients of a given family), the strategy is developed in such a way as to allow the user to specify several different types of constraints, both on the input variables (e.g., excipients families, type and ratio; the API dose) and on the output variables (the blend properties). The effectiveness of the proposed approach is tested experimentally in the development of a pharmaceutical blend for direct compression for an in-house API.

5.2 Case study and methodology

5.2.1 Available data

The datasets used in this study are similar to those previously described in the work of Polizzi and García-Muñoz (2011). They were obtained from a wide database of physical properties of pharmaceutical powders, which includes individual excipients, APIs, formulated blends and granulations. As described in the original work, data on excipients (e.g., binders, fillers, disintegrants, lubricants), APIs and historical formulated blends were collected from the database and gathered in four different datasets, which were later used to build the model:

- a dataset $\mathbf{X}^{\mathrm{EXC}}$, including 45 physical properties and particle size distribution (PSD) measurements for 64 excipients (including disintegrants);
- a dataset $\mathbf{X}^{\mathrm{API}}$, including 51 physical properties and PSD measurements for 118 APIs.
- a dataset $\mathbf{R}$ including the fractions of each considered raw material in 24 historical blended formulations (both active and inactive). This matrix can be seen as formed by and $\mathbf{R}^{\mathrm{EXC}}$ matrix of the fractions of each excipient, and a $\mathbf{R}^{\mathrm{API}}$ matrix of the fractions of each API in each individual formulation ($\mathbf{R} = \left[\mathbf{R}^{\mathrm{EXC}}\ \mathbf{R}^{\mathrm{API}}\right]$).
- a vector $\mathbf{r}^{\mathrm{MgSt}}$ of the magnesium stearate fractions in the 24 historical formulations. The reason why the magnesium stearate fractions (that is a lubricant) are not considered in $\mathbf{R}$ will be explained later on.
- a dataset $\mathbf{Y}$, including 51 physical properties and the PSD measurements for the 24 historical blends.

These data are representative of the range of materials typically seen in solid dosage form development. The interested reader is referred to the original manuscript of Polizzi and García-Muñoz (2011) for the complete list of the measured material physical properties and for details on the measurement procedures. It is important to note that not all the same properties have been measured for each raw material. Furthermore, even if some physical measured properties were the same for different materials (e.g., the PSD regimes), the testing method may not have been the same. In fact, all the APIs were tested using smaller (0.5 g, $3/8'' \times 3/8'' \times 3/16''$) compacts compared to the excipients (5 g, $3/4'' \times 3/4'' \times 3/8''$), thus causing some differences between the correlations among properties within the two distinct populations (Polizzi and García-Muñoz, 2011). This is the reason for which the WSPLS approach (García-Muñoz and Polizzi, 2012) has been proposed to model these data.

Table 5.1. *Families in which the excipients in the database are divided and number of excipients per family.*

Excipient families	**# of excipients per family**
Function	
Alkalizing	2
Binder	2
Buffering	1
Disintegrant	5
Filler	51
Gum	1
Lubricant	1
Orally Disintegrating Tablet disintegrant (ODT)	2
Stabilizer	1
Mechanical behavior	
Brittle	30
Duttile	27
Main compound	
Calcium Carbonate (CaCo3)	4
Calcium	7
Calcium Phosphate (CaPho4)	2
Calcium Silicate	1
Hydrophilic Polymer Matrix (HPM)	4
Hydroxy-Propyl-Methyl-Cellulose (HPMC)	3
Lactose	14
Lactose Anhydrous	11
Lactose Hydrous	3
Microcrystalline Cellulose (MCC)	8
Polyethylene Oxide	2
PolyVinylPyrrolidone	1
Sodium	6
Starches	2
Sugars	13

To assist the formulation design problem, the materials included in the excipient database $\mathbf{X}^{\mathrm{EXC}}$ have been divided in "families" according to their function in the solid mixture, their mechanical behavior and their main chemical compound. In Table 5.1, the list of the 26 identified excipient families is reported together with the number of excipients per family. In general, families representing the function of an excipient in the blend (e.g., filler, disintegrant) include families of materials with different mechanical behavior (e.g., brittle, ductile), which in turn include families of materials based on different compounds (e.g., MCC, lactose). Accordingly, an excipient may belong to more than one family.

This material categorization in families is not unique; nevertheless it is useful in the formulation design problem, as it helps to identify the materials that need to be considered for the formulation. The constraints and the level of knowledge the formulators may have, drive the selection of the families of excipients that can in turn narrow more or less significantly the range of candidate materials among which searching the most suitable ones for the desired formulation. This idea will be exploited later, in the implementation of the blend prediction model inversion for the *in-silico* formulation design.

5.2.2 Blend prediction model

The datasets described in the previous section were related through the WSPLS method (García-Muñoz and Polizzi, 2012), which allows to include mixture data involving materials of different nature and with different characteristics in a whole modeling framework. This results in a blend property prediction model, in which the relationships between the raw material properties and the final blend properties are mapped and understood.

The application of the WSPLS technique (García-Muñoz and Polizzi, 2012) to build the blend prediction model is exemplified in Figure 5.1. The procedure goes through 3 main steps. First, two separate PCA models are built on the API ($\mathbf{X}^{\mathrm{API}}$) and excipient ($\mathbf{X}^{\mathrm{EXC}}$) datasets, with the aim of describing the complex network of relationships (the correlation structure) between the API and the excipient properties using a few interpretable principal components (PCs). This allowed all materials to be positioned in a multivariate "design space" (Polizzi and García-Muñoz, 2011). In this case $A = 6$ PCs have been found to be significant for each model. As mentioned above, the use of separate models is justified by the different correlation structure found among measured properties due to the difference in variables or to different testing methods employed. In the second step, the PCA scores of the excipients ($\mathbf{T}^{\mathrm{EXC}}$) used in the 24 historical formulations are weighted according to their fractions in the formulations ($\mathbf{R}^{\mathrm{EXC}}$). A matrix $\mathbf{T}_{\mathrm{W}}^{\mathrm{EXC}}$ is obtained, whose rows are the sums of the weighted scores of each excipient entering each formulation. The same operation is performed for the API scores ($\mathbf{T}^{\mathrm{API}}$), by weighting them through their fractions in the historical formulations ($\mathbf{R}^{\mathrm{API}}$), thus giving the matrix $\mathbf{T}_{\mathrm{W}}^{\mathrm{API}}$ of the sums of the weighted API scores for each formulation.

Eventually, the weighted scores matrices $\mathbf{T}_{\mathrm{W}}^{\mathrm{EXC}}$ and $\mathbf{T}_{\mathrm{W}}^{\mathrm{API}}$ are concatenated each other and with the vector $\mathbf{r}^{\mathrm{MgSt}}$ of the magnesium stearate fractions in the historical formulations ($\mathbf{T}_{\mathrm{W}} = \left[\mathbf{T}_{\mathrm{W}}^{\mathrm{EXC}}\ \mathbf{T}_{\mathrm{W}}^{\mathrm{API}}\ \mathbf{r}^{\mathrm{MgSt}}\right]$). The magnesium stearate fractions have been added in this case to account for the effect of the lubricant amount on the blend properties. Since this was the only lubricant used in the historical formulations, its physical properties have not been considered in $\mathbf{R}$ for the model building. Eventually, in the last step of the procedure, matrix $\mathbf{T}_\mathrm{W}$ is related to the $\mathbf{Y}$ dataset of the historical blended formulation properties through a PLS model:

$$\mathbf{T}_\mathrm{W} = \mathbf{T}\mathbf{P}^\mathrm{T} + \mathbf{E}_\mathbf{X} \tag{5.1}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^\mathrm{T} + \mathbf{E}_\mathbf{Y} \tag{5.2}$$

$$\mathbf{T}^\mathrm{PLS} = \mathbf{T}_\mathrm{W}\mathbf{W}^* \quad , \tag{5.3}$$

where the regressor matrix is the weighted scores matrix $\mathbf{T}_\mathrm{W}$, while the meaning of the other symbols is the same as for a typical PLS model (Chapter 2, Section 2.1.2). For the PLS model design 6 LVs have been considered.

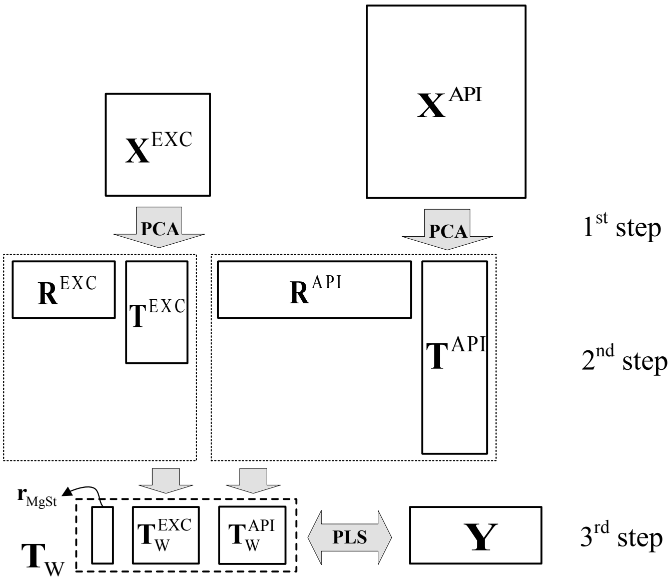


Figure 5.1. *Schematic of the Weighted-Scores PLS procedure applied to the blend prediction model design.*

5.2.2 Blend prediction model inversion

The blend prediction model described in the previous section can be used in a *forward* way to predict the properties of the blended mixture once the materials and their fractions (i.e. the product formulation) are known. However, in a pharmaceutical formulation design case study the interest is in determining the most suitable formulation for one or more given APIs in order to obtain a blend of desired properties. To solve this problem, a procedure to invert the blend prediction model is proposed. The rationale behind this procedure is shown in Figure 5.2: once a set $\mathbf{y}^\mathrm{DES}$ of desired blend properties has been established and the APIs have been specified, the PLS model linking the blend properties with the weighted scores of the input materials properties (3rd step in Figure 5.1) is inverted to find the weighted score vector $\mathbf{t}_\mathrm{W} = \left[\mathbf{t}_\mathrm{W}^\mathrm{API}\,\mathbf{t}_\mathrm{W}^\mathrm{EXC}\,r^\mathrm{MgSt}\right]$. In $\mathbf{t}_\mathrm{W}$, $\mathbf{t}_\mathrm{W}^\mathrm{API}$ includes the elements of $\mathbf{t}_\mathrm{W}$ related to the API, whereas

$\mathbf{t}_{\mathrm{W}}^{\mathrm{EXC}}$ includes the elements of $\mathbf{t}_{\mathrm{W}}$ related to the excipients. From $\mathbf{t}_{\mathrm{W}}$, the magnesium stearate fraction r^{MgSt} can be directly determined (if not constrained). The weighting scores operation of the blend prediction model (2nd step in Figure 5.1) is then inverted to determine the API fraction(s) ($\mathbf{r}^{\mathrm{API}}$), and the excipient types and fractions ($\mathbf{r}^{\mathrm{EXC}}$) to be mixed with the selected API/s in order to give the weighted score vector $\mathbf{t}_{\mathrm{W}}$. The selected excipient types and their fractions are represented by the gray boxes in Figure 5.2.

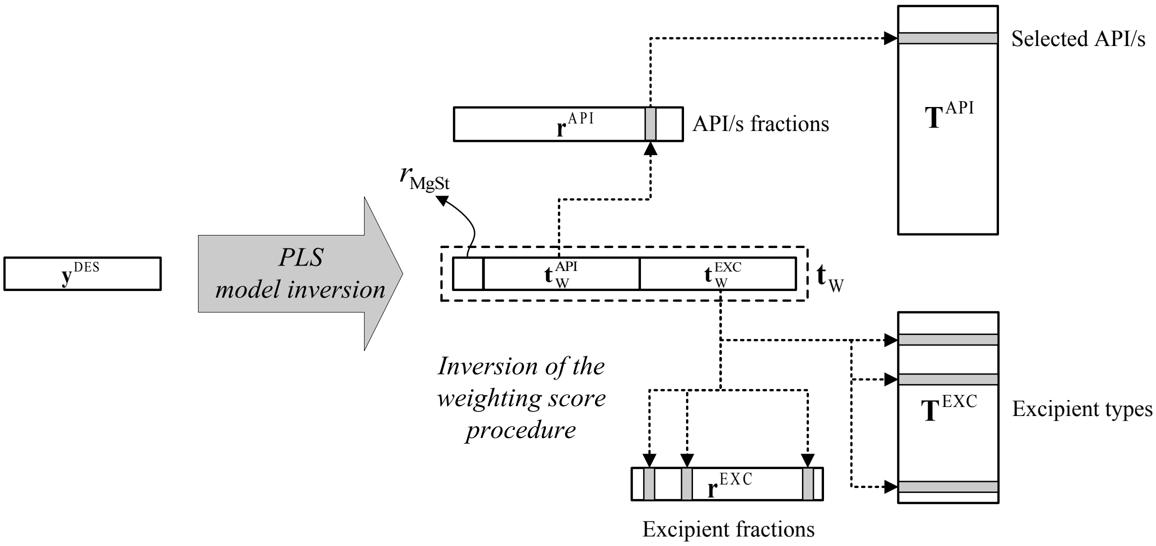

Selected API/s
$\mathbf{T}^{\mathrm{API}}$
$\mathbf{r}^{\mathrm{API}}$
API/s fractions
r_{MgSt}
$\mathbf{y}^{\mathrm{DES}}$
PLS model inversion
$\mathbf{t}_{\mathrm{W}}^{\mathrm{API}}$
$\mathbf{t}_{\mathrm{W}}^{\mathrm{EXC}}$
$\mathbf{t}_{\mathrm{W}}$
Inversion of the weighting score procedure
$\mathbf{T}^{\mathrm{EXC}}$
Excipient types
$\mathbf{r}^{\mathrm{EXC}}$
Excipient fractions

Figure 5.2. *Schematic of the blend prediction model inversion procedure.*

The blend prediction model inversion represented in Figure 5.2 would then require to perform a PLS model inversion and the weighting score operation inversion. As seen in Chapter 2 (Section 2.2), from a mathematical point of view, the PLS model inversion could have infinite solutions (Jaeckle and MacGregor, 1998). Furthermore, there may exist an infinite number of combinations of raw materials and their fractions that can give the same weighted score vector $\mathbf{t}_{\mathrm{W}}$. An optimization approach is therefore needed to find the optimal combination of raw material types and fractions that satisfy the constraints provided for the blend properties, while adhering to the historical data covariance structure represented by the PLS model.

Due to the large number of involved materials, the domain inside which the optimizer should search the optimal formulation can be very wide and may lead to unrealistic solutions if the problem is not correctly stated. Therefore, the following constraints have been considered for the implementation of the optimization framework; these can be easily adapted and modified as needed:

1. the desired blend formulation is assumed to be made by one API (assigned by the user), two excipients of different families and one disintegrant, in addition to magnesium stearate as a lubricant. This reflects the composition of the historical formulations used to build the blend prediction model and is typical of direct compression blends.

2. Even if the blend prediction model returns the predictions for all the 51 blend physical properties considered in the model building phase, the interest in this case study is that the final blend meets the desired requirements for only a few core properties, which are considered most important in the formulation of blends for direct compression. These core properties, forming the set $\mathbf{y}^{\mathrm{DES}}$ of the desired properties used for the model inversion (Figure 5.2), include $\mathrm{B_{MID}}80$ (which is a measure of the width of the blend PSD; Katdare and Chaubal, 2006), compression stress, tensile strength, brittle fracture index (BFI) and the flow function coefficient (FFC). Design specifications (i.e. the constraints for the model inversion optimization problem) are therefore assigned for each of these properties.
3. Given that excipients in the database are divided in "families", only one excipient per family can be selected by the procedure. This requires to implement some logical constraints to discard all the excipients in a family, once an excipient of that family has been selected.
4. The inversion solution in terms of selected excipient types and fractions must satisfy the overall mass balance. This means that:

$$\sum_{i=1}^{NEXC} r^{\mathrm{EXC}(i)} + r^{\mathrm{API}} + r^{\mathrm{MgSt}} = 1 \quad , \tag{5.4}$$

where *NEXC* is the number of excipients included in the formulation, while $r^{\mathrm{EXC}(i)}$ is the fraction of the *i*-th excipient in the formulation. Namely, the sum of the material fractions in the formulation must be equal to one.

On the basis of the above-mentioned remarks, an optimization framework to perform the blend prediction model inversion for formulation design has been implemented. The framework combines in a unique optimization problem the LVRM inversion strategies described in Chapter 4 (Section 4.2), with discrete variable constraints and mass balance equations, giving rise to a mixed-integer nonlinear programming (MINLP) problem (Quesada and Grossman, 1992), whose formal statement is discussed in Section 5.2.2.1. In addition to the formulation design, the framework enables to cope with a wide range of applications and constraints a formulator may encounter in a new formulation design problem; for example, it allows to determine or maximize *in-silico* the API dose in the new product formulation, or to minimize the blend total weight (with the aim of reducing the material consumption) or the final tablet weight. Therefore, in addition to the API that needs to be formulated, the following constraints need to be specified for the optimizer to find a solution:

- The API dose, which can be assigned (equality constraint) or calculated by the procedure. In the latter case, a lower bound for the dose can be provided, or the procedure can determine the formulation with the maximum API dose.

- The blend (or tablet) total weight, which can be assigned (equality constraint) or calculated by the procedure. In the latter case, an upper bound (inequality constraint) can be provided for it or the procedure can minimize it.
- The families within which the optimal excipients for the formulation (including disintegrants) should be searched. For each excipient family, upper boundaries (inequality constraints) can be set for the fraction of the selected excipient in the final blend. Moreover, initial guesses for each excipient type and fraction can be specified to aid the optimizer finding an optimal solution. If the selection of one or more excipients in the final formulation is fixed (e.g., for safety, processability or availability reasons), the mandatory excipient can be set as an equality constraint.
- The magnesium stearate fraction in the final blended formulation r^{MgSt} (equality constraint).
- The desired values for the (five) final blend properties. These can be assigned (equality constraints) or allowed to vary in assigned ranges (inequality constraints).

In output, the framework returns the *in-silico* designed formulation for the given API, which, according to the model, ensures the achievement of the desired blend properties, together with their predictions and the calculated API dose and tablet weight (if not constrained).

5.2.2.1 Problem statement

The procedure for the blend prediction model inversion is based on the following MINLP formalization (Quesada and Grossmann, 1992). The objective function of the problem is:

$$\min_{\mathbf{t}}\left[\left(\hat{\mathbf{y}}^{\mathrm{NEW}}-\mathbf{y}^{\mathrm{DES}}\right)^{\mathrm{T}}\mathbf{\Gamma}\left(\hat{\mathbf{y}}^{\mathrm{NEW}}-\mathbf{y}^{\mathrm{DES}}\right)+g_1\cdot T^2+g_2\cdot \mathrm{SPE}_{\mathbf{t}_{\mathrm{W}}}+g_4\cdot m^{\mathrm{API}}+g_5\cdot m^{\mathrm{TOT}}\right] \quad .(5.5)$$

It is composed by different terms, according to the formulation problem that has to be solved. The optimization variable is represented by the score vector $\mathbf{t}$ in the historical formulation latent space (i.e., in the latent space of the PLS part of the blend prediction model; Figure 5.1). The first three terms of the objective function are the same terms of the objective function for Scenario 4 of the general framework presented in Chapter 4 (Eq.(4.4)), where $\mathbf{t}_{\mathrm{W}}$ (the weighted score vector) represents the regressor set to be determined through the inversion. The fourth term represents the API dose (m_{API}), and the fifth one the tablet weight (m_{TOT}). These terms are weighted according to g_3 and g_4, and they appear in the objective function only if the optimization procedure is used to maximize the API dose in the formulation (for which $g_4<0$) or minimize the final tablet weight. Otherwise $g_4=0$ and $g_5=0$.

The optimization problem is subject to several other constraints, which can be summarized as follows.

In Eqs.(5.6)-(5.8) the equations inherent to the PLS model part of the blend prediction model (see Figure 5.1) are reported. The meaning of the symbols is the same as for Eqs.(5.1)-(5.3).

$$\hat{\mathbf{y}}^{\mathrm{NEW}} = \mathbf{Qt} \tag{5.6}$$

$$\mathbf{t}_{\mathrm{W}} = \mathbf{Pt} \tag{5.7}$$

$$\mathbf{t} = \mathbf{W}^{*^{\mathrm{T}}} \mathbf{t}_{\mathrm{W}} \tag{5.8}$$

In Eqs.(5.9)-(5.10) the equations and constraints of the solution statistics T^2 and $\mathrm{SPE}_{\mathbf{t}_{\mathrm{W}}}$ are reported. In Eq.(5.9), t_a is the a-th element of the score set $\mathbf{t}$, which is composed by A elements corresponding to the number of LVs used to build the PLS model; $s^2_{\mathbf{T}_a^{\mathrm{PLS}}}$ is the variance of the a-th column of the historical formulation score matrix ($\mathbf{T}$). In Eq.(5.10), $\hat{\mathbf{t}}_{\mathrm{W}}$ is the reconstruction of the solution weighted scores through the model. As in Scenario 3 and Scenario 4 of the framework proposed in Chapter 4 (Eq.(4.3) and Eq.(4.4)), an inequality constraint is assigned to $\mathrm{SPE}_{\mathbf{t}_{\mathrm{W}}}$, which has to be lower than its 95% confidence limit ($\mathrm{SPE}_{\mathbf{T}_{\mathrm{W}},95\%\mathrm{lim}}$), properly lowered by weight g_3.

$$T^2 = \sum_{a=1}^{A} \frac{t_a^2}{s^2_{\mathbf{T}_a^{\mathrm{PLS}}}} \tag{5.9}$$

$$\mathrm{SPE}_{\mathbf{t}_{\mathrm{W}}} = \left(\hat{\mathbf{t}}_{\mathrm{W}} - \mathbf{t}_{\mathrm{W}}\right)^{\mathrm{T}} \left(\hat{\mathbf{t}}_{\mathrm{W}} - \mathbf{t}_{\mathrm{W}}\right) \tag{5.10}$$

$$\mathrm{SPE}_{\mathbf{t}_{\mathrm{W}}} \le g_3 \cdot \mathrm{SPE}_{\mathbf{T}_{\mathrm{W}},95\%\,\mathrm{lim}} \tag{5.11}$$

In Eqs.(5.12)-(5.14) the equations for the calculation of the regressor set $\mathbf{t}_{\mathrm{W}}$ (the weighted score section of the blend prediction model, as represented in Figure 5.1) are reported. $\mathbf{t}_{\mathrm{W}}$ (Eq.(5.12)) has already been described in Section 5.2.2. In Eq.(5.13), $t^{\mathrm{API}}_{\mathrm{W},k}$ is the k-th element of $\mathbf{t}^{\mathrm{API}}_{\mathrm{W}}$, $r_{\mathrm{V}}^{\mathrm{API}(i)}$ is the volumetric fraction of the i-th API in the formulation while $t_{\mathrm{PCA}}^{\mathrm{API}(i,k)}$ is the k-th score for the i-th API in the PCA model built on the API database, being $\mathrm{PC}^{\mathrm{API}}$ the number of PCs considered in the model and *NAPI* the total number of APIs in the historical database. In Eq.(5.14), $t^{\mathrm{EXC}}_{\mathrm{W},t}$ is the t-th element of $\mathbf{t}^{\mathrm{EXC}}_{\mathrm{W}}$, $r_{\mathrm{V}}^{\mathrm{EXC}(s)}$ is the volumetric fraction of the s-th excipient in the formulation, while $t_{\mathrm{PCA}}^{\mathrm{EXC}(s,t)}$ is the t-th score for the s-th excipient in the PCA model built on the excipient database, being $\mathrm{PC}^{\mathrm{EXC}}$ the number of PCs considered in the model and *NEXC* the total number of excipients in the historical database (including disintegrants).

$$\mathbf{t}_{\mathrm{W}} = \left[\mathbf{t}^{\mathrm{API}}_{\mathrm{W}} \; \mathbf{t}^{\mathrm{EXC}}_{\mathrm{W}} \; r_{\mathrm{MgSt}}\right] \tag{5.12}$$

$$t^{\mathrm{API}}_{\mathrm{W},k} = \sum_{i=1}^{NAPI} r_{\mathrm{V}}^{\mathrm{API}(i)} \cdot t_{\mathrm{PCA}}^{\mathrm{API}(i,k)} \qquad k = 1,\ldots,\mathrm{PC}^{\mathrm{API}} \tag{5.13}$$

$$t^{\mathrm{EXC}}_{\mathrm{W},t} = \sum_{s=1}^{NEXC} r_{\mathrm{V}}^{\mathrm{EXC}(s)} \cdot t_{\mathrm{PCA}}^{\mathrm{EXC}(s,t)} \qquad t = 1,\ldots,\mathrm{PC}^{\mathrm{EXC}} \tag{5.14}$$

In Eqs.(5.15)-(5.16) the equations to transform the material massive fractions in volumetric ones are reported. $r^{\mathrm{API}(i)}$ is the massive fraction of the *i*-th API, while $r^{\mathrm{EXC}(d)}$ is the massive fraction of the *d*-th excipient in the formulation. $\boldsymbol{\rho}^{\mathrm{API}}$ and $\boldsymbol{\rho}^{\mathrm{EXC}}$ are the vectors of the densities of the *NAPI* APIs and of the *NEXC* excipients included in the historical databases .

$$\mathbf{r}_{\mathrm{V}}^{\mathrm{API}} = \left(\mathbf{r}^{\mathrm{API}} / \boldsymbol{\rho}^{\mathrm{API}}\right) \cdot 100 \Bigg/ \left(\sum_{i=1}^{NAPI} r^{\mathrm{API}(i)} / \rho^{API(i)} + \sum_{d=1}^{NEXC} r^{\mathrm{EXC}(d)} / \rho^{\mathrm{EXC}(d)} \right) \tag{5.15}$$

$$\mathbf{r}_{\mathrm{V}}^{\mathrm{EXC}} = \left(\mathbf{r}^{\mathrm{EXC}} / \boldsymbol{\rho}^{\mathrm{EXC}}\right) \cdot 100 \Bigg/ \left(\sum_{i=1}^{NAPI} r^{\mathrm{API}(i)} / \rho^{\mathrm{API}(i)} + \sum_{d=1}^{NEXC} r^{\mathrm{EXC}(d)} / \rho^{\mathrm{EXC}(d)} \right) \tag{5.16}$$

Note that Eqs.(5.6)-(5.16) represents the two blend prediction model steps, namely the weigthed score calculation procedure and the PLS model step between the weighted scores and the product properties. Logical constraints are then added to select the most suitable materials according to the hypotheses described above (Section 5.2.2). In Eq.(5.17) these logical constraints for the selection of only one excipient per family are reported. $b^{\mathrm{EXC}(z_j),L_j}$ is a binary variable which indicates if the *z*-th excipient of the L_j family is ($b^{\mathrm{EXC}(z_j),L_j} = 1$) or not ($b^{\mathrm{EXC}(z_j),L_j} = 0$) inside the formulation. $\mathbf{L}$ is the set of the *J* families, specified by the user, each including *NEXC$_j$* excipients among which to select the most suitable ones.

$$\sum_{z_j=1}^{NEXC_j \in L_j} b^{\mathrm{EXC}(z_j),L_j} = 1 \qquad \mathbf{L} = \{\mathrm{Exc}_1, \mathrm{Exc}_2, ..., \mathrm{Exc}_J\} \qquad j = 1,...,J \tag{5.17}$$

In Eqs.(5.18)-(5.21) the constraints for the massive fractions of the materials are reported. In particular the fractions of all the APIs but the selected one are set to 0 (Eq.(5.19)), while the magnesium stearate fraction is assigned (Eq.(5.21)). In Eq.(5.20) U^{EXC,L_j} represents the upper limit defined by the user for the massive fraction of the *z$_j$*-th excipient of the *j*-th family inside the formulation.

$$r^{\mathrm{API}(l)} = \frac{m^{\mathrm{API}}}{m^{\mathrm{TOT}}} \qquad l = \text{selected API} \tag{5.18}$$

$$r^{\mathrm{API}(i)} = 0 \qquad i \neq \text{selected API} \tag{5.19}$$

$$r^{\mathrm{EXC}(z_j),L_j} \le b^{\mathrm{EXC}(z_j),L_j} \cdot U^{\mathrm{EXC},Lj} \qquad \begin{cases} z_j = 1,..., NEXC_j & j = 1,...,J \\ \mathbf{L} = \{\mathrm{Exc}_1, \mathrm{Exc}_2, ..., \mathrm{Exc}_J\} \end{cases} \tag{5.20}$$

$$r^{\mathrm{MgSt}} = 1 \tag{5.21}$$

In Eqs.(5.22)-(5.24), the constraints due to mass balances are shown. In Eq.(5.23), $\mathbf{m}^{\mathrm{EXC}}$ represents the vector of the masses of the all the excipients inside the blend, while m^{MgSt} in Eq.(5.24) is the magnesium stearate mass.

$$\mathbf{r}^{\mathrm{API}} + \mathbf{r}^{\mathrm{EXC}} + r^{\mathrm{MgSt}} = 1 \tag{5.22}$$
$$\mathbf{m}^{\mathrm{EXC}} = \mathbf{r}^{\mathrm{EXC}} \cdot m^{\mathrm{TOT}} \tag{5.23}$$
$$m^{\mathrm{MgSt}} = r^{\mathrm{MgSt}} \cdot m^{\mathrm{TOT}} \tag{5.24}$$

In Eqs.(5.25)-(5.26), the constraints for the API dose in the formulation and for the final tablet total mass are reported. These constraints are active in the case m^{API} and m^{TOT} do not appear in the objective function. In Eq.(5.25) and Eq.(5.25), M_1 and M_2 are the boundary values for the dose and the tablet weight defined by the user.

$$g_4 \cdot m^{\mathrm{API}} = (\geq) g_4 \cdot M_1 \tag{5.25}$$
$$g_5 \cdot m^{\mathrm{TOT}} = (\leq) g_5 \cdot M_2 \tag{5.26}$$

Finally, in Eqs.(5.27)-(5.31) the constraints for each of the variables in $\hat{\mathbf{y}}^{\mathrm{NEW}}$ (described above) predicted by the model are reported. lb and ub represent respectively the lower and upper bound of each variable. These bounds are considered in the inversion problem only if specified by the user.

$$lb_{Bmid80} \leq \hat{y}_{Bmid80} \leq ub_{Bmid80} \tag{5.27}$$
$$lb_{TensileStrength} \leq \hat{y}_{TensileStrength} \leq ub_{TensileStrength} \tag{5.28}$$
$$lb_{FFC} \leq \hat{y}_{FFC} \leq ub_{FFC} \tag{5.29}$$
$$lb_{BFI} \leq \hat{y}_{BFI} \leq ub_{BFI} \tag{5.30}$$
$$lb_{CompressionStres} \leq \hat{y}_{CompressionStress} \leq ub_{CompressionStres} \tag{5.31}$$

The above-described procedure has been implemented using Matlab® (the MathWorks Inc., Natick, MA) and GAMS (GAMS Development Corp., Washington, D.C.). In particular, the data handling, the blend prediction model building and the results analysis have been performed in MATLAB using an in-house developed multivariate analysis toolbox (phi v1.7). The mixed-integer nonlinear optimization problem has been implemented and solved in GAMS using the BARON solver (Sahinidis, 1996).

5.3 Results

The blend prediction model inversion framework has been tested to design the formulation for a proprietary API, which for confidential reasons in the following text will be referred to as API-A. The proposed strategy is applied in two different case studies in which the objective is to maximize the API dose in the formulation. To this end, the objective function for the optimization problem (Eq.(5.5)) reduces to the term representing the mass of API in the formulation (m^{API} in Eq.(5.5), with a negative value for the weight g_4). The procedure will thus give in output the list of material types and fractions that have to be included in the formulation, together with the API fraction which, according to the model, ensures the desired blend with the maximum API load. In particular, in Case study 1 it is assumed that the formulation is composed by two excipients, that have to be selected within the lactose and the MCC families, and a disintegrant which has to be chosen among the materials in its relevant family (Table 5.1). In Case study 2, one of the excipients in the formulation is assumed to be known (i.e., it is constrained in the optimization framework) and corresponds to calcium phosphate dibasic anhydrous (A-Tab). The second excipient is constrained to be selected within the lactose excipient families. Following the hypothesis reported in Section 5.2.2 (step 1.), a disintegrant has to be selected as well inside the disintegrant family (Table 5.1). In both case studies, the magnesium stearate fraction is considered fixed and set to 1 wt%.

In both case studies, the formulations were designed with the aim of reaching a blend suitable for direct compression. Accordingly, design specification ranges were assigned for each blend property, as reported in Table 5.2. These ranges act as constraints for the model predictions in the inversion problem. Note that the ranges were established based on experience, and they have been made dimensionless in Table 5.2 and in the remainder of the Chapter for confidentiality reasons. Following the industrial practice, a rating category system has also been applied for each blend property. The categories classify blend properties as “Attribute”, “Marginal”, “Deficient” and “Severely deficient”, and were developed for each property to give an indication of processing performance for solid-dosage form development (Polizzi and García-Muñoz, 2011).

Table 5.2. *Ranges of blend properties which define a blend suitable for direct compression. These ranges provide the constraints for the blend prediction model inversion problem.*

Property	**Constraints**
$B_{MID}80$	< 0.36
Compression stress	> 0.20 and < 0.50
Tensile strength	> 0.40
BFI	< 0.22
FFC	> 0.57

The formulations obtained from the optimization exercise were experimentally tested to validate the model-based design. Experimental results and model predictions were compared based on the numerical results and on the rating category. The goodness of the results was also assessed through a Student’s *t*-test (Montgomery and Runger, 2010) with the aim of verifying that, for each blend predicted property, the difference between model predictions and experimental values was not statistically significant for the tested blends within uncertainties. The comparisons were performed univariately as the rating categories for the properties are defined univariately. Results of the statistical hypothesis testing are reported in terms of *p*-values: the tested null hypothesis that there is not a significant difference between model predictions and experimental values within experimental and prediction uncertainty is rejected when the *p*-value is below the significance level $\alpha = 0.05$. Experimental uncertainty has been estimated from the properties of the blends in the historical dataset with the same formulation (same material type and very similar material fractions). Model prediction uncertainty has been estimated through jackknifing (Duchesne and MacGregor, 2001).

5.3.1 Case study 1: formulation design and experimental validation

In Table 5.3 results are reported for the Case study 1, dealing with the *in-silico* formulation design for the maximum dose of API-A when no constraints were assigned to the input materials (except for the lubricant). In this case the optimization problem involves 961 variables (of which 23 integers) and 951 constraints. The list of the estimated excipients is reported together with their fractions and the API fraction expressed in wt%. Note that this represents the formulation that, according to the model, has the highest API dose. However, this does not mean that a formulation with a higher API load could not exist. The formulation with the maximum API load calculated through the model is the one for which the model is still valid. Otherwise stated, the model can estimate a formulation with a higher API load but at the expenses of the reliability in its estimates, which may not adhere to the correlation structure represented by the PLS model. As described in Chapter 2, the squared prediction error (SPE) is used to quantify the lack of representativeness of the model. In the case of the formulation reported in Table 5.3, SPE = 4.75, that is equal to the value assigned for its constraint (= 0.9×5.28 in Eq.(5.11) above). This means that the constraint on SPE is an active constraint in the optimization and the obtained formulation is a “limit” formulation, with respect to the model correlation. This is not surprising, since the objective is to maximize the API dose.

Table 5.3. *Case study 1. Model-based formulation design for API-A when no constraints are assigned to the input materials*: *estimated material types and fractions.*

Model-based formulation	Material fraction in the formulation (wt.%)
API-A	14.47
Avicel PH200	63.31
Lactose Anhydrous (direct tabletting)	18.21
Sodium starch glycolate (Explotab)	3.01
Magnesium Stearate	1.00

The fact that the estimated API dose seems not to be very high (14.47 wt.%) is a consequence of the suboptimal properties of the selected API and of the historical data range used to build the model. The selected API (API-A) scores are quite far from the center of the multivariate API design space, meaning that its properties are different from the multivariate API mean.

The optimizer thus compensates for this difference (which impacts on the calculation of the weighted scores) by lowering the API fraction. Accordingly, the estimated formulation in Table 5.3 is found at a significant distance from the center of the historical formulation design space (represented by the PLS model latent space), even if still inside the relevant confidence limit. As a matter of example, in Figure 5.3 the space of the LV scores on the first and on the third LVs of the PLS model is reported. In the plot, the black dots (●) correspond to the historical formulations, while the empty triangle (△) represents the projections of the formulation in Table 5.3 (the empty square □ is instead representative of the solution for Case study 2 reported in the next section).

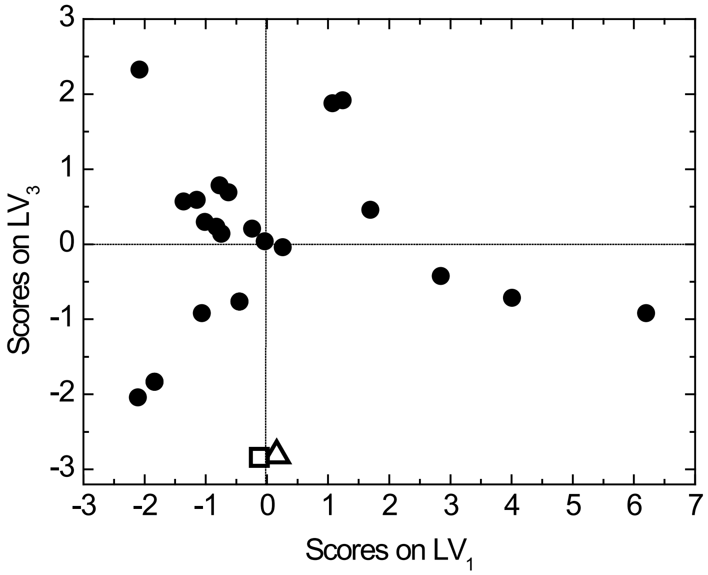


Figure 5.3. *Score space on the first and on the third LVs of the PLS model step of the blend prediction model. The black dots represent the historical formulation projections, the empty triangle (△) the projections of the solution for Case study 1, while the empty square (□) the projections of the solution for Case study 2.*

The distance of the formulation projections from the origin of the latent space calculated considering all the LVs of the model is quantified by the Hotelling's T^2 statistics (Chapter 2, Section 2.1.1.3), which for the formulation in Table 5.3 equals 11.45, below the relevant confidence limit calculated from the historical data (= 17.45).

The model-based designed formulation reported in Table 5.3 was prepared in laboratory to test experimentally if the blend designed *in-silico* achieved the desired properties assigned as constraints to the optimization problem (Table 5.2). Note that the blend prediction model used in this study does not account for the effect of the number of blender revolutions on tablet properties (Kushner and Moore, 2010), as there was no information on the blending time/revoutions for the historical blends used to build the model. For this reason, two blends were prepared and tested at different blending times (~ 3 minutes and ~ 60 minutes), and the mean value of each obtained blend property was considered for the experimental validation.

Table 5.4 reports the comparison between the model predictions for the properties of the designed blend of Table 5.3 and the experimental values obtained from the prepared blend. For each model prediction, the uncertainty is quantified by half the width of the 95% respective confidence intervals, which are reported in parentheses (95% CI). Similarly, for each experimental value the experimental uncertainty, estimated from the historical data, is reported in parentheses. The last column represents the *p*-value of the Student's *t*-test performed to compare model predictions with the experimental values.

Table 5.4. *Case study 1. Model predictions and experimental validation of the blend properties for the in-silico formulation of Table 5.3. Uncertainties are reported for model predictions and experimental values (95% CI), together with the p-value from the t-tests to compare them.*

Property	**Model predictions**	**Experimental validation**		***p*-value**[#]
	Value (95% CI)	**Value (95% CI)**	**Rating category**	
$B_{MID}80$	0.33 (0.29)	0.31(8.0e-2)	Attribute	0.514
Compression stress	0.50 (0.20)	0.31 (7.7e-2)	Attribute	0.018
Tensile strength	0.58 (0.21)	0.32 (0.37)	Marginal	0.267
BFI	3.82e-2 (7.31e-2)	6.19e-2 (5.37e-2)	Attribute	0.392
FFC	0.70 (5.4e-2)	0.75 (5.6e-2)	Attribute	0.227

# cutoff: 0.05

The first important remark to note when observing the results in Table 5.4 is that all the properties measured for the real blend satisfy the constraints assigned to the optimization problem and listed in Table 5.2, except for tensile strength. Analogously, the rating category is found to be "Attribute" for all the properties but tensile strength. This is an important result, as blends with similar ratings are considered to be similar in a real-world manufacturing environment. Nevertheless, by examining the *p*-values, it can be seen that the difference between the prediction and the measured value for the tensile strength is not found to be significant. This is mainly due to the uncertainty in the experimental data, which is even larger than the prediction uncertainty for tensile strength. Therefore, despite the experimental value for this property is not satisfying the assigned constraint, in this case the model is actually doing a good job, as model predictions (and the value of the constraint) are well within the experimental uncertainty.

Conversely, from Table 5.4 it can be noticed that, even if the compression stress measured for the real blend falls within the range desired for it, the difference between the model prediction and the measured value is significant (*p*-value below 0.05, even if the null hypothesis would not be rejected if the significance level were 0.01). It was found that this property is the one for which the model shows the least predictive ability, and therefore an inferior prediction performance was somewhat expected for this property. Furthermore, as mentioned earlier, the formulation reported in Table 5.3 is a limit formulation with respect to model representativeness. This can impact on the uncertainties and therefore also on the reliability of the model predictions.

To get a visual representation and a thorough understanding of the differences between measurements (◇) and predictions (●), Figure 5.4 shows a graphical comparison between the predicted and the measured values for each blend property. In particular, each predicted and experimental value is reported together with the relevant 95% confidence limit (the vertical bars). The regions corresponding to the different rating categories are reported as well with different gray shadings, going from the "Attribute" (in white) to the "Severely Deficient" region (in dark gray). The axes scales of each plot have been coded for confidentiality reasons. As can be seen, for $B_{MID}80$, compression stress and BFI the prediction uncertainties are higher than the experimental ones. The reverse is true for tensile strength and FFC, where experimental uncertainties are larger than the prediction ones. These variables are however the ones most affected by the different blending time/number of revolutions. Since the blends in the historical dataset were obtained at different lubrication levels, which the model does not take into account, the effect of this parameter is included on the estimation of the experimental uncertainty, which is indeed more significant for the variables that are most affected by the lubrication itself. Nonetheless, for Case study 1 each plot of Figure 5.4 confirms that the values of all the experimental properties fall within the model prediction uncertainty (except for tensile strength, as seen above). Moreover, the experimental values for most of the properties fall inside the Attribute region, even considering the uncertainty limits.

5.3.2 Case study 2: formulation design and experimental validation

Table 5.5 reports the *in-silico* designed formulation when one of the formulation excipients is constrained to be calcium phosphate dibasic anhydrous (A-Tab). Again, the objective was to design a formulation allowing the maximum API-A dose. In this case the optimization problem involves 950 variables (of which 12 integers) and 946 constraints. The estimated formulation is again found to be a limit formulation, being its SPE equal to 4.75, which is the value assigned for the corresponding constraint (Eq.(5.11)). The formulation is inside the multivariate design space of the historical formulations, as demonstrated by the value of the Hotelling's T^2 = 10.41, even if at a significant distance from the center of the space (as seen from the projections on the first and on the third LVs of the PLS model represented by the empty square □ in Figure 5.3).

The blend designed *in-silico* and reported in Table 5.5 was then prepared in laboratory as done for Case study 1, and the properties of interest were measured to validate the results obtained from the proposed procedure.

Table 5.6 reports the comparisons between the model predictions and the experimental measurements. As above, predictions and real values are reported with the relevant 95% confidence intervals; the rating categories for the experimental values and the *p*-values of the

Student's *t*-test are reported as well in the table. As in the previous case study, the properties measured for the prepared blend satisfy the assigned constraints (Table 5.2), thus giving a blend of desired properties, except from tensile strength. Accordingly, the rating category is found to be "Attribute" for all the measurements, except tensile strength.

Table 5.5. *Case study 2. Model-based formulation design for API-A when the choice of an excipient (calcium phosphate dibasic anhydrous) is constrained: estimated material types and fractions.*

Model-based formulation	Material fraction in the formulation (wt.%)
API-A	13.67
Avicel PH200	72.03
Calcium phosphate dibasic anhydrous (A-Tab)	10.30
Sodium starch glycolate (Explotab)	3.00
Magnesium Stearate	1.00

Table 5.6. *Case study 2. Model predictions and experimental validation of the blend properties for the in-silico formulation of Table 5.5. Uncertainties are reported for model predictions and experimental values (95% CI), together with the p-value from the t-tests to compare them.*

Property	Model predictions	Experimental validation		*p*-value#
	Value (95% CI)	Value (95% CI)	Rating category	
$B_{MID}80$	0.34 (0.26)	0.28 (8.0e-2)	Attribute	0.165
Compression stress	0.49 (0.19)	0.36 (7.7e-2)	Attribute	0.047
Tensile strength	0.67 (0.19)	0.38 (0.37)	Marginal	0.247
BFI	~0 (6.32e-2)	5.24e-2 (5.37e-2)	Attribute	0.178
FFC	0.63 (4.4e-2)	0.79 (5.6e-2)	Attribute	0.064

# cutoff: 0.05

From the *p*-values it is interesting to note that compression stress shows again a statistically significant difference between measurements and predictions, even if close to the 0.05 cutoff value. For all the other properties, the difference between the predicted and the real properties of the blend is not found to be significant, even if for FFC, if the *p*-value of the test is quite near to the 0.05 significance level.

A graphical illustration of the results is reported in Figure 5.4 (right portion of the figures). The same remarks reported for Case study 1 applies also for Case study 2.

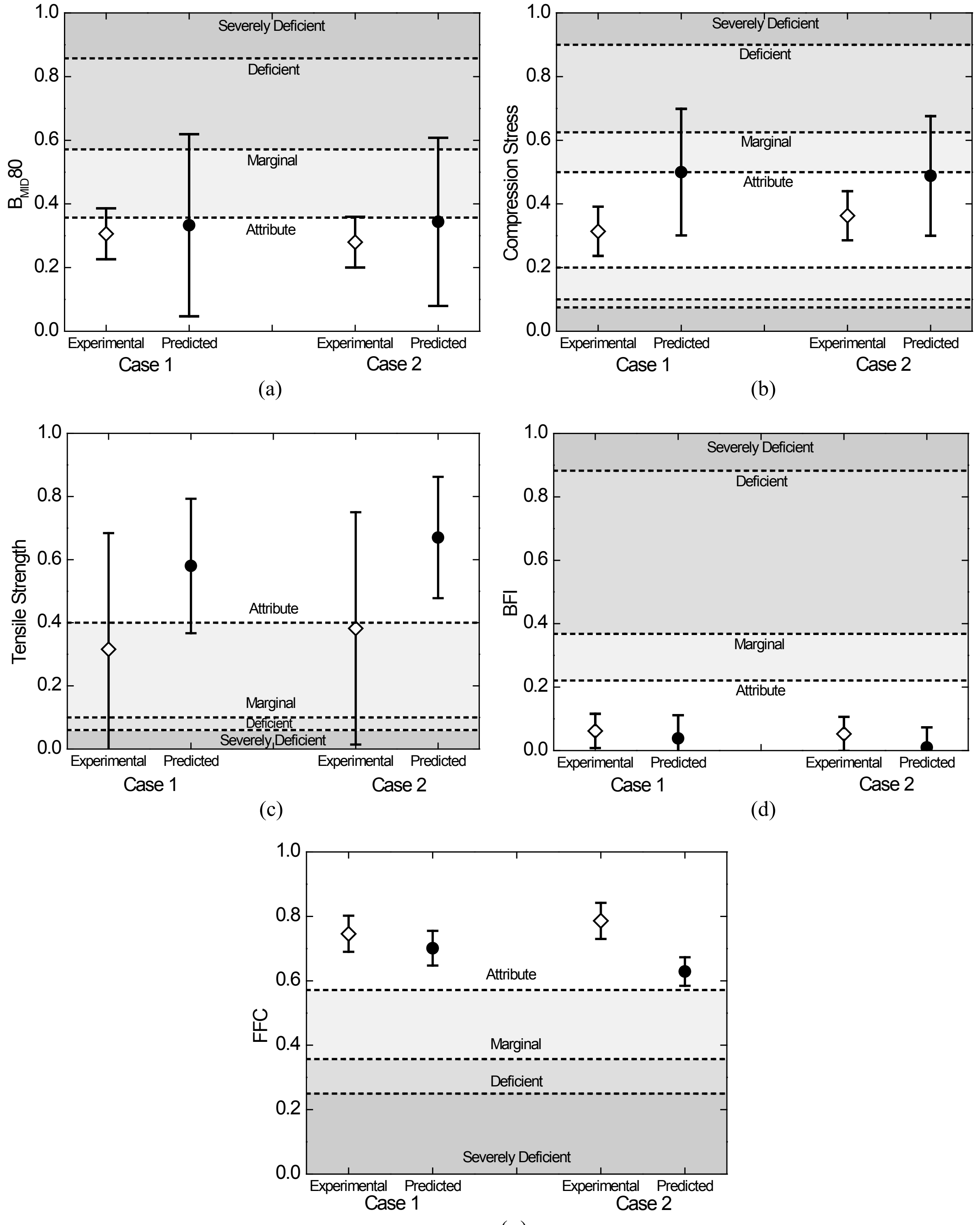
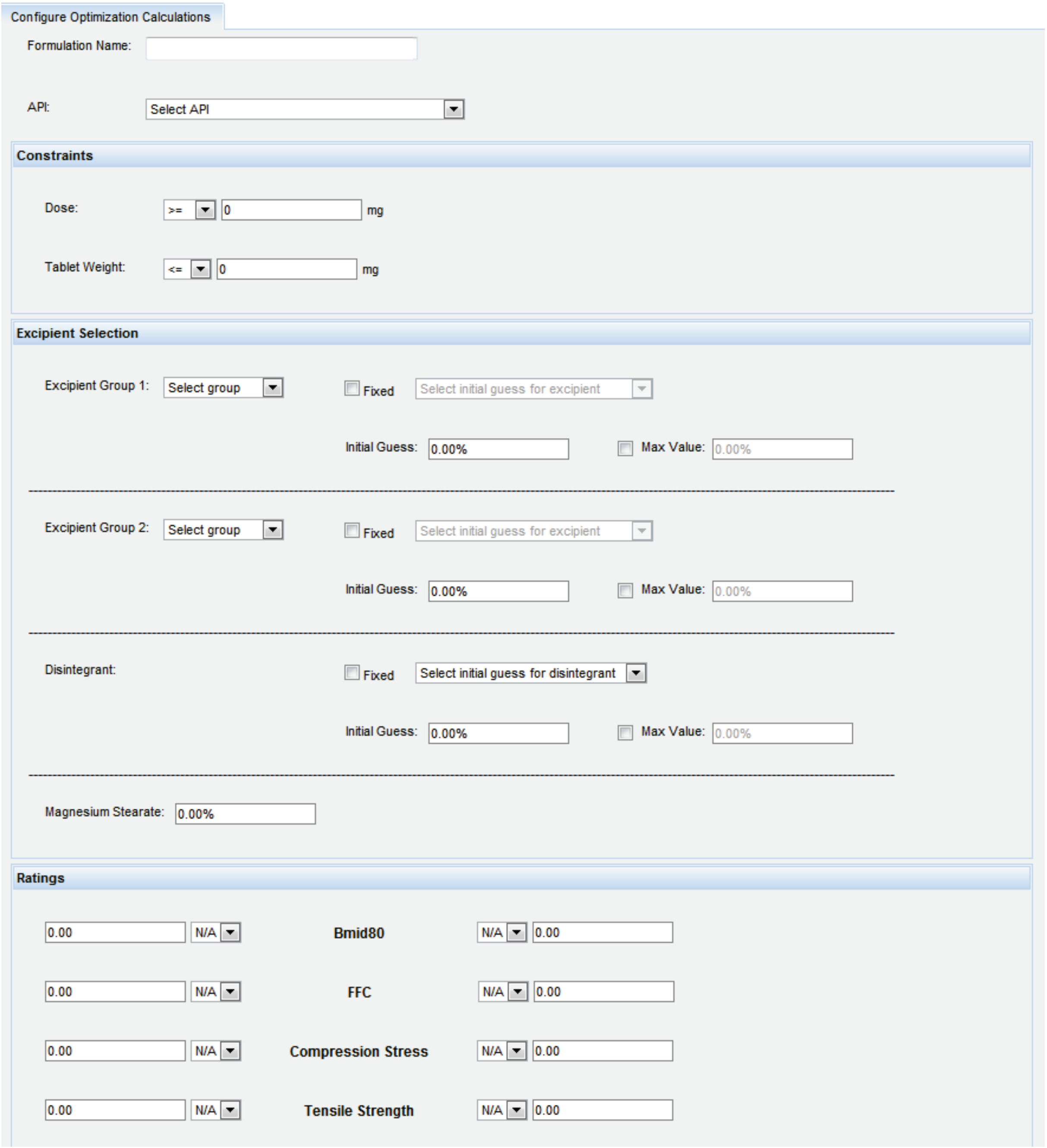


Figure 5.4. *Comparisons between model predictions (●) and experimental values (◇) for the properties of the blends designed in-silico in Case study 1 and Case study 2. (a) $B_{MID}80$; (b) compression stress; (c) tensile Strength; (d) BFI; (e) FFC. In each plot, regions corresponding to different rating categories have been reported with different grey shadings.*

5.4 A user-friendly interface

Figure 5.5 reports a snapshot of the user-friendly interface that was developed to allow users (e.g., formulation scientists) to set the inputs required for the blend prediction model inversion. The interface provides the inputs to the Matlab and GAMS codes, in which the optimization problem is solved and results are processed.



Configure Optimization Calculations

Formulation Name:

API: Select API

Constraints

Dose: >= 0 mg

Tablet Weight: <= 0 mg

Excipient Selection

Excipient Group 1: Select group □ Fixed Select initial guess for excipient

Initial Guess: 0.00% □ Max Value: 0.00%

Excipient Group 2: Select group □ Fixed Select initial guess for excipient

Initial Guess: 0.00% □ Max Value: 0.00%

Disintegrant: □ Fixed Select initial guess for disintegrant

Initial Guess: 0.00% □ Max Value: 0.00%

Magnesium Stearate: 0.00%

Ratings

0.00	N/A	Bmid80	N/A	0.00
0.00	N/A	FFC	N/A	0.00
0.00	N/A	Compression Stress	N/A	0.00
0.00	N/A	Tensile Strength	N/A	0.00

Figure 5.5. *Interface developed for the blend prediction model inversion exercise. The interface allows users to input several objectives and constraints in terms of materials to be included in the final formulation, their fractions and desired properties for the final blend.*

The structure of the interface reflects the inputs and the constraints needed by the procedure, described in Section 5.2.2. First, the user needs to provide a name for the formulation and to select the API.

The second section of the interface is the one in which the user should set the constraints for the API dosage and the tablet weight. As described in Section 5.2.2, the API dose can be constrained to be equal or greater than an assigned value, or maximized. Similarly, the tablet weight can be minimized or constrained to be equal or lower than an assigned value.

The third section of the interface refers to the excipient selection constraints. As can be seen from Figure 5.5, the user has to specify the group (family) within which the procedure has to select the best excipient. Options are provided to assign a given excipient in the formulation (as in Case study 2; Section 5.3.2) and to bound the maximum fraction of each excipient in the formulation (U^{EXC,L_j} in Eq.(5.20)). Initial guesses both for the excipient type and fraction may be provided as well. As can be seen, following the initial hypothesis on the formulation composition (Section 5.2.2), the excipient selection section has been programmed in order to include 2 different excipients, a disintegrant and magnesium stearate in the formulation. The magnesium stearate fraction can be provided by the user as well.

The last section of the interface refers to the ratings specifications. In this section, the user may provide the desired values/ranges for each of the five blend properties described in Section 5.2.2. For each property, the user can assign specific values (equality constraints) or ranges (inequality constraints). Furthermore, he/she may also decide to exclude one or more properties from the analysis.

5.5 Conclusions

In this Chapter, a novel method for the *in-silico* design of new pharmaceutical formulations has been proposed and experimentally validated. The aim of the work was to propose a systematic procedure, which, on a scientifically sound basis, is able to suggest which materials and in which amount should be tested in a formulation, in order to obtain a blend of desired properties.

The method relies on the inversion of a latent variable regression model (LVRM) recently proposed to model mixture data of complex formulations. The proposed procedure relies on an optimization framework which, given an API, selects the best excipients types and amounts that, according to the model, would ensure the achievement of the desired blend. The methodology has been implemented in order to manage the different constraints a formulation scientist may have to face for the formulation of a new product. These may include constraints on the excipient family among which the most suitable excipient should be chosen, constraints on the excipient type itself, on the API dose, or on the final tablet weight.

Furthermore, the procedure can deal with different objectives the user may have, such as the maximization of the API dose in the final tablet, or the minimization of the tablet weight.

The effectiveness of the method has been tested in the design of the formulation for an assigned proprietary API (API-A), in order to obtain a blend suitable for direct compression. In particular, the objective was to design the formulation for the maximum API dose, in two different case studies. In Case study 1, the excipient choice was not constrained; however, a constraint was active on the families of the excipients among which to address the search. The optimizer was used to design the whole formulation, in terms of both materials types and amounts. In Case study 2, a constraint on one of the formulation excipients was assigned, while the optimizer was asked to select all of the other excipients and their amounts.

The blends designed *in-silico* have been prepared in laboratory and characterized, in order to compare the real properties of the blend with the desired ones, which formed the design specifications for the optimization problem. Experimental results showed a good agreement with model predictions. Almost all of the real blend properties fall within the model prediction uncertainty and, above all, they are ranked “Attribute”, namely optimal for direct compression. Some issues have been found for tensile strength, whose achieved value does not match the constraints assigned to the *in-silico* procedure. However, it was found that this property is strongly affected by the number of blender revolutions, whose effect is not accounted for by the model. Nonetheless, the difference between model predictions and real values is not found to be statistically significant. More significant were the differences found in the comparisons between the real and the predicted values of compression stress, which were however due to the modest performance of the model in predicting this property. To overcome these limitations, the model could be improved by augmenting the historical datasets, as more materials and especially formulations will be available. Furthermore, results showed that to obtain reliable predictions (and thus to have a robust model to be used in formulation design), information on the blending process should be included in the model. At the same time, the approach can be extended to consider other constraints and targets (such as stability), as long as there is a mathematical way to relate them (e.g., degradation extent) to the incoming formulation. These are further steps that will be taken into account in the future.

The proposed strategy provides a systematic tool to support pharmaceutical development personnel in the design of new product formulations, by guiding experiments and suggesting *a priori* the most suitable materials to test, based on a scientific basis. Different and complex databases can be managed in a straightforward way through the use of LVRMs, which enable a more transparent model structure and simpler interpretation compared to other data-based methods (e.g., black-box models). Accordingly, the optimization framework ensures to find the optimal formulation, on the basis of the historical knowledge. This allows accelerating the decision making process in product development activities and overcome issues related to resorting to traditional procedures in pharmaceutical development.

Chapter 6

Transfer of products between different plants*

This Chapter presents a methodology to address the issue of transferring in a target plant a product already manufactured in one or more source plants, which may differ for size, lay-out or involved units. The procedure is based on two steps: first, data from the source plant(s) are related to the (usually few) available data from the target plant through a latent variable regression model (LVRM). Then, the model is inverted, following the general framework proposed in Chapter 4, to suggest the optimal process conditions which, according to the model, ensure to manufacture the desired product in the target plant. The methodology is tested experimentally on a process for the manufacturing of nanoparticles for pharmaceutical applications through a solvent displacement process. The experiments confirm the effectiveness of the proposed procedure and provide an experimental validation of the theoretical concept of null space.

6.1 Introduction

One of the most burdensome problems in product and process development is the transfer of technology between plants. The ultimate objective of the transfer is to obtain in a target plant a product of desired quality, which has usually already been obtained in a source plant. This problem is commonly encountered in process scale-up or in the transfer of production between different manufacturing sites, where the involved equipment may be different for size or layout. Commonly, in the source plant (e.g., a small-scale plant) extended experimental campaigns are carried out to gain process understanding and disclose potential pitfalls in the process equipment, operating conditions or control configurations. This experimentation eventually leads to the definition of the process equipment layout, as well as of a set of operating conditions that can guarantee the required product quality with acceptable variability. When the production has to be moved to a different plant (e.g., a large-scale plant), an extended experimental campaign may be impossible or economically

*Tomba E., N. Meneghetti, P. Facco, T. Zelenková, D.L. Marchisio, A.A. Barresi, F. Bezzo and M. Barolo (2013). Product transfer between different plants through latent variable model inversion. Submitted to *AIChE J.*.

unsustainable. An issue therefore arises on whether it is possible to exploit the data available from the experiments performed in the source plant to assist the transfer of technology to the target plant. To this end, appropriate methodologies are needed to exploit the knowledge available for the source plant in order to guide the experimentation in the target plant with the aim of accelerating the transfer and subsequently the time-to-market for new products. This Chapter focuses on the product transfer problem, namely the problem of estimating the process operating conditions in the target plant, wherein the manufacturing is being started, to obtain a product of desired properties, by exploiting the knowledge acquired from the source plant.

Approaches to guide technology transfer activities have already been proposed for specific matters, as the transfer of models between instruments (Feudale *et al.*, 2002) or plants (Lu and Gao, 2008a and 2008b; Lu *et al.*, 2009). Model transfer approaches will be reviewed in Chapter 7, in which a procedure to transfer monitoring models between plant is proposed.

In general, model-based transfer approaches exploit features that the process may have in common in the different plants to address the transfer. A similar rationale is exploited also in product transfer when dimensional analysis is applied. Dimensional analysis is commonly used to identify plant-independent variables (e.g., dimensionless numbers), which indicate the relevance of the physical phenomena occurring in the process. Usually, the transfer is driven by criteria that are set on the plant-independent variables and aim to ensure that the process in the different plants is operated under similar physical regimes (Zlokarnik, 2006). However, this approach often requires a deep mechanistic knowledge of the process under investigation, which for many processes may not be available.

A feasible alternative to tackle the product transfer problem is to exploit the historical datasets that are usually available in product and process development environments, for example from screening experiments or from studies on products already manufactured. Jaeckle and MacGregor (2000b) pioneered to use historical databases of products already manufactured in the target and in the source plant to guide the experimentation in order to simplify and accelerate the transfer of a new product in the target plant. The authors related the datasets of the process conditions in each plant through the data of the historical common products manufactured in the plants, and used LVRM inversion to estimate the process conditions for the target plant to manufacture a new product of assigned properties. This strategy has been further refined by García-Muñoz *et al.* (2005), who proposed the joint-Y projection to latent structures (JY-PLS) method to relate data from different plants (Chapter 2, Section 2.1.3.2). Assuming that the correlation between the properties of the historical products manufactured in different plants is similar, JY-PLS exploits the latent space generated by the joint dataset of the product properties to relate the corresponding process data. The model can then be inverted to estimate the optimal process conditions that, according to the model, ensure the achievement of the desired product in the target plant. Some contributions on the application of these techniques to support the scale-up of critical operations in the pharmaceutical industry have recently appeared in the literature and have been reviewed in Chapter 1, Section 1.4.1.3 (Liu *et al.*, 2011b; Muteki *et al.*, 2011).

In this Chapter a general methodology is proposed and tested experimentally to support product transfer between different plants based on JY-PLS modeling and LVRM inversion. To this purpose, the general framework for LVRM inversion proposed in Chapter 4 is used to invert the JY-PLS model built on the historical datasets available for the plants, with the aim of suggesting an appropriate set of operating conditions to be tested in the target plant, in order to manufacture therein a product with desired quality specifications. The proposed methodology explicitly addresses the issue of managing constraints in both inputs (operating conditions) and outputs (quality specifications). It can also cope with the differences that may occur in the experimental setup over time (due for example to maintenance, changes in the ancillary equipment, sensors or operators). By proceeding this way, the knowledge available from the experiments carried out in the source plant(s) is optimally exploited to streamline the product transfer.

The proposed strategy is applied to an experimental case study dealing with a nanoparticle production process through solvent displacement in passive mixers (Lince *et al.*, 2009). This process is widely used in the pharmaceutical industry to manufacture polymer nanoparticles that can be used as drug carriers for controlled drug delivery. The problem under investigation is to transfer the nanoparticle production system from a reference passive mixer (also called plant in the following), for which a large historical database is available, to a target passive mixer of different scale (or geometry), where limited historical data are available. The problem is further complicated by the fact that the historical dataset available for the target plant was developed by running the plant under a different experimental setup than the one under which the current experimentation can be carried out. By using new experiments designed and carried out in this study, the first experimental validation of the concept of null space (Chapter 2, Section 2.2.1), introduced theoretically by Jaeckle and MacGregor (1998), is also provided.

6.2 Process and datasets

An experimental precipitation process to manufacture pharmaceutical nanoparticles through solvent displacement is considered in this study. This process is used to manufacture polymer nanoparticles that are used for drug delivery and controlled drug release. It consists of a dissolution phase (in which the drug is dissolved in a solvent together with polymer and other additives), a mixing phase (in which the prepared solution is mixed with an anti-solvent, usually water, in a mixing chamber), and a solvent elimination phase to give the final product (Lince *et al.*, 2008). The process is operated continuously in passive mixers, where the solution of the drug and polymer and the anti-solvent are injected through separate inlets, and mixed. As soon as the two solutions mix, nanoparticles are formed and then collected at the mixer output. One of the most important issues related to this process is to control the nanoparticle size. In fact, depending on the administration route, the potential of nanoparticles as drug delivery systems depends on their particle size distribution. For parenteral administration, in order to avoid negative interactions with the reticulo-endothelial system, particles must have an assigned dimension (or range of dimensions), which guarantee an adequate life in the blood stream and a continuous and controlled drug release (Moghimi *et al*., 2001; Alexis *et al*., 2008).

In the case study considered in this paper, the process occurs in confined impinging jets mixers (CIJMs). The objective is to manufacture nanoparticles of desired size in a target device B (CIJM-d2), by exploiting the data available from experiments performed on a source device A of different size (CIJM-d1) and from experiments performed on the device B itself, but under a different overall experimental setup. The mixers are schematically shown in Figure 6.1 and differ for the size of the inlet pipes: CIJM-d1 is characterized by a mixing chamber diameter of about 5 mm and inlet pipes inner diameter of 1 mm, whereas CIJM-d2 has the same chamber geometry and dimension and the same pipe length as CIJM-d1, but inlet pipes inner diameter of 2 mm.

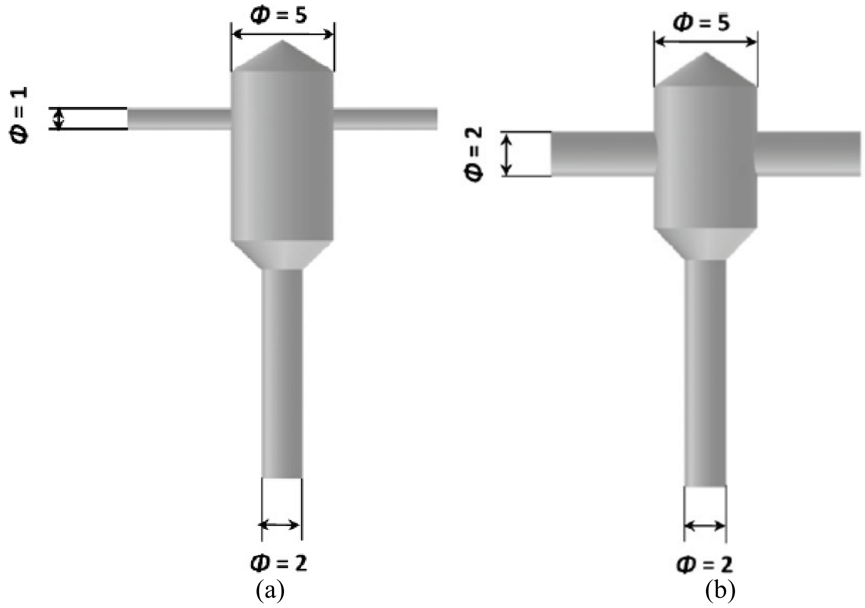


Figure 6.1. *Sketch of the different mixers considered in this work for the product transfer problem. (a) Device A: CIJM-d1. (b) Device B: CIJM-d2 (adapted from Lince et al., 2011a). All dimensions are in millimeters.*

Although the difference in the geometries of the two mixers may seem negligible, being the process highly mixing sensitive it has a drastic impact on the final nanoparticle characteristics. There are several reasons justifying the investigation of these two different geometrical configurations. As an example, two of them are pressure drop optimization (clearly device B is characterized by smaller pressure drops than device A) and the requirement of improving process performances with respect to fouling or flow instability. Moreover, notwithstanding what the actual geometrical difference is, from the practical point of view any two different devices (with a measurable impact on product quality) can in principle be used to prove the capability of the presented methodology to successfully carry out process transfer activities.

Data are available from historical experiments carried out earlier in both devices to study the influence of the process parameters on the nanoparticle size. Part of the available datasets refer to the actual chemical systems used in real applications, namely a drug used for breast cancer treatment (doxorubicin) and different polymers, including a PEGylated polymer that forms stealth nanoparticles (i.e. poly(methoxypolyethyleneglycolcyanoacrylate-co-hexadecylcyanoacrylate)) (Lince *et al.*, 2011b). However, a simpler system was considered in this study. In fact, since the final drug loading is often relatively low, in many cases the overall particle formation process is controlled by the polymer nanoparticle formation process (i.e. polymer molecules self-assembly into nanoparticles). For this reason, the analysis is limited to the chemical system for which the largest historical database was already available. The experiments on the mixers were therefore performed by manipulating four variables: the polymer concentration in the initial solution (c_{pol}), the inlet water flowrate (*FR*), the anti-solvent/solvent flow rate ratio (*W/A*) and the polymer type (*Type*). All experiments were performed using poly-ε-caprolactone (PCL) as a polymer, but considering two lots of polymers of different molecular weight (MW). These lots are named PCL_{14} for the lot with low MW (MW=14000 kg/kmol) and PCL_{80} for the lot having higher MW (MW=80000 kg/kmol). Accordingly, the variable *Type* indicating the polymer lot is binary (*Type* = 0 for PCL_{14} and *Type* = 1 for PCL_{80}). As mentioned above, all experiments were carried out with no drug in the polymer solution, and using acetone (HPLC grade by Sigma Aldrich) as the polymer solvent, while distilled and micro filtrated water (Millipore System, Milli-Q RG, millipack® R 4.0 sterile pack, 0.22:m, Holliston MA, US) was used as the anti-solvent. The mean particle size (d_p) was the only property considered for the characterization of the nanoparticles obtained from the experiments.

The experiments were performed according to the following protocol (Lince *et al.*, 2011a). The polymer solution in acetone and bi-distilled water were fed into the mixers by means of a syringe pump (KDS200 syringe pump; KD Scientific, Massachusetts, US); 2 mL of water feed were generally employed for each test. The outlet stream was then collected in a small volume (10 mL) of distilled-microfiltered water under gentle stirring and then sampled for particle size measurements. This dilution quenched the particles and prevented the occurrence of secondary processes (e.g., aggregation) after the stream left the mixer. Experiments were typically repeated three times and the nanoparticle size distribution was then measured through dynamic light scattering (Zetasizer Nanoseries ZS90, Malvern Instruments, Worchestershire, UK).

The system under investigation can be unstable, hence it was not easy to ensure perfect repeatability across the experiments. Taking also into account the uncertainties in the preparation of the solutions, in the characteristics of the syringes, in the feeding rate of the pumping system and in the d_p measurements, variations in the mean nanoparticle size from repeated runs are considered acceptable if they range within 15% of the average value.

6.2.1 Dataset organization

The data available from the experiments have been organized in three datasets.

- Dataset A. $\mathbf{X}_\mathrm{A}$ (348×4) and $\mathbf{Y}_\mathrm{A}$ (348×1) refer to the experimental campaign performed on device A (CIJM-d1). $\mathbf{X}_\mathrm{A}$ includes the operating conditions of 348 experimental runs, whereas $\mathbf{Y}_\mathrm{A}$ collects the mean diameters d_p measured for the nanoparticles obtained. Experiments did not follow a structured experiment design campaign. Three process settings (c_{pol}, FR, W/A) were partially manipulated according to a one-factor-at-a-time strategy, but not all of the experiments were repeated with both available polymers. The extended experimental campaign on device A was carried out over a time window of about 24 months.
- Dataset B. $\mathbf{X}_\mathrm{B}$ (39×4) and $\mathbf{Y}_\mathrm{B}$ (39×1) refer to the experimental campaign originally performed on device B (CIJM-d2). $\mathbf{X}_\mathrm{B}$ includes the operating conditions of 39 experimental runs, whereas $\mathbf{Y}_\mathrm{B}$ collects the mean diameters d_p measured for the obtained nanoparticles. As for device A, three process settings (c_{pol}, FR, W/A) were partially manipulated according to a one-factor-at-a-time strategy, but on a small number of polymer concentration levels. Furthermore (and unfortunately), the polymer type was changed according to the c_{pol} level, thus confounding its effect on d_p. This introduced an artificial collinearity between these two operating variables. The original experimental campaign on device B was carried out on a time window of 12 months.
- Dataset C. $\mathbf{X}_\mathrm{C}$ (17×4) and $\mathbf{Y}_\mathrm{C}$ (17×1) refer to a more recent experimental campaign purposely designed and performed on device B (CIJM-d2) for this study. $\mathbf{X}_\mathrm{C}$ includes the operating conditions of 17 experiments, whereas $\mathbf{Y}_\mathrm{C}$ collects the mean diameters d_p measured for the nanoparticles obtained in the experiments. Nine of these experiments had been carried out using sets of experimental conditions already explored in dataset B (they were located close to the center of the historical experimental design space). However, some differences were observed in the obtained nanoparticle diameters with respect to the experiments included in dataset B. It should be noted that, compared to the experiments collected in historical datasets A and B, dataset C experiments were performed at a much later time and under a slightly different experimental setup (in terms of syringes used, pumping procedure and involved operators). This was believed to be a potential cause for

the observed differences in d_p. Therefore, to take this issue into account 8 new experiments were carried out on CIJM-d2, in which all of the four operating parameters were manipulated. By these new runs, some form of orthogonality was introduced between the operating parameters, in order to study their independent effect on d_p under the new experimental setup.

Table 6.1 reports the type and level of the operating parameters used in the experiments, and the ranges of particle diameters that were obtained. As can be seen, the number of levels assigned to the operating parameters in dataset A is significantly larger than the one in dataset B (especially for c_{pol}).

Table 6.1. *Operating parameters manipulated in the experiments, with the levels assigned for each operating parameter on each experimental campaign.*

	Level					
Parameter	$\mathbf{X}_A$ (348×4)	$\mathbf{Y}_A$ (348×1)	$\mathbf{X}_B$ (39×4)	$\mathbf{Y}_B$ (39×1)	$\mathbf{X}_C$ (17×4)	$\mathbf{Y}_C$ (17×1)
c_{pol} [mg/mL]	0.026, 0.21, 0.22, 0.42, 0.82, 1.39, 1.47, 2.28, 2.65, 3.66, 5.04, 5.05, 6.17, 10.46, 15.07, 24.83	–	1.47, 5.04	–	1.47, 3.25, 5.04	–
FR [mL/min]	3, 20, 40, 60, 80, 120	–	3, 40, 60, 80, 120	–	3, 40, 60, 67, 80, 120	–
W/A	1, 1.83, 2.88, 6.08, 8.06	–	1, 1.83, 2.88, 8.06	–	1, 1.92, 2.84, 2.94, 5.18	–
Type	PCL$_{14}$ (= 0), PCL$_{80}$ (= 1)	–	PCL$_{14}$ (= 0), PCL$_{80}$ (= 1)	–	PCL$_{14}$ (= 0), PCL$_{80}$ (= 1)	–
d_p range [nm]	–	[98.9 ÷ 1194]	–	[181.2 ÷ 587.7]	–	[144.8÷437.5]

6.3 Product transfer methodology

As stated previously, the objective of this work is to obtain nanoparticles of desired mean size d_p in device B under the new experimental setup, by exploiting the historical data available from experiments in device A and device B and the few new experiments performed in device B under the new setting. In this case, the transfer problem is therefore complicated by two different issues.

The first issue is related to the difference in the device geometries (sizes of the inlet tubes), which determine a completely different mixing behavior and performances of the devices. Despite the limited size difference, the transfer from device A to device B can increase the productivity thanks to the larger amount of material processed in device B (at the same pressure drop and therefore at the same operating costs) or for the same productivity can significantly reduce the pressure drops (see Lince *et al.*, 2011a). Criteria to scale the production from one device to the other have been suggested based on dimensionless numbers, such as the Reynolds number and the Damkhöler number (Valente *et al.*, 2012), or on the estimation of multi-scale mixing times (Lince *et al.*, 2011a). Although in many cases definitive conclusions are difficult to drawn due to the complexity of the phenomena involved and the inherent experimental uncertainties, it seems that the Reynolds numbers of the inlet jets can be used for scale-up only when the device geometry and size ratios are maintained. On the other hand, relying on multi-scale mixing time principles requires the estimation of the characteristic time-scales involved in the process, which may be quite complicated if appropriate tools are not available (e.g., computational fluid dynamics; Lince *et al.*, 2011a). Therefore, the complexity of the physical phenomena occurring in the mixers, the limitations to their predictability in different devices and the current lack of a fully predictive model (Di Pasquale *et al.*, 2012) justify the use of multivariate statistical approaches (like LVRMs) to guide the transfer and gain a better understanding of the process providing useful insights for mechanistic model development.

The second issue to take into account in the transfer is that, for the target device B, two different datasets referred to experimental campaigns carried out at different times and under a slightly different experimental setup are available. This complicates the understanding and the quantification of the operating parameters effects on the mean nanoparticle size, particularly if the two experimental campaigns carried out in device B were merged. The difference observed in the mean sizes of the nanoparticles obtained in the two campaigns suggests instead to analyze the relevant datasets separately, as if they came from two different devices (or sites).

In light of the above, a procedure to jointly analyze all the available data to draw information for product transfer is proposed in the following. The procedure is based on two main steps. In the first step, a multi-site JY-PLS model (García-Muñoz *et al.*, 2005) is built considering the three operating parameter datasets ($\mathbf{X}_\mathrm{A}$, $\mathbf{X}_\mathrm{B}$, $\mathbf{X}_\mathrm{C}$) separately, and joining them through the common space generated by the nanoparticle diameter datasets ($\mathbf{Y}_\mathrm{A}$, $\mathbf{Y}_\mathrm{B}$, $\mathbf{Y}_\mathrm{C}$). This model allows describing the relationships between operating parameters and particle diameter, which is proper of each device and experimental setup, while relating at the same time the identity of the datasets coming from each device. In the second step, the JY-PLS model is used within the LVRM inversion framework proposed in Chapter 4, in order to estimate the optimal operating conditions to be used in device B to manufacture nanoparticles of desired mean size.

6.3.1 Multi-site JY-PLS

In order to optimally exploit the information available from the different devices and from experiments carried out under a different experimental setup, JY-PLS (García-Muñoz *et al.*, 2005) has been applied.

From a practical perspective, the rationale behind JY-PLS is that, if similar products are produced in different plants (i.e. sites) exploiting the same chemical and physical process, the product properties should share a common correlation structure, represented by their latent variables. Assuming that the regressors are correlated with the product quality within each plant (within-plant correlation), the latent spaces of the regressor datasets will span a common region, represented by the latent space of the product properties (or a subset of it). This latent space can therefore be used to relate regressor datasets from different sources (between-plant correlation).

In the system under investigation, the product quality space is univariate, because it is represented only by the mean particle size d_p. This variable identifies the only direction along which the latent structures of the regressor datasets of different devices should be aligned in order to be related. Considering the issues mentioned earlier, the three available datasets should be analyzed separately to explore the latent structures typical of each device/experimental setup.

The three datasets were therefore organized as in Figure 6.2, and a multi-site JY-PLS strategy was implemented to model the relationships between them (García-Muñoz *et al.*, 2005).

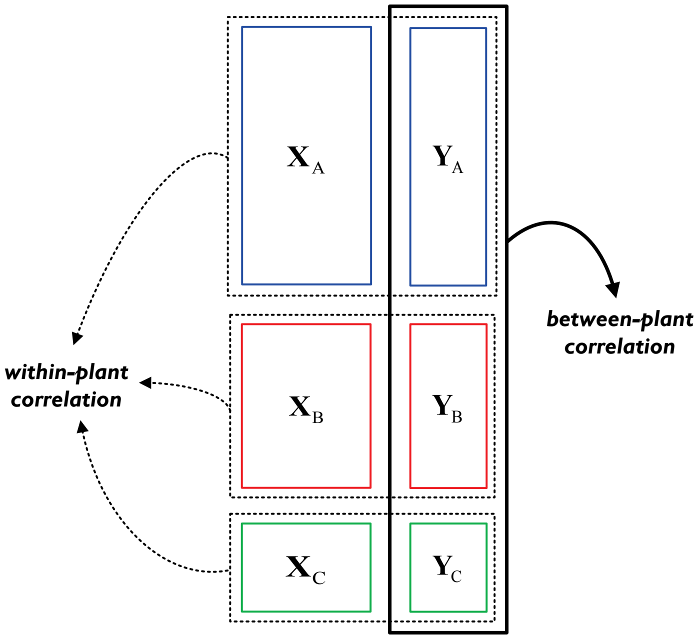


Figure 6.2. *Schematic of the multi-site JY-PLS approach applied for the product transfer problem in this study.*

Accordingly, multi-site JY-PLS decomposes the datasets for each site into their latent structures and overlaps them with the direction identified by the matrix $\mathbf{Y}_\mathrm{J}$ of the joint mean nanoparticle size datasets:

$$\mathbf{Y}_\mathrm{J} = \begin{bmatrix} \mathbf{Y}_\mathrm{A} \\ \mathbf{Y}_\mathrm{B} \\ \mathbf{Y}_\mathrm{C} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_\mathrm{A} \\ \mathbf{T}_\mathrm{B} \\ \mathbf{T}_\mathrm{C} \end{bmatrix} \mathbf{Q}_\mathrm{J}^\mathrm{T} + \mathbf{E}_{\mathbf{Y}_\mathrm{J}} \tag{6.1}$$

$$\mathbf{X}_\mathrm{A} = \mathbf{T}_\mathrm{A}\mathbf{P}_\mathrm{A}^\mathrm{T} + \mathbf{E}_{\mathbf{X}_\mathrm{A}} \tag{6.2}$$

$$\mathbf{X}_\mathrm{B} = \mathbf{T}_\mathrm{B}\mathbf{P}_\mathrm{B}^\mathrm{T} + \mathbf{E}_{\mathbf{X}_\mathrm{B}} \tag{6.3}$$

$$\mathbf{X}_\mathrm{C} = \mathbf{T}_\mathrm{C}\mathbf{P}_\mathrm{C}^\mathrm{T} + \mathbf{E}_{\mathbf{X}_\mathrm{C}} \tag{6.4}$$

$$\mathbf{T}_\mathrm{A} = \mathbf{X}_\mathrm{A}\mathbf{W}_\mathrm{A}^* \tag{6.5}$$

$$\mathbf{T}_\mathrm{B} = \mathbf{X}_\mathrm{B}\mathbf{W}_\mathrm{B}^* \tag{6.6}$$

$$\mathbf{T}_\mathrm{C} = \mathbf{X}_\mathrm{C}\mathbf{W}_\mathrm{C}^* \quad , \tag{6.7}$$

where the meaning of the symbols is the same as described in Chapter 2 for the classical two-sites model (Section 2.1.3.2). As stated above, to apply a JY-PLS model it should be first verified that the different product properties datasets share a common correlation structure (García-Muñoz *et al.*, 2005). Note that, for the system under investigation, this is not necessary, since $\mathbf{Y}_\mathrm{J}$ is univariate, thus identifying a unique direction of variability (i.e. a single LV), which the latent spaces of the regressors datasets are aligned with.

It must be emphasized that the JY-PLS modeling approach has been shown to be particularly useful for transfer activities compared to other statistical methodologies, especially when the latent structures of the target site/plant/device are not fully observable from the available data (e.g., because the number of available data is not adequate to describe it) or the LVs effect is different in the modeled sites (García-Muñoz, 2004). This can be due to differences in the process parameters between the plants or to the different effect the process parameters may have on the LVs (and on the response variables), if the same parameters are considered in the different plants (Liu *et al.*, 2011b). For the system under investigation, previous studies have demonstrated the different importance of the operating parameters on the different devices used (Lince *et al.*, 2008 and 2011a). This consideration, together with the known differences in the experimental setup between old and recent experimental campaigns and the reduced number of data available from the experiments in the target device, justifies the use of JY-PLS to support the transfer over other methodologies.

6.3.2 JY-PLS inversion for product transfer

Once the JY-PLS model has been built between the different datasets, it can be used to support product transfer, namely to estimate the process conditions $\mathbf{x}_\mathrm{C}^\mathrm{NEW}$ (4×1) for device B that, according to the model, provide nanoparticles of desired mean size d_p. To this purpose, model inversion can be applied.

Assuming that $\mathbf{y}^{\mathrm{DES}}$ $(M \times 1)$ is a generic set of M desired product properties and that values are assigned to all the M elements of $\mathbf{y}^{\mathrm{DES}}$ (i.e., only equality constraints are set on $\mathbf{y}^{\mathrm{DES}}$), a JY-PLS model can be directly inverted to reconstruct $\hat{\mathbf{x}}_{\mathrm{C}}^{\mathrm{NEW}}$ (Jaeckle and MacGregor, 1998):

$$\hat{\mathbf{t}}^{\mathrm{DES}} = \left(\mathbf{Q}_{\mathrm{J}}^{\mathrm{T}}\mathbf{Q}_{\mathrm{J}}\right)^{-1}\mathbf{Q}_{\mathrm{J}}^{\mathrm{T}}\mathbf{y}^{\mathrm{DES}} \tag{6.8}$$

$$\hat{\mathbf{x}}_{\mathrm{C}}^{\mathrm{NEW}} = \mathbf{P}_{\mathrm{C}}\hat{\mathbf{t}}^{\mathrm{DES}} \quad , \tag{6.9}$$

being $\hat{\mathbf{t}}^{\mathrm{DES}}$ the $(A \times 1)$ score vector of the direct inversion solution $\hat{\mathbf{x}}_{\mathrm{C}}^{\mathrm{NEW}}$. As stated in Chapter 4, direct model inversion does not always provide a viable solution, for example when constraints are set for the solution $\mathbf{x}_{\mathrm{C}}^{\mathrm{NEW}}$ or when inequality constraints are assigned to the elements of $\mathbf{y}^{\mathrm{DES}}$. In fact, some of these constraints may not be satisfied through direct inversion. Furthermore, in the presence of a null space (Chapter 2, Section 2.2.1), the model inversion has infinite solutions.

The possible presence of the null space and of constraints in either the regressor or the product quality space requires solving the JY-PLS inversion problem within an optimization framework. To this end, the general framework for LVRM inversion proposed in Chapter 4 has been extended in this study to consider the inversion of a JY-PLS model. Namely, the most general scenario of the framework (Scenario 4; Section 4.2.1.2) has been considered and reported in Eq.(6.10) for a general case in which the transfer is intended between a plant A and a plant B:

$$\begin{aligned}
&\min_{\mathbf{x}_{\mathrm{B}}^{\mathrm{NEW}}}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{\mathrm{NEW}} - \mathbf{y}^{\mathrm{DES}}\right)^{\mathrm{T}}\boldsymbol{\Gamma}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{\mathrm{NEW}} - \mathbf{y}^{\mathrm{DES}}\right) + g_1 \cdot \left(\sum_{a=1}^{A}\frac{t_a^2}{s_a^2}\right) + g_2 \cdot \mathrm{SPE}_{\mathbf{x}_{\mathrm{B}}^{\mathrm{NEW}}} \\
&s.t. \\
&\hat{\mathbf{y}}_{\mathrm{B}}^{\mathrm{NEW}} = \mathbf{Q}_{\mathrm{J}}\mathbf{t} \\
&\hat{\mathbf{x}}_{\mathrm{B}}^{\mathrm{NEW}} = \mathbf{P}_{\mathrm{B}}\mathbf{t} \\
&\mathbf{t} = \mathbf{W}_{\mathrm{B}}^{*\mathrm{T}}\mathbf{x}_{\mathrm{B}}^{\mathrm{NEW}} \\
&\mathrm{SPE}_{\mathbf{x}_{\mathrm{B}}^{\mathrm{NEW}}} = \left(\hat{\mathbf{x}}_{\mathrm{B}}^{\mathrm{NEW}} - \mathbf{x}_{\mathrm{B}}^{\mathrm{NEW}}\right)^{\mathrm{T}}\left(\hat{\mathbf{x}}_{\mathrm{B}}^{\mathrm{NEW}} - \mathbf{x}_{\mathrm{B}}^{\mathrm{NEW}}\right) \le g_3 \cdot \mathrm{SPE}_{\mathbf{X}_{\mathrm{B}},95\%\,\mathrm{lim}} \\
&\hat{y}_{j,\mathrm{B}}^{\mathrm{NEW}} \le b_j \\
&x_{r,\mathrm{B}}^{\mathrm{NEW}} = c_r \\
&x_{f,\mathrm{B}}^{\mathrm{NEW}} \le d_f \\
&lb_k^y \le \hat{y}_{k,\mathrm{B}}^{\mathrm{NEW}} \le ub_k^y \qquad lb_l^x \le x_{l,\mathrm{B}}^{\mathrm{NEW}} \le ub_l^x
\end{aligned} \quad . \tag{6.10}$$

Since the objective of the transfer is the estimation of the process conditions for plant B, only the relevant model parameters ($\mathbf{W}_\mathrm{B}^*$, $\mathbf{P}_\mathrm{B}$) appear in the inversion problem, together with the joint loadings $\mathbf{Q}_\mathrm{J}$. The meaning of the rest of the notation in Eq.(6.10) is the same as used in Chapter 4 (Eq.(4.4)), but referred to plant B.

As shown in Chapter 4 (Section 4.4), the use of the soft constraint for $\hat{\mathbf{y}}_\mathrm{B}^\mathrm{NEW}$ as in Eq.(6.10) does not ensure that the solution of the model inversion satisfies the equality constraints assigned for the desired value of the product quality. Therefore, the problem in Eq.(6.10) has been modified, for its application to the present case study, by setting the equality constraint for the desired value of d_p ($\mathbf{y}^\mathrm{DES}$) as a hard rather than a soft constraint. This is also possible because the desired product property set $\mathbf{y}^\mathrm{DES}$ is univariate and there are not drawbacks due to the need of simultaneously satisfying the equality constraints for $\mathbf{y}^\mathrm{DES}$, while adhering to the covariance structure of the historical product data (as if $\mathbf{y}^\mathrm{DES}$ were multivariate; see Section 4.4). Accordingly, the problem in Eq.(6.10) has been reformulated and adapted to the present case study, where the symbols related to the mean particle size d_p are indicated in italics ($\hat{y}_\mathrm{C}^\mathrm{NEW}$, y^DES), being the response variable univariate:

$$\begin{aligned}
&\min_{\mathbf{x}_\mathrm{C}^\mathrm{NEW}}\left[g_1\cdot\left(\sum_{a=1}^{A}\frac{t_a^2}{s_a^2}\right)+g_2\cdot\mathrm{SPE}_{\mathbf{x}_\mathrm{C}^\mathrm{NEW}}\right]\\
&s.t.\\
&\hat{y}_\mathrm{C}^\mathrm{NEW}=y^\mathrm{DES}\quad\text{or}\quad\hat{y}_\mathrm{C}^\mathrm{NEW}\le b\\
&\hat{y}_\mathrm{C}^\mathrm{NEW}=\mathbf{Q}_\mathrm{J}\mathbf{t}\\
&\hat{\mathbf{x}}_\mathrm{C}^\mathrm{NEW}=\mathbf{P}_\mathrm{C}\mathbf{t}\\
&\mathbf{t}=\mathbf{W}_\mathrm{C}^{*\mathrm{T}}\mathbf{x}_\mathrm{C}^\mathrm{NEW}\\
&\mathrm{SPE}_{\mathbf{x}_\mathrm{C}^\mathrm{NEW}}=\left(\hat{\mathbf{x}}_\mathrm{C}^\mathrm{NEW}-\mathbf{x}_\mathrm{C}^\mathrm{NEW}\right)^\mathrm{T}\left(\hat{\mathbf{x}}_\mathrm{C}^\mathrm{NEW}-\mathbf{x}_\mathrm{C}^\mathrm{NEW}\right)\le g_3\cdot\mathrm{SPE}_{\mathbf{X}_\mathrm{C},95\%\,\mathrm{lim}}\\
&x_{r,\mathrm{C}}^\mathrm{NEW}=c_r\\
&x_{f,\mathrm{C}}^\mathrm{NEW}\le d_f\\
&lb^y\le\hat{y}_\mathrm{C}^\mathrm{NEW}\le ub^y\qquad lb_l^x\le x_{l,\mathrm{C}}^\mathrm{NEW}\le ub_l^x
\end{aligned}\quad. \tag{6.11}$$

It can be seen in Eq.(6.11) that the parameters of the model referred to the target device *in the new experimental setup* ($\mathbf{P}_\mathrm{C}$, $\mathbf{W}_\mathrm{C}^*$) are used. Comparing the problem in Eq.(6.11) to problem in Eq.(6.10), the first term of the objective function (soft constraint) has been deleted in Eq.(6.11), and the equality constraint for the desired mean particle size has been set as a hard constraint ($\hat{y}_\mathrm{C}^\mathrm{NEW}=y^\mathrm{DES}$). This is used as an alternative to the inequality constraint $\hat{y}_\mathrm{C}^\mathrm{NEW}\le b$, which is used if the mean particle size is required to be below an assigned threshold b.

6.4 Results and discussion

This Section reports the results of the application of the proposed product transfer methodology to the process described in Section 6.2. Results are organized in three sub-sections. First, the multi-site JY-PLS model built on the available datasets is presented, together with its diagnostics and parameter interpretation. The model is then inverted to design the process conditions of device B in the new experimental setup for two different problems. In the first problem, the objective is to obtain in device B nanoparticles of an assigned mean size d_p (i.e. an equality constraint is set for it). Since the existence of a null space can be postulated, different operating conditions are estimated and tested in order to experimentally verify its existence. In the second problem, the objective is to obtain in device B nanoparticles whose diameter is below an assigned threshold, for their final application as carriers. In both problems, the conditions estimated *in-silico* through the JY-PLS model inversion are tested experimentally to validate the procedure.

6.4.1 Model design, diagnostics and interpretation

The first step of the procedure is to build the JY-PLS model on the available datasets. Table 6.2 and Table 6.3 report the diagnostics of the model in calibration (Table 6.2) and cross-validation (Table 6.3). Namely, Table 6.2 presents the explained variance (R^2) and the cumulative explained variance (R^2_{CUM}) per LV and per dataset considered in model design. As can be seen, for the response variable matrices (**Y**'s) the variances explained after the first LV are not significant, which is expected since the response datasets are univariate. Differently, from the analysis of the variances explained for the regressor datasets, it can be noted that at least 3 LVs are needed to adequately describe the systematic variability in the $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{C}}$ datasets, whereas the $\mathbf{X}_{\mathrm{B}}$ dataset variability is *fully* described with 3 LVs, due to the artificial collinearity introduced by experimentation between the polymer concentration and the polymer type variables. These results indicate that a null space is present, due to the different rank of the regressor and of the response variable matrices.

Table 6.2. *Diagnostics of the JY-PLS model. Explained variance (R^2) and cumulative explained (R^2_{CUM}) variance of the considered datasets per LV in model calibration.*

LV	$R^2\mathbf{X}_{\mathrm{A}}$	$R^2_{\mathrm{CUM}}\mathbf{X}_{\mathrm{A}}$	$R^2\mathbf{Y}_{\mathrm{A}}$	$R^2_{\mathrm{CUM}}\mathbf{Y}_{\mathrm{A}}$	$R^2\mathbf{X}_{\mathrm{B}}$	$R^2_{\mathrm{CUM}}\mathbf{X}_{\mathrm{B}}$	$R^2\mathbf{Y}_{\mathrm{B}}$	$R^2_{\mathrm{CUM}}\mathbf{Y}_{\mathrm{B}}$	$R^2\mathbf{X}_{\mathrm{C}}$	$R^2_{\mathrm{CUM}}\mathbf{X}_{\mathrm{C}}$	$R^2\mathbf{Y}_{\mathrm{C}}$	$R^2_{\mathrm{CUM}}\mathbf{Y}_{\mathrm{C}}$
1	0.293	0.293	0.548	0.548	0.465	0.465	0.712	0.712	0.232	0.232	0.658	0.658
2	0.261	0.554	0.062	0.610	0.288	0.754	0.056	0.768	0.217	0.449	0.032	0.690
3	0.196	0.750	0.001	0.611	0.246	1.000	~0	0.768	0.364	0.813	7e-4	0.691
4	0.250	1.000	~0	0.611	0	1.000	~0	0.768	0.187	1.000	~0	0.691

To get a better indication of the number of LVs to use in order to build the JY-PLS model, cross-validation has been applied with a jackknife approach (Duchesne and MacGregor, 2001). The cross-validation has been performed considering the three different "portions" of the JY-PLS model structure (Figure 6.2) as three different PLS models, thus cross-validating each portion separately (García-Muñoz, 2004). In Table 6.3, the variances explained by the model in cross-validation per LV and per dataset are reported. Namely, diagnostics are reported both for the regressors (P^2 and P^2_{CUM}; Chapter 4, Section 4.3.1) and for the response (Q^2 and Q^2_{CUM}) datasets.

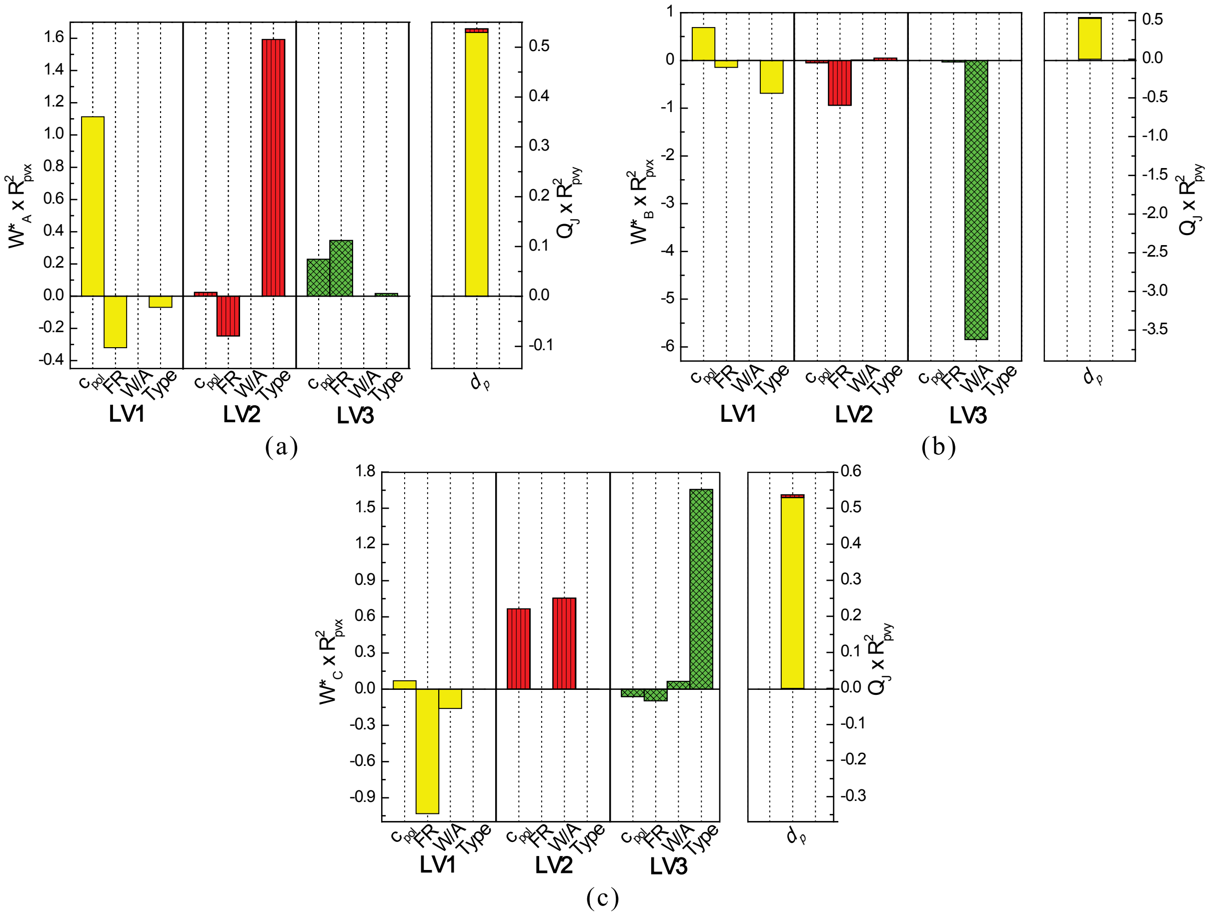
The analysis of the results reported in Table 6.3 confirms what was observed from the model calibration diagnostics. Given that the objective of the product transfer is the estimation of the process conditions for device B in the new experimental setup, it is important to select a number of LVs that allows to adequately describe the variability of the $\mathbf{X}_{\mathrm{C}}$ and $\mathbf{Y}_{\mathrm{C}}$ datasets. On the basis of the P^2 values (and in particular of value of $P^2\mathbf{X}_{\mathrm{C}}$), three LVs were selected to build the JY-PLS model. Given that the $\mathbf{Y}_{\mathrm{J}}$ space is univariate, this means that there exists a bi-dimensional null space that has to be considered in the inversion of the model, to estimate the process conditions for device B.

Table 6.3. *Diagnostics of the JY-PLS model. Explained variance and cumulative explained variance of the regressor (P^2, P^2_{CUM}) and the response (Q^2, Q^2_{CUM}) datasets per LV in model cross-validation.*

LV	$P^2\mathbf{X}_{\mathrm{A}}$	$P^2_{\mathrm{CUM}}\mathbf{X}_{\mathrm{A}}$	$Q^2\mathbf{Y}_{\mathrm{A}}$	$Q^2_{\mathrm{CUM}}\mathbf{Y}_{\mathrm{A}}$	$P^2\mathbf{X}_{\mathrm{B}}$	$P^2_{\mathrm{CUM}}\mathbf{X}_{\mathrm{B}}$	$Q^2\mathbf{Y}_{\mathrm{B}}$	$Q^2_{\mathrm{CUM}}\mathbf{Y}_{\mathrm{B}}$	$P^2\mathbf{X}_{\mathrm{C}}$	$P^2_{\mathrm{CUM}}\mathbf{X}_{\mathrm{C}}$	$Q^2\mathbf{Y}_{\mathrm{C}}$	$Q^2_{\mathrm{CUM}}\mathbf{Y}_{\mathrm{C}}$
1	0.282	0.282	0.520	0.520	0.435	0.435	0.679	0.679	0.098	0.098	0.413	0.413
2	0.261	0.543	0.075	0.595	0.295	0.730	0.079	0.758	0.327	0.425	0.150	0.563
3	0.207	0.750	0.008	0.603	0.270	1.000	-0.012	0.745	0.422	0.847	0.120	0.683
4	0.250	1.000	-0.003	0.600	0	1.000	0.022	0.767	0.153	1.00	0.008	0.691

In Figure 6.3, the weights of the JY-PLS model per LV are reported as bar plots together with the joint loadings for the response variable d_p ($\mathbf{Q}_{\mathrm{J}}$) for device A (Figure 6.3a), device B in the old experimental setup (Figure 6.3b), and device B in the new experimental setup (Figure 6.3c). Both the weights and the loadings have been reported per LV and weighted according to the variance explained by the model per original variable (R^2_{pvx} and R^2_{pvy}). The joint loadings $\mathbf{Q}_{\mathrm{J}}$ are of course the same in each plot, even if they appear different for scaling reasons.

The plots in Figure 6.3 are particularly useful to gain understanding on the physics of the system and on the effect of the different process parameters on the mean nanoparticle size in the different devices and settings. Figure 6.3 clarifies that the impact of the operating parameters is quite different in the different devices, although the general trend is the same. To understand the effects on d_p, especially the weights of the variables on LV1 have to be considered (see the $R^2\mathbf{Y}$ values in Table 6.2). As can be seen, the main variables affecting the nanoparticle mean size are the polymer concentration (c_{pol}) and the water flowrate (*FR*). In particular, the higher the polymer concentration, the larger is expected to be the mean nanoparticle size. This can be inferred from the positive weight c_{pol} has on LV1 in all the bar plots of Figure 6.3, which is concordant with the joint loading of d_p on LV1 (meaning that there is a positive correlation between them). The effect of *FR* is the opposite instead: in all plots, *FR* has a negative weight on LV1, which is opposite to the d_p joint loading on LV1. This means that larger nanoparticles are obtained at lower flowrates. The effects of the polymer type (*Type*) and of the anti-solvent/solvent ratio (*W/A*) generally seem to be less significant on LV1, being their weights on LV1 much lower (shorter bars).



(a) (b)

(c)

Figure 6.3. *JY-PLS model weights and of the joint-Y loadings* $\mathbf{Q}_J$ *per LV for (a) device A, and device B in the (b) old experimental setup and (c) new experimental setup.*

While this analysis provides an idea of the general effect of the operating variables on d_p, from a detailed analysis of each plot it can also be concluded that the importance of the operating variables in each device is very different. Figure 6.3a suggests that the dominant driving force in device A is due to c_{pol}, and that *FR* has a lower impact on d_p. Additionally, an indication on the effect of the polymer type can be drawn. In fact, from the value of the weight of *Type* it seems that, when this variable assumes a "low" value (a negative bar length means *Type* = 0, i.e. PCL_{14} is used), the nanoparticles obtained in device A are larger. The weights on LV2 explain the variability in the data due to *Type* and *FR*. Whereas the effect of *FR* is the same as described for LV1, the weight of *Type* seems to give contrasting information compared to what was concluded from LV1. However, it must be noted that the

LV weights are orthogonal, thus representing independent effects. Therefore, the weights on LV2 are not providing conflicting information to those on LV1, but are indicating a second driving force, which explains the second highest part of the variability in the data. This is mainly due to the combination of the polymer type and *FR*: a (very little) part of the variability in d_p (on LV2; 6.2%, Table 6.2) is due to the experiments performed with PCL_{80} (*Type* = 1) and low *FR*, which provide nanoparticles of size larger than the historical mean. The weights on LV3 accounts for a third part of variability, which is mainly in the regressor dataset, being LV3 not significant at all for d_p (0.1%, Table 6.2).

From Figure 6.3b it can be seen that the most important variables for the process in device B with the original experimental setup are c_{pol} and the type of polymer used. However, it must be noted that, as mentioned above, a collinearity between these two variables was forced in the historical data from device B, due to the way in which the experiments were carried out. This collinearity is described by the weights of c_{pol} and *Type* on LV1 in Figure 6.3b (which are opposite, meaning inverse correlation between them). As a consequence, the effect of these two variables on d_p is confounded, even if it is likely that the c_{pol} effect is prevailing. The variability due to *FR* is described mainly by LV2 and it looks less important than c_{pol}. It is also interesting to note that *W/A* shows a significant weight on LV3, which is however not significant to describe d_p (see Table 6.2). In this case, LV3 is useful only to describe the variability in $\mathbf{X}_{\mathrm{B}}$ due to *W/A*.

Finally, from Figure 6.3c it can be noted that the latent structure of $\mathbf{X}_{\mathrm{C}}$ (i.e., of the data from device B in the new experimental setup) is quite different both from the one of $\mathbf{X}_{\mathrm{A}}$ and (especially) from that of $\mathbf{X}_{\mathrm{B}}$. This justifies the separate analysis of the datasets B and C, and the use of the JY-PLS model. The first important thing to note is that the most important variable on LV1 is the water flowrate (*FR*), differently from the other datasets. It is known that the dependence of d_p from *FR* is strong at low water flowrates, whereas d_p and *FR* are substantially unrelated at high flowrates (Lince *et al.*, 2011a). Due to the different inlet diameters (and therefore to the different inlet jet velocities), the increased importance of *FR* in device B (with the new experimental setup) may be caused by the fact that in this device the relationship between *FR* and d_p is strong on a wider range of flowrates than in device A. This could not be seen from the analysis of the $\mathbf{X}_{\mathrm{B}}$ dataset only, as LV1 was biased by the artificially introduced collinearity between c_{pol} and *Type*.

Furthermore, in Figure 6.3c *W/A* is found to have an impact on LV1 (hence on d_p), whereas c_{pol} is poorly described by LV1. The effect of *FR* and c_{pol} follows the trend observed earlier, but it is interesting to note that *W/A* is found to be inversely related to d_p. This means that the nanoparticle mean size is expected to decrease at higher *W/A* values; under these conditions in fact less solvent is mixed with the same amount of anti-solvent, probably inducing the formation of smaller nanoparticles. It should be mentioned however that the effect of *W/A* on these systems has not been completely clarified on a physical basis (Lince *et al.*, 2011a;

Valente *et al*., 2012), since under *W*/*A* values different from unity poor mixing conditions are generally obtained, resulting in complicated effects of the final nanoparticle size. However, its effect should be read in light of the polymer concentration: at high polymer concentration, the nanoparticle size decreases at higher *W/A*, probably due to the achievement of supersaturation conditions, which are favored if less solvent and more concentrated solutions are introduced as already mentioned, while at very low polymer concentration the role of the mixing efficiency may prevail. This would explain the pattern of the weights on LV1 in Figure 6.3c and its relation to the joint loading of d_p. At the same time, the weights on LV2 account for the second (although minor; $R^2\mathbf{Y}_\mathrm{C}$ = 3.2% for LV2, Table 6.2) driving force in the data, which is due to the variability in c_{pol} and *W/A*. Finally, LV3 describes mainly the variability associated to the type of polymer, that in this dataset seems to be not significant for d_p.

This first analysis already shows the utility of this multivariate statistical approach, as some of these conclusions would be very difficult to be drawn following a one-factor-at-a-time experimental strategy. In the following, the JY-PLS model described in this Section is inverted to design the process conditions in plant B to produce nanoparticles of a desired assigned mean size (Problem 1) or with a size below an assigned threshold of interest (Problem 2).

6.4.2 Problem 1: transfer results and null space validation

The objective is to manufacture nanoparticles of mean size $y^{\mathrm{DES}} = 280$ nm in device B under the new experimental setup. Nanoparticles of this size were already obtained in the experimental campaigns performed in device A and in device B with the old experimental setup (datasets A and B, respectively). The desired size is well within the ranges of the historical data (Table 6.1), and the JY-PLS model can be feasibly used to support the design of the process conditions to obtain the nanoparticles.

Table 6.4 reports the results obtained by direct inverting the JY-PLS model through Eqs.(6.8)-(6.9).

Table 6.4. *Problem 1; operating conditions in device B determined by direct inversion of the JY-PLS model to obtain nanoparticles of mean size* y^{DES} = *280 nm. The 95% confidence limits are:* $T^2_{95\%\,\mathrm{lim}}$ = *11.46;* $\mathrm{SPE}_{95\%\,\mathrm{lim}}$ = *4.37e-2.*

	c_{pol} [mg/mL]	*FR* [mL/min]	*W/A*	*Type*	T^2	SPE
$\hat{\mathbf{x}}_\mathrm{C}^{\mathrm{NEW}}$	3.2	53	2.27	0.58	0.62	0

The direct inversion solution provides realistic results for c_{pol} and *FR*, but the value of the variable *Type* is meaningless (this variable is binary). Therefore the value of this variable has to be assigned, unless resorting to optimization algorithms for mixed-integer problems. Furthermore, due to limitations of the experimental apparatus, *W/A* can assume values within a given interval. This provides an additional constraint to the problem, which the direct inversion solution in Table 6.4 cannot satisfy. Table 6.4 reports also the T^2 and SPE values for the solution, from which it can be concluded that the solution lies onto the space of the LVs of the model (SPE = 0) and is quite close to the mean of the nanoparticle sizes included in dataset $\mathbf{Y}_\mathrm{C}$ ($T^2 = 0.62$).

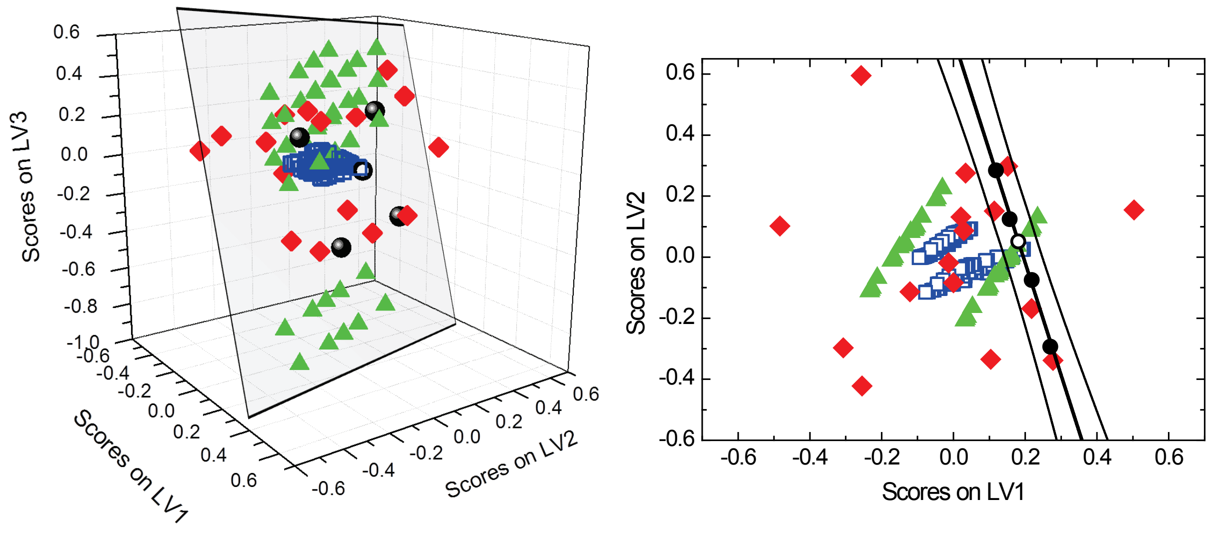
In addition to the issue on constraints, the existence of a bi-dimensional null space should be considered in the inversion. Due to the null space, the model inversion problem has infinite solutions, all of which (according to the model) correspond to the same y^DES. To calculate the best set of process conditions $\mathbf{x}_\mathrm{C}^\mathrm{NEW}$ along the null space (i.e. those minimizing the objective function and satisfying at the same time all constraints), the optimization problem in Eq.(6.11) was solved.

First, the very existence of the null space was validated experimentally. This was done by evaluating different solutions to the JY-PLS model inversion problem along the null space, and performing in device B the experiments suggested by these solutions. To move the solutions along the null space, the problem in Eq.(6.11) was solved by setting equality constraints to variables *Type* and *W/A* (i.e. by assigning the polymer and the anti-solvent/solvent ratio to be used in the experiments), and changing the constraint values in order to generate a set of solutions. Furthermore, $g_1 = 0$ was set, in order to prevent the optimizer from moving the solutions towards the origin of the historical data score space. The following boundaries (inequality constraints) were set for the other variables:

- polymer concentration: $0.026 < c_{pol} < c_{pol,max}$;
- water flowrate: $3 < FR < 120$ mL/min.

These ranges represent the experimental domain defined by the physical limits of the experimental apparatus (for *FR*) and by the historical data range ($c_{pol,max}$ = 6.17 mg/mL for PCL$_{80}$, and $c_{pol,max}$ = 24.83 mg/mL for PCL$_{14}$,).

Four different operating conditions sets $\mathbf{x}_\mathrm{C}^\mathrm{NEW}$ were then calculated using this strategy. Figure 6.4 shows the representation of the null space, as calculated from the $\mathbf{Q}_\mathrm{J}$ loadings through singular value decomposition (Jaeckle and MacGregor, 2000b; Chapter 2, Section 2.2). Namely, Figure 6.4a shows the representation of the bi-dimensional null space in the three-dimensional score space of the JY-PLS model. The projections of the points estimated by optimization along the null space are reported (●), together with the scores of the direct inversion solution (○), and of the historical data used to build the model and included in $\mathbf{X}_\mathrm{A}$ (□), $\mathbf{X}_\mathrm{B}$ (▲) and $\mathbf{X}_\mathrm{C}$ (◆). For the sake of clarity, Figure 6.4b reports the score space on the first two LVs of the model. The meaning of the symbols is the same as in Figure 6.4a, but the null space is represented by the thick line, which represents the intersection between the bi-dimensional null space and the plane of the first two LV scores. The relevant 95% confidence limits for the null space are reported as thin lines. These limits have been calculated through a bootstrapping algorithm, following the procedure described in Appendix C.



(a) (b)

Figure 6.4. *Problem 1; null space validation. Representation of the null space for y^{DES} = 280 nm, and of the projections (●) on the score space of the JY-PLS model of the operating conditions in device B as estimated by model inversion along the null space: (a) score space of the 3 LVs of the model and (b) of first 2 LVs of the model. In each plot, the scores of the data in* $\mathbf{X}_A$ *(□),* $\mathbf{X}_B$ *(▲) and* $\mathbf{X}_C$ *(◆) are reported together with the scores of the direct inversion solution (○). The null space is represented in (a) by the gray plane and in (b) by the thick line with the relevant 95% confidence limits (thin lines).*

Table 6.5 reports the four different process operating conditions sets estimated along the null space through the JY-PLS model inversion. As can be seen, a reasonably wide region of the null space could be explored by simply changing the equality constraints for *W/A* and for the polymer type. This can be observed also from the projections of the solution sets into the score space and from the solution Hotelling's T^2 in Table 6.5. Furthermore, despite the constraints assigned to some of the variables, the solution SPEs are very low, meaning that the solutions are quite close to the LV model space, thus improving their reliability.

Table 6.5. *Problem 1, null space validation. Operating conditions in device B determined by inversion of the JY-PLS model to obtain nanoparticles with* y^{DES} *= 280 nm, and comparison with the mean diameters obtained experimentally. Variables W/A and Type assigned as equality constraints. The 95% confidence limits are:* $T^2_{95\%\,\mathrm{lim}}$ *= 11.46;* $\mathrm{SPE}_{95\%\,\mathrm{lim}}$ *= 4.37e–2.*

Run no.	c_{pol} [mg/mL]	*FR* [mL/min]	*W/A*	*Type*	T^2	SPE	d_p^{EXP} [nm]	Error [%]
1	1.5	3	1.00	PCL_{80}	3.16	2.45e-4	289.5	–3.4
2	2.9	24	2.94	PCL_{80}	1.56	4.63e-6	287.4	–2.6
3	4.1	53	1.00	PCL_{14}	2.33	3.13e-4	268.6	+4.1
4	5.0	68	2.84	PCL_{14}	2.40	4.72e-4	247.9	+11.5

To validate the existence of the null space and to verify that the desired nanoparticle size were actually obtained, the operating conditions estimated by JY-PLS model inversion were actually implemented in a series of experiments on device B under the new experimental setup (i.e. the same used to obtain the data included in datasets $\mathbf{X}_{\mathrm{C}}$ and $\mathbf{Y}_{\mathrm{C}}$). The new experimental results are reported in the last two columns of Table 6.5, in terms of mean size of the obtained nanoparticles (d_p^{EXP}) and of percentage error compared to the desired value $y^{\mathrm{DES}} = 280$ nm. As can be seen, the obtained nanoparticles have a mean size very close to the target one, especially for the experiments performed with PCL$_{80}$; slightly larger errors are observed when using PCL$_{14}$. However, the observed errors are well within the repeatability threshold (15%). Hence, it can be concluded that particles of roughly the same size are obtained by running device B at very different operating conditions. These five sets of operating conditions (which include those obtained from direct model inversion) all lie on the same space on the score space (the null space), as Figure 6.4b clearly shows.

Note that the errors for PCL$_{80}$ are both negative, whereas for PCL$_{14}$ they are both positive. Although the number of experimental runs is not sufficient to draw general conclusions, this seems to indicate that a polymer effect does exist, but it is not completely captured by the model. Notwithstanding this, the general trend observed in the data, according to which PCL$_{14}$ is associated with smaller nanoparticles compared to PCL$_{80}$, is captured. Also note that the largest error between experimental and expected values is observed in run 4, when the largest water flowrate *FR* value was used. As mentioned earlier, from first-principles knowledge on the process it is known that the dependence of d_p from the water flowrate is stronger at low values of *FR*, whereas at higher values the effect is less significant. The *FR* value used in run 4 (68 mL/min) is an intermediate value (see Table 6.1), which is representative of a transition zone in the relationship between *FR* and d_p. This may justify the increased error observed for this run. Since the JY-PLS modeling technique is linear, a solution to this issue may consist in the use of a nonlinear transformation for the *FR* variable. Alternatively, an iterative approach could be used, by designing the model after new experiments have been carried out, and performing the inversion with the updated model, until convergence on the desired mean nanoparticle size is reached (García-Muñoz *et al.*, 2005). Local modeling approaches (Dayal and MacGregor, 1997) may also be used to cope with possible nonlinearities.

The results validate experimentally the existence of the null space, and clearly show how different operating conditions along it indeed provide the same desired mean particle size. This can be useful in defining the *design space* of the process, under a QbD framework (Chapter 1, Section 1.2.2), confirming what was stated in Chapter 4 on the link between the null space and the design space concepts (Section 4.3.2.1). Therefore, this has very important implications in pharmaceutical engineering, where the desired product is often defined within very narrow characteristic property windows.

6.4.3 Problem 2: transfer results and validation

Manufacturing nanoparticles whose dimension is smaller than an assigned threshold is a typical requirement when the particles are to be used as drug carriers (Section 6.2). To test the proposed methodology for the solution of this problem, the JY-PLS model inversion approach was applied with the objective of manufacturing nanoparticles of mean size d_p < 190 nm in device B under the new experimental setup. The optimization problem in Eq.(6.11) was solved by setting an inequality constraint for $\hat{y}_{\mathrm{C}}^{\mathrm{NEW}}$, and by assigning $b = 190$ nm as the constraint value. As in Problem 1, the polymer type and the anti-solvent/solvent ratio values were assigned and set as equality constraints in the optimization.

In Table 6.6, three different sets of operating conditions determined by inversion of the JY-PLS model at different values of *W/A* and *Type* have been reported, together with the relevant T^2 and SPE statistics, and with the values of the mean nanoparticle size predicted by the model for the calculated solution sets ($\hat{d}_p$). The experimental conditions obtained by optimization present high values of *FR* and (relatively) low values of c_{pol}. This is not surprising, as the optimizer is asked to find operating conditions suitable to manufacture nanoparticles with small size compared to the historical available data. Therefore, to satisfy the constraints, the optimizer is forced to find solutions near the boundaries of the operating variable domain. In fact, in runs 1 and 3 the calculated value for *FR* hits its upper bound. As a consequence, the values of the SPE statistic are larger than in Problem 1 (although still acceptable), indicating that in order to satisfy the constraints, the solution has to be moved out of the LV model space.

Table 6.6. *Problem 2. Operating conditions in device B determined by inversion of the JY-PLS model to obtain nanoparticles with d_p < 190 nm, and comparison with the mean diameters obtained experimentally. Variables W/A and Type assigned as equality constraints. The 95% confidence limits are: $T^2_{95\%\,\mathrm{lim}}$ = 11.46; $\mathrm{SPE}_{95\%\,\mathrm{lim}}$ = 4.37e–2.*

Run no.	c_{pol} **[mg/mL]**	***FR*** **[mL/min]**	***W/A***	***Type***	T^2	**SPE**	$\hat{d}_p$ **[nm]**	d_p^{EXP} **[nm]**	**Error [%]**
1	3.0	120	1.00	PCL_{80}	1.2	3.28e-2	185.4	183.4	+1.1
2	1.5	83	2.94	PCL_{80}	1.5	7.93e-5	183.1	192.4	–5.1
3	2.4	120	2.94	PCL_{14}	2.8	5.30e-3	161.4	165.8	–2.7

As in Problem 1, the operating conditions calculated by inversion of the JY-PLS model were implemented on device B, under the new experimental setup, to experimentally validate the results. The last two columns of Table 6.6 report the value of the mean size of the nanoparticles obtained experimentally (d_p^{EXP}), together with the error accounting for the difference between model predictions and experimental values. For run 1 and run 3 the experimental values of the nanoparticle mean size satisfy the inequality constraint assigned in the inversion problem ($d_p \leq 190$ nm). For run 2 the experimental value does not satisfy the constraint, but the obtained value for the nanoparticle size (192.4 nm) is very close to the threshold. Nonetheless, the errors between experimental and predicted values are very small. It should be emphasized that at the high *FR* values used in runs 1 and 3 the behavior of the system is more easily predicted by the model, as data at similar flowrates were obtained also from the historical experiments.

Figure 6.5 represents the projections of the operating condition sets reported in Table 6.6 in the space of the scores on the first two LVs of the JY-PLS model. The meaning of the symbols is the same as in Figure 6.4. It can be observed that the estimated process conditions project in the region of negative scores for both LV1 and LV2, which, by considering the weights in Figure 6.3c, corresponds to the region of high *FR*, low c_{pol} and high *W/A* values. The (relatively) small values of the constraints assigned to *W/A* to estimate the sets of Table 6.6 tend therefore to pull the solution projections to the center of the model latent space.

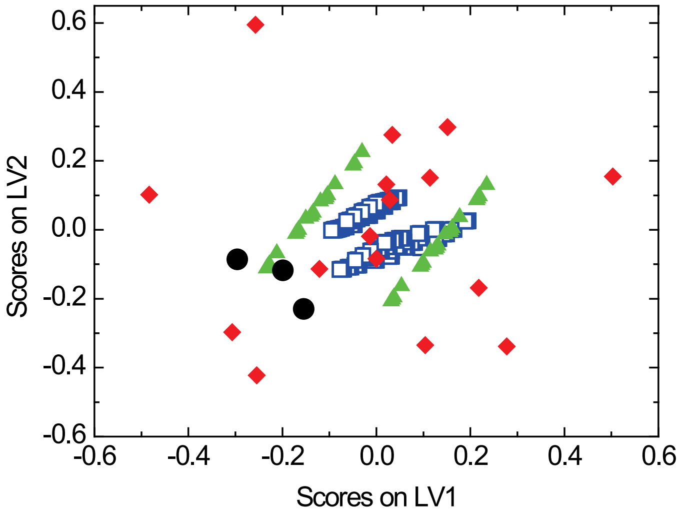


Figure 6.5. *Problem 2. Projections (●) on the score space of the first two LVs of the JY-PLS model of the operating conditions in device B as estimated by model inversion to obtain nanoparticles with d_p < 190 nm. The scores of the data in* $\mathbf{X}_A$ *(□),* $\mathbf{X}_B$ *(▲) and* $\mathbf{X}_C$ *(◆) are reported as well.*

6.5 Conclusions

In this Chapter, the problem of the transfer of a product between different equipment, plants or manufacturing sites has been tackled. This problem is commonly encountered in process development environments, especially when scaling a production between a source plant (usually of small size) to a target plant, where the production is supposed to start (e.g., a large-scale plant). When dealing with transfer problems, a large amount of data is usually available from the experiments performed on the source plant, where the process has been widely studied, while few data are available from the target plant. In this Chapter, a procedure has been proposed to jointly analyze all the available datasets from experimental campaigns already carried out on different products or different plants configurations, in order to suggest, on the basis of the historical knowledge, the most appropriate experimental conditions to test in order to address the transfer.

The procedure is based on the use of a JY-PLS model to relate all the data available from the different sources in a whole modeling framework, in order to identify the common latent structures between the different datasets. The JY-PLS model is then inverted through the general LVRM inversion framework proposed in Chapter 4, in order to suggest the experiments to perform.

The proposed procedure has been applied experimentally to a process involving the preparation of nanoparticles for pharmaceutical applications through a solvent displacement process in passive mixers. The objective was to produce in a target device, where limited data were available, nanoparticles of desired size, by exploiting the data available from a different device of smaller size as well as from the same target device but under a different experimental setup.

Two historical datasets were available, which were related to experiments that had been carried out in both devices. A third dataset, of limited size, was built *ad-hoc*, by performing the experiments in such a way as to cope with the different experimental setup that the target plant experienced over time (mainly due to maintenance operations).

The three datasets were first modeled through JY-PLS to gain understanding on the physics of the systems. The model parameters were interpreted from first-principles, confirming what was known only partially from other studies on the systems and providing very useful new insights on what determines the difference between the devices. The model was then inverted to suggest the optimal experimental sets to implement in the target device to produce nanoparticles of desired mean size.

Two specific problems were studied. In the first one, JY-PLS inversion was used to estimate the conditions in device B with the new experimental setup to manufacture nanoparticles with an assigned mean size. A bi-dimensional null space had to be considered in model inversion. Different process operating conditions were estimated along the null space and experimentally tested. Experiments confirmed the existence of the null space and showed how different process settings were able to provide the same desired mean nanoparticle size, within the experimental uncertainty.

In the second problem, JY-PLS inversion was used to design the experiments in order to obtain nanoparticles with mean size below an assigned threshold. Again, experiments confirmed the effectiveness of the proposed procedure in designing the target device operating conditions in such a way to obtain nanoparticles of assigned size range.

The proposed procedure can be easily extended to problems where the product quality is characterized by a multivariate set of property specifications. It can be feasibly used to support product transfer in product and process development, especially in those industries (as the pharmaceutical one), where transfer activities can be critical in terms of time, resources and regulatory oversight.

Chapter 7

Transfer of process monitoring models between different plants[*]

In this Chapter, a general procedure is proposed based on LVMs to tackle the issue of transferring process monitoring models between different plants. The procedure identifies five different scenarios, which are described and discussed. The proposed methodology is applied on a benchmark problem related to the scale-up of the monitoring model for an industrial continuous spray-drying process. Then, the methodology is extended to batch process monitoring.

7.1 Introduction

When the manufacturing of a product with assigned quality specifications is transferred from a source plant A to a target plant B, it would be highly desirable to have a reliable monitoring system available for plant B as quickly as possible, in order to detect incipient faults and possibly to diagnose them since the beginning of the operation in plant B. Multivariate statistical process control techniques have been applied successfully in several industrial applications for online process monitoring and fault detection (Nomikos and MacGregor, 1994; Wise and Gallagher, 1996; Wold *et al.*, 1998). In order to build a reliable process monitoring model, these techniques require that data representing the common cause variability (CCV) to which the process is subject be available.

In production transfer activities (e.g. plant scale-up), experiments in the target plant B are limited to those needed to define the normal process conditions of the operation, and process data are therefore usually insufficient to build a monitoring model for this plant based on multivariate statistical techniques. Experimental campaigns designed to produce CCV data in plant B are carried out very rarely, especially if the cost of raw materials is high or the product

[*] Tomba, E., P. Facco, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Combining fundamental knowledge and latent variable techniques to transfer process monitoring models between plants. *Chemom. Intell. Lab. Syst.*, **116**, 67-77.

Facco, P., E. Tomba, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Transfer of process monitoring models between different plants using latent variable techniques. *Ind. Eng. Chem. Res.*, **51**, 7327-7339.

Facco, P., M. Largoni, E. Tomba, F. Bezzo and M. Barolo (2013). Transfer of process monitoring models between plants: batch systems. *In preparation*.

manufacturing is subject to a rigid regulatory environment (as in the case of pharmaceutical industries). At the same time, if the product has already been manufactured in plant A, several operating data are usually available from this plant, and a set of normal operating conditions (NOC; MacGregor and Kourti, 1995) may have been identified that guarantee that the product quality meet the specifications with acceptable variability.

It would be therefore useful to transfer the knowledge already available for plant A in order to monitor the manufacturing process in plant B until a sufficient amount of data are collected in this plant to design a process monitoring model entirely based on the plant B data. In this Chapter, a possible strategy to solve this problem (which is referred to as a process monitoring model transfer or simply *model transfer* problem) is presented.

The model transfer issue can be considered as part of the much wider technology transfer problem, which has been partly reviewed in Chapter 5. However, model transfer is fundamentally different from the product transfer problem that was considered in Chapter 5 and was based on LVRM inversion (Jaeckle and MacGregor, 2000b; García-Muñoz *et al.*, 2005).

So far, the model transfer issue has been investigated mainly with reference to instrument calibration models, in particular in spectroscopy (Feudale *et al.,* 2002). The underlying idea of calibration model transfer approaches is that if the same sample is analyzed using different instruments, there should be a correspondence between the spectra measured in each instrument. However, transfer approaches developed for these models are not suitable for the transfer of process monitoring systems, since it is hard or even impossible to find a correspondence between samples coming from different plants. Methodologies for transferring a model to a new process have been recently proposed by Lu and coworkers (2008a, 2008b and 2009). Although these procedures are effective, they basically refer to the transfer of predictive models (e.g., soft sensors) rather than to the transfer of monitoring models, and therefore are not appropriate for the problem under investigation.

A first contribution to the transfer of monitoring models was presented by Chiang and Colegrove (2007), who implemented a method to monitor the quality of products manufactured in different plants and with different production targets, but showing similar correlation among the quality variables. Even if the procedure is very effective, it is applied for product quality control, thus not considering the online process measurements.

The complexity of the model transfer problem arises from the fact that several issues intersecting with each other need to be accounted for when transferring a monitoring model between plants. First, one should consider the type of information initially available, namely if process measurements only, or process measurements as well as (perhaps limited) fundamental process knowledge (e.g. in the form of physical laws to which the process is known to obey) shall be used. The appropriate model transfer approach also depends on the source of the available process data, namely if plant A data only, or plant A data as well as plant B data are available. Finally, the process variables that are used to design the monitoring model has to be considered as well. In fact, some variables measured in both plants might be similar in nature (*common variables*), but some other may not. Assuming that the fundamental driving forces of the process do not change between the plants, common variables deserve special attention because they may reflect similar signatures of the process in each plant, and therefore may provide a link between the plants. Therefore, whether common variables only should be used to design the plant B monitoring model, or both common variables as well as other variables are to be used is a matter of decision.

A general framework is therefore proposed to tackle the problem of transferring a process monitoring model between different plants that manufacture the same product. The framework is illustrated in Figure 7.1 and is based on five different scenarios, depending on the combination of the issues mentioned above. For each scenario, a solution strategy based on LV modeling is proposed.

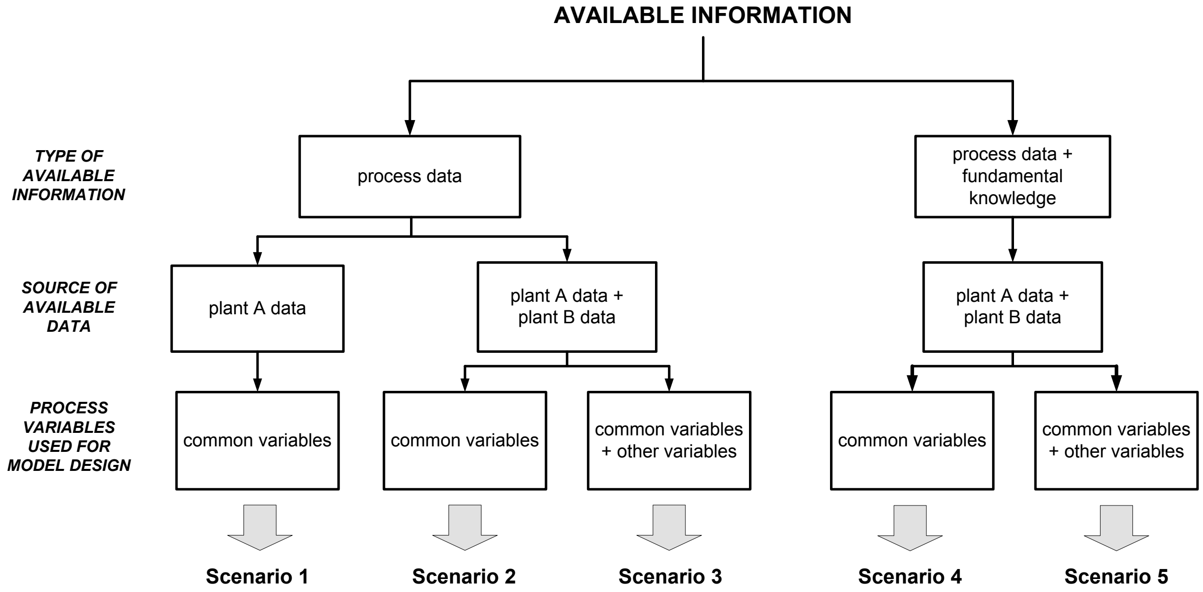

AVAILABLE INFORMATION
TYPE OF AVAILABLE INFORMATION
SOURCE OF AVAILABLE DATA
PROCESS VARIABLES USED FOR MODEL DESIGN
process data
process data + fundamental knowledge
plant A data
plant A data + plant B data
plant A data + plant B data
common variables
common variables
common variables + other variables
common variables
common variables + other variables
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5

Figure 7.1. *Proposed framework for the development of latent variable approaches to the transfer of process monitoring models between different plants.*

In the following, the strategies conceived for each Scenario of the framework are described and applied to a case study concerning the transfer of the monitoring model for an industrial continuous spray-drying process between two plants that differ in the production scale. First, scenarios that exploit only process data for the transfer are presented, and monitoring results are shown (Scenario 1, Scenario 2 and Scenario 3). Secondly, scenarios that combine the use of process data with fundamental engineering knowledge in terms of conservation laws are described and the relevant results presented (Scenario 4 and Scenario 5). Finally, the techniques proposed for Scenario 2 and Scenario 3 of the proposed framework are extended to batch processes by application to a case study dealing with the transfer of the monitoring model for a penicillin batch fermentation process, and preliminary results are discussed.

7.2 Spray-drying process and available data

The continuous process considered in this study is a pharmaceutical spray-drying process (Figure 7.2). Spray-drying is widely used in the pharmaceutical industry, not only for the preparation of solid amorphous dispersions, but also for excipient manufacturing, biotherapeutic particle engineering, drying of crystalline active pharmaceutical ingredients and encapsulation (Dobry *et al.*, 2009). A schematic of the process is shown in Figure 7.2 (García-Muñoz and Settell, 2009).

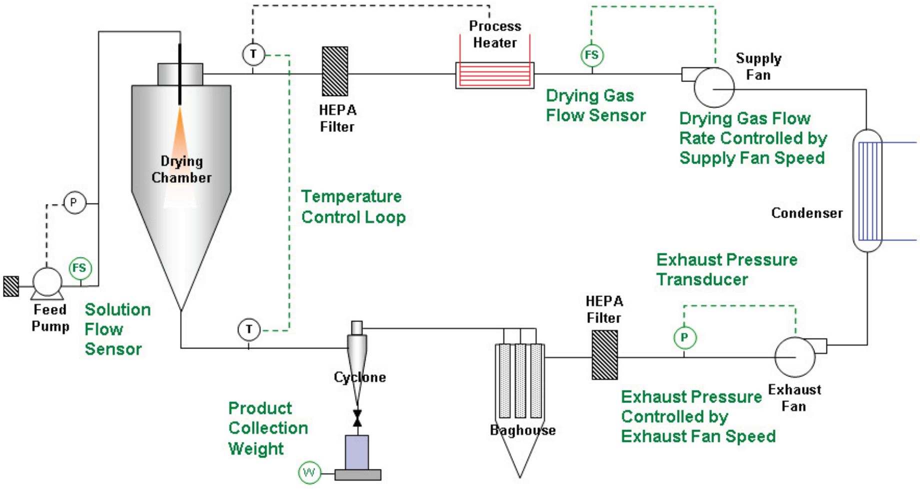


Figure 7.2. *Schematic of the spray-drying plants (adapted from García-Muñoz and Settell, 2009).*

Two industrial plants of different size are considered, namely a pilot-scale unit (plant A) and a production-scale unit (plant B). The plants are designed with similar (although not identical) layouts. The objective is to develop a model to monitor the performance of the production-scale plant using information from the pilot-scale plant, i.e. a way to scale-up the process monitoring model is sought for.

Process data are available on normal operating conditions (NOC) in plant A and plant B, as well as on a real fault occurred in plant B. Data are organized in the following datasets:

- $\mathbf{X}^{\mathrm{A}}$, which includes $I^{\mathrm{A}} = 15031$ NOC samples from plant A, for which $V^{\mathrm{A}} = 16$ process variables were measured.
- $\mathbf{X}^{\mathrm{B}}$, which includes $I^{\mathrm{B}} = 4224$ NOC samples from plant B, for which $V^{\mathrm{B}} = 10$ process variables were measured. The sampling interval in the $\mathbf{X}^{\mathrm{B}}$ dataset is larger than that in the $\mathbf{X}^{\mathrm{A}}$ dataset.
- $\mathbf{X}^{\mathrm{BF}}$, which includes $I^{\mathrm{BF}} = 81$ samples from a real fault occurred in plant B, for which V^{B} process variables were recorded. The fault was due to a little wandering metal piece that clogged one of the swirl nozzle channels of the production-scale plant. Because it is known that the fault onset at sample no. 25, the faulty dataset is split into two phases: phase 1

corresponds to the first 24 normal operating condition samples of the faulty sequence, whereas phase 2 includes samples from no. 25 to no. 81 (actual appearance of the fault).

In Table 7.1 the process variables measured in each plant are listed. Note that, between the plants, the measured process variables differ in several respects, e.g. number, sampling frequency, variability, measurement units, and actual location of the measurement sensor in the plant. However, some variables (indicated in italics in Table 7.1) share the same physical meaning in both plants, and may therefore be thought as *common* between them.

Table 7.1. *Process variables measured in plant A and in plant B. Process variables that are common between the plants are indicated in italics; response variables that are common between the plants (an issue discussed in Section 7.4.2) are marked by* †.

Plant A (pilot-scale unit)			**Plant B (production-scale unit)**		
Var. no.	**Measured variables**		**Var. no.**	**Measured variables**	
1	*Pressure 1* (psig)	†	*1*	*Pressure 1* (barg)	†
2	*Temperature 1* (°C)		*2*	*Temperature 1* (°C*)*	
3	Temperature 2 (°C)		*3*	*Temperature 3* (°C)	†
4	*Temperature 3* (°C)	†	*4*	*Flowrate 1* (kg/h)	
5	*Flowrate 1* (kg/h*)*		*5*	*Flowrate 2* (kg/h)	
6	Temperature 4 (°C)		*6*	*Pressure 2* (mbar)	†
7	*Flowrate 2* (kg/h)		*7*	*Pressure 3* (mbar)	†
8	*Pressure 2* (mm_{H2O})	†	*8*	*Pressure 4* (mbar)	†
9	*Pressure 3* (mm_{H2O})	†	9	Pressure 7 (mbar)	
10	*Pressure 4* (mm_{H2O})	†	*10*	*Pressure 6* (mbar)	†
11	Pressure 5 (mm_{H2O})				
12	*Pressure 6* (mm_{H2O})	†			
13	Temperature 5 (°C)				
14	Speed 1 (%)				
15	Speed 2 (%)				
16	Speed 3 (%)				

Namely, process variables $v'^{\text{A}} = \{1, 2, 4, 5, 7, 8, 9, 10, 12\}$ of plant A have the same physical meaning of process variables $v'^{\text{B}} = \{1, 2, 3, 4, 5, 6, 7, 8, 10\}$ of plant B[**]. Therefore, a set $\mathscr{V}'$ of $V' = 9$ measured process variables are common between the plants. The within-plant correlation and between-plant correlation of common variables can provide very valuable information related to the transfer of a monitoring model from one plant to the other one.

An additional classification of the measured variables will be considered in the case of Scenario 3 (Section 7.4.2), and the selected variables are marked in Table 7.1 by the † symbol. An overview of the variable classification criteria and of the datasets used for each of the model transfer Scenarios of Figure 7.1 is presented in Table 7.2, together with the relevant used notation.

[**] Superscripts ' and " are used in this Chapter to denote two different classifications of the measured process variables. Namely, superscript ' refers to the classification (common variables vs. other variables) used in model transfer Scenarios 1 and 2, whereas superscript " is used to refer to the classification (common response variables vs. other variables) used in Scenario 3.

Table 7.2. *Classification of the measured variables and their notation according to the model transfer scenario. The meaning of symbols k, S and W will be provided in the relevant sections describing each Scenario.*

Scenario	**Variable classification**	**Plant**	**Number of variables**	**Data matrix**	**Dimension of data matrix**
Scenario 1 & Scenario 2	common variables	A	V'	$\mathbf{X}'^{\mathrm{A}}$	$I^A \times V'$
		B	V'	$\mathbf{X}'^{\mathrm{B}}_k$	$(k–1) \times V'$
Scenario 3	other variables	A	V''^{A}	$\mathbf{X}''^{\mathrm{A}}$	$I^A \times V''^{\mathrm{A}}$
		B	V''^{B}	$\mathbf{X}''^{\mathrm{B}}_k$	$(k–1) \times V''^{\mathrm{B}}$
	common response variables	A	V''	$\mathbf{Y}''^{\mathrm{A}}$	$I^A \times V''$
		B	V''	$\mathbf{Y}''^{\mathrm{B}}_k$	$(k–1) \times V''$
Scenario 4	common variables	A	V'	$\mathbf{X}'^{\mathrm{A}}_{\mathrm{SUB}}$	$S \times V'$
		B	V'	$\mathbf{X}'^{\mathrm{B}}_{\mathrm{W}}$	$W \times V'$
Scenario 5	none	A	V^{A}	$\mathbf{X}^{\mathrm{A}}$	$I^A \times V^{\mathrm{A}}$
		B	V^{B}	$\mathbf{X}^{\mathrm{B}}_{\mathrm{W}}$	$W \times V^{\mathrm{B}}$

Furthermore, to assess the effect of uncertainty onto the proposed model transfer methods, one hundred different realizations of the plant B real fault were generated artificially. To this purpose, the original faulty data in $\mathbf{X}^{\mathrm{BF}}$ were first filtered with a median filter (with a window size of 3 samples), and the noise of each of the original variables was characterized by estimating the variance of the difference between the original signal and the filtered one. The one hundred different realizations of the fault were then generated by adding to each of the filtered faulty variables a Gaussian random noise, with the same variance as the one estimated from the real data. Each fault realization includes 81 samples and maintains the same division in phase 1 and phase 2 as the original faulty dataset.

7.3 Transfer based on process data only

As long as the data collected from plant B are not enough to build a monitoring model based entirely on these data, a way to transfer to plant B the plant A dataset is sought for. Note that, when process data are becoming available from the operation of plant B, the plant B monitoring model may or may not be adapted using these incoming data. In both cases, a time will be reached when the model transfer will be stopped and the process will be monitored using plant B data only.

In this section, the first three proposed model transfer scenarios of Figure 7.1 will be presented. Note that, for convenience, only a subset of the plant B dataset $\mathbf{X}^{\mathrm{B}}$ was considered to build the monitoring model, which includes $I^{\mathrm{B}} = 3750$ data.

With reference to the adaptive approaches, we will indicate with *k* the model updating instant, i.e. the time at which the incoming plant B measurements are used to adapt the monitoring model. As will be clarified later, all the proposed methodologies rely on the assumption that the correlation structure between common variables remains essentially the same in both plants.

7.3.1. Data pretreatment

One key issue for all the proposed transfer procedures is data pretreatment. The differences existing in common variables (e.g. in measurement units, measurement sensor location, …) can be compensated for by autoscaling, i.e. by mean-centering the measured process variable measurements and scaling them to unit variance. However, the *plant* difference must be compensated for as well, and for this reason the pretreatment of data related to one plant must be performed *within the same plant*. Therefore, plant A data must be mean-centered and scaled on the mean and standard deviation measured on plant A, whereas plant B data must be mean-centered and scaled on the mean and standard deviation measured on plant B (García-Muñoz, 2004; Chiang and Colegrove, 2007).
With respect to plant A, the mean and standard deviation of any process variable v are known because the plant A dataset does not change over time. As for the plant B data, they can be autoscaled on values of mean and standard deviation that are adaptively updated any time new normal operating condition samples become available from this plant[†]. In this case, the mean-centering and scaling of process variable v are performed at updating instant k based on the mean and standard deviation of the samples available for plant B up to sample $(k-1)$. Throughout this Chapter, reference is always made to measured data that have been autoscaled within the appropriate plant.

7.3.2 Model transfer using common process variables only

Usually, several of the process variables measured in one plant are correlated, the correlation structure being related to the (possibly unknown) fundamental mechanisms driving the process in that plant. For variables that are common between the plants, the correlation structure in one plant is expected to be nearly the same as that in the other plant, because the fundamental mechanisms driving the process do not change across the plants and therefore leave similar signatures on the measured common variables. Following this rationale, two different strategies are proposed to transfer the monitoring model from plant A to plant B based on process data only:

- Scenario 1: this strategy builds the plant B monitoring model using plant A data only, and results in the design of a PCA monitoring model;

[†] Alternatively, autoscaling on fixed expected values of the mean and standard deviation for plant B may prove a viable alternative.

- Scenario 2: this strategy builds the plant B monitoring model using plant A data as well as incoming plant B data, and results in the design of an adaptive PCA monitoring model.

A comprehensive discussion on the application of PCA to process monitoring is provided by MacGregor and Kourti (1995) and Nomikos and MacGregor (1994).

7.3.2.1 Scenario 1: process monitoring in plant B using PCA

The PCA monitoring model is built on matrix $\mathbf{X}'^{\mathrm{A}}$ of the common process variables measured in plant A. Under the assumption that the directions of maximum variability in both plants are due to the same driving forces, projecting the incoming plant B data onto the latent space of the plant A principal components can effectively survey the operation of plant B. Accordingly, faults can be detected by projecting the incoming plant B data onto the PCA model designed on plant A data, and observing how each sample locates in the Hotelling T^2 and squared prediction error (SPE) control charts (Chapter 2, Section 2.1.4).

7.3.2.2 Scenario 2: process monitoring in plant B using adaptive PCA

The PCA monitoring model can be made adaptive (Rännar *et al.*, 1998; Qin, 1998; Li *et al.*, 2000) if it is designed based not only on the available plant A dataset, but also on the incoming plant B samples. Model adaptation is carried out at updating instant *k*. At this instant, the monitoring model is built on a data matrix $\mathbf{X}'_k$ where the common variable data measured in plant B up to sample $(k-1)$ are included:

$$\mathbf{X}'^{\mathrm{B}}_k = \begin{bmatrix} \mathbf{x}'^{\mathrm{B}^{\mathrm{T}}}_1 \\ \mathbf{x}'^{\mathrm{B}^{\mathrm{T}}}_2 \\ \vdots \\ \mathbf{x}'^{\mathrm{B}\;\mathrm{T}}_{k-1} \end{bmatrix} = \begin{bmatrix} x'^{\mathrm{B}}_{1,1} & x'^{\mathrm{B}}_{1,2} & \cdots & x'^{\mathrm{B}}_{1,V'} \\ x'^{\mathrm{B}}_{2,1} & x'^{\mathrm{B}}_{2,2} & \cdots & x'^{\mathrm{B}}_{2,V'} \\ \vdots & \vdots & \ddots & \vdots \\ x'^{\mathrm{B}}_{k-1,1} & x'^{\mathrm{B}}_{k-1,2} & \cdots & x'^{\mathrm{B}}_{k-1,V'} \end{bmatrix} . \tag{7.1}$$

Matrix $\mathbf{X}'^{\mathrm{B}}_k$ is concatenated vertically to the available plant A common variable data ($\mathbf{X}'^{\mathrm{A}}$):

$$\mathbf{X}'_k = \begin{bmatrix} \mathbf{X}'^{\mathrm{A}} \\ \mathbf{X}'^{\mathrm{B}}_k \end{bmatrix} . \tag{7.2}$$

Note that, if some of the incoming plant B samples are found or known to be far from the plant B normal operating conditions (e.g. because of purposely different settings of the plant), they should be removed from $\mathbf{X}'^{\mathrm{B}}_k$.

At updating instant *k*, the algorithm goes through the following steps:

1. design a PCA model on matrix $\mathbf{X}'_k$;

2. build the control charts for both the Hotelling's T^2 statistic and the SPE statistic with the relevant (say, $1-\alpha=95\%$) confidence limits $T^2_{(1-\alpha)\lim}$ and $\text{SPE}_{(1-\alpha)\lim}$ (Chapter 2, Section 2.1.4);
3. autoscale the new incoming sample $\mathbf{x}'^{\text{B}}_{k}$ from plant B on the current values of mean and standard deviation for all the V' common variables of plant B available so far;
4. project the new plant B sample onto the space of principal components determined by the PCA model designed at point 1;
5. calculate T^2_k and SPE_k for the incoming plant B data and compare them to the confidence limits in the relevant control chart. If at time instant *k* the new incoming sample is found to be out of the confidence limits, the model should not be updated.

7.3.3 Model transfer using common variables as well as other variables

Although the measurements of some variables that are similar in nature (common variables) may be available in both plants, in most cases some variables are measured in one plant but not in the other. In these cases, it may be not beneficial to discard *a priori* all the variables that are not measured in both plants, as the information embedded in these variables might be useful for process monitoring purposes, especially if the correlation between common and other variables within a plant is not very strong (for example, because the variables that are not common are representative of additional driving forces acting on the process). For these reasons, Scenario 3 considers *all* the measured variables in each plant to transfer knowledge between them for monitoring purposes.

7.3.3.1 Scenario 3: process monitoring in plant B using adaptive JY-PLS

This approach is based on the use of JY-PLS (Garcia-Muñoz *et al.*, 2005) which is extended here to the transfer of process monitoring models. In this context, JY-PLS models the space of common variables in conjunction with the space of variables that have not been labeled as common and are specific of each single plant (Figure 7.3). This allows analyzing the between-plant correlation structure jointly with the within-plant correlation structure for each plant. Therefore, the use of JY-PLS allows monitoring the process in a space of reduced dimension made of latent variables that take into account the correlation of common variables between the plants as well as the correlation between *all* the variables within each plant. The space of the directions of maximum joint variability between common variables in the different plants (the joint space) is used to monitor the plant B process through control charts, which include information on the relation between common variables and all other variables within each plant. The JY-PLS model and the control charts are adaptively updated each time a new incoming sample from plant B becomes available.

It should be noted that the operation of labeling a set of variables as common between plants has some degree of arbitrariness. In fact, any subset $\mathscr{V}_{\text{sub}} \subset \mathscr{V}$ of variables can be labeled as common, which implies that, in the context of JY-PLS model transfer, defining what "other variables" are in a given plant may be a matter of convenience. As an example, let us consider the process under investigation and the measurements available in plant B (Table 7.1): only variable #9 is not common between the plants, and labeling this single variable as "other variable" in plant B would make matrix $\mathbf{X}_k''^{B}$ in Figure 7.3 a column vector, with limited predictability over the space $\mathscr{V}''$ of common variables. Therefore, to show the potential of model transfer via JY-PLS, in this study some common variables were moved to the space of other variables (in each plant). A simple criterion was used to move the variables: the subspace $\mathscr{V}_{\text{sub}}$ of common variables was defined such that it is made only by those elements (i.e. variables) of $\mathscr{V}''$ that are also controlled by the control system in both plants. These variables, which are referred to as "common response variables" in the following, are indicated by symbol † in Table 7.1. All remaining variables in each plant are referred to as "other variables" for that plant. Therefore, the following variable classification is used in the context of JY-PLS (Figure 7.3 and Table 7.2):

- common response variables ($\mathbf{Y}''^{\text{A}}$ and $\mathbf{Y}_k''^{\text{B}}$): they correspond to variables $v''^{\text{A}} = \{1,4,8,9,10,12\}$ in plant A, and to variables $v''^{\text{B}} = \{1,3,6,7,8,10\}$ in plant B;
- other variables ($\mathbf{X}''^{\text{A}}$ and $\mathbf{X}_k''^{\text{B}}$): they correspond to variables $v^{\text{A}} = \{2,3,5,6,7,11,13,14,15,16\}$ in plant A, and to variables $v^{\text{B}} = \{2,4,5,9\}$ in plant B.

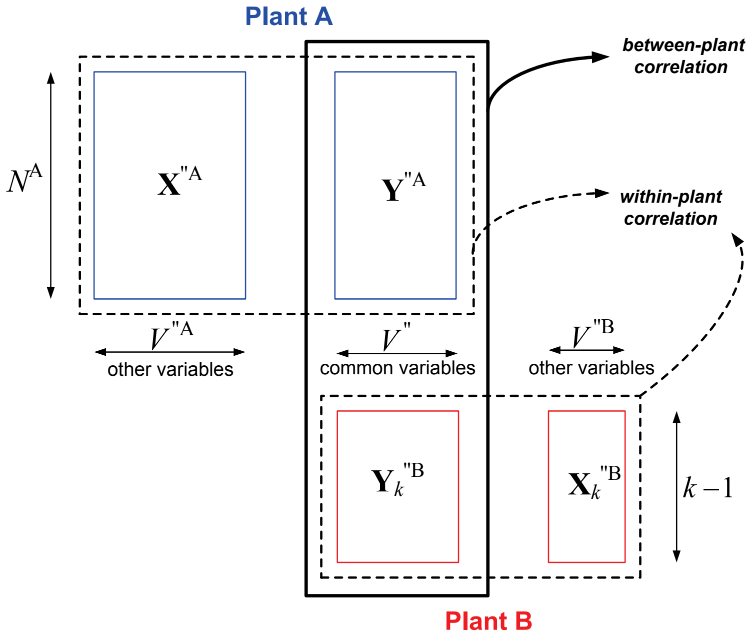


Figure 7.3. *Scenario 3. Schematic of the adaptive JY-PLS model at updating instant k.*

The use of adaptive JY-PLS for model transfer is based on the design of the monitoring model using the entire plant A dataset ($\mathbf{X}''^{\text{A}}$ and $\mathbf{Y}''^{\text{A}}$) as well as the samples incoming from plant B up to instant k–1 ($\mathbf{X}_k''^{\text{B}}$ and $\mathbf{Y}_k''^{\text{B}}$). At each updating instant k the algorithm goes through the following steps:

1.design a JY-PLS model (Chapter 2, Section 2.1.3.2) on all the data from plant A plus data available from plant B up to sample $\left(k-1\right)$.;

2.build the control charts for both the Hotelling's T^2 statistic and the SPE statistic, with (say) 95% confidence limits;

3.autoscale the incoming data from plant B ($\mathbf{y}''^{\mathrm{B}}_k$ as well as $\mathbf{x}''^{\mathrm{B}}_k$) on the current values of mean and standard deviation for all the variables measured in plant B;

4.project the incoming plant B data onto the joint space of the JY-PLS model designed at point 1:

$$\hat{\mathbf{t}}^{\mathrm{B}^{\mathrm{T}}}_k = \mathbf{x}''^{\mathrm{B}^{\mathrm{T}}}_k \mathbf{W}^{*\mathrm{B}}_k \quad ; \tag{7.3}$$

5.in such a way as to obtain the prediction $\hat{\mathbf{y}}''^{\mathrm{B}}_k$ of the common response variables:

$$\hat{\mathbf{y}}''^{\mathrm{B}}_k = \mathbf{Q}_{\mathrm{J},k}\hat{\mathbf{t}}^{\mathrm{B}}_k \quad ; \tag{7.4}$$

6.calculate T^2_k and SPE_k for the new data, and compare them to the confidence limits in the relevant control chart of the joint space. If at time instant *k* the new incoming sample is found to be out of the confidence limits, the model should not be updated.

The control chart design and interrogation procedures for the adaptive JY-PLS model are discussed in Appendix D.

7.3.4 Results and discussion

The proposed strategies for the transfer of the spray-drying monitoring model from plant A (pilot-scale unit) to plant B (production-scale unit) have been tested using the following types of data:

- plant B normal operating condition data;
- faulty data from the plant B real fault;
- faulty data from the 100 artificial realizations of the plant B real fault.

The alarm rate has been used to provide a quantitative evaluation of the fault detection performance of the model. To define the alarm rate, it has been assumed that an alarm is warned ($A_i = 1$) by the monitoring model when ($\Delta - 1$) out of Δ consecutive plant B samples lie outside the 95% confidence limit in either the T^2 or the *SPE* monitoring chart. The alarm rate AR has been defined as the ratio between the total number A_{tot} of alarms warned on *N* samples projected onto the monitoring model and the number *N* of projected samples, i.e. (in percentage terms):

$$\mathrm{AR} = 100 \times \frac{A_{tot}}{N} \quad , \tag{7.5}$$

where:

$$A_{tot} = \sum_{i=1}^{I} A_i \quad . \tag{7.6}$$

The alarm A_i warned at sample n can take two values, i.e.:

$$A_i = \begin{cases} 1, \text{ if } \sum_{i=k-\Delta+1}^{k} \psi_i \geq \Delta - 1 \\ 0, \text{ if } \sum_{i=k-\Delta+1}^{k} \psi_i < \Delta - 1 \end{cases} , \tag{7.7}$$

where $k = \Delta, \Delta+1, \ldots, I$, and:

$$\psi_i = \begin{cases} 1, \text{ if } T_i^2 > T_{95\%\lim}^2 \text{ or } SPE_i > SPE_{95\%\lim} \\ 0, \text{ if } T_i^2 \leq T_{95\%\lim}^2 \text{ and } SPE_i \leq SPE_{95\%\lim} \end{cases} . \tag{7.8}$$

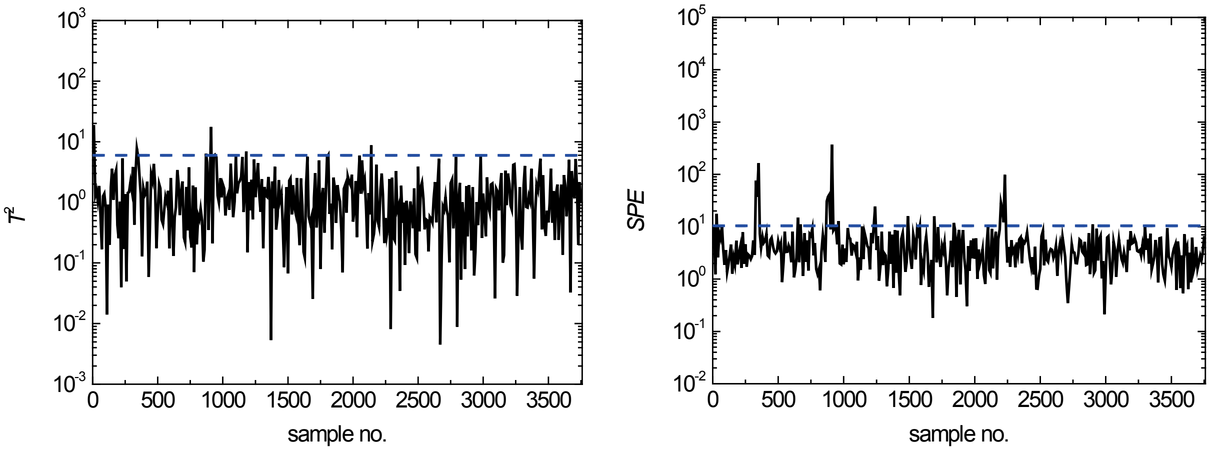
AR is expected to be close to zero for normal operating condition samples, whereas it should be close to 100% for faulty samples.

The monitoring model detection performance has further been assessed by evaluating the time needed to detect a fault. To this purpose, the time to detection index has been used, which is defined as the number of samples after the real fault occurs and until an alarm is generated (García-Muñoz *et al.*, 2004). For quick fault detection, the time to detection should be as short as possible.

The results presented in the following refer to the case of $\Delta = 5$ samples. Results referring to the 100 artificial realizations of the plant B fault have been averaged over all the fault realizations, i.e. a mean AR and a mean time to detection have been evaluated in this case.

7.3.4.1 Results for Scenario 1

A two-principal component monitoring model has been designed using only the entire plant A dataset, and a preliminary study has been carried out to evaluate its performance when the available plant B normal operating condition dataset is presented to (i.e. projected onto) the model. Figure 7.4 presents the resulting T^2 and SPE control charts. Both plots show that the plant B normal operating condition samples are correctly assessed as normal by the monitoring model designed using plant A data only (2.4% of them exceed the T^2 limit and 5.9% exceed the SPE limit, which is not too far to the 5% control limit violations expected for the plant A calibration samples). Following the definition of alarm rate, AR = 0.81% for this plant B dataset.



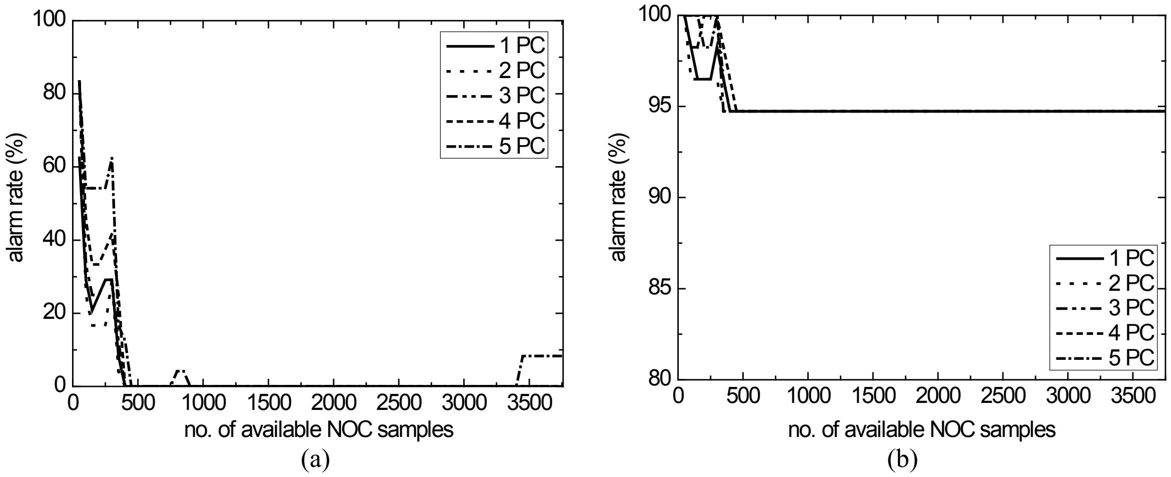
(a) (b)

Figure 7.4. *Scenario 1: monitoring performance on the plant B normal operating condition dataset. (a) Hotelling's T2 chart, (b) SPE chart.*

Next, we study how many NOC samples need to become available from plant B before the monitoring model can perform satisfactorily in the detection of a fault (recall that in Scenario 1 the available plant B normal operating condition samples are not used to update the monitoring model, but they only serve to update the mean and standard deviation of the plant B incoming samples).

Results referring to the projection of the plant B real faulty dataset onto the PCA model are presented in Figure 7.5 in terms of alarm rate for different numbers of PCs retained in the monitoring model. Each point in the *x*-axis indicates the number of plant B normal operating condition samples that the monitoring model "has seen" before the faulty dataset is presented to the monitoring model. Recall that the fault onsets at sample no. 25; therefore, the faulty dataset is split into two phases: phase 1 corresponds to the first 24 normal operating condition samples of the faulty sequence (Figure 7.5a), whereas phase 2 includes samples from no. 25 to no. 81 (actual appearance of the fault; Figure 7.5b).

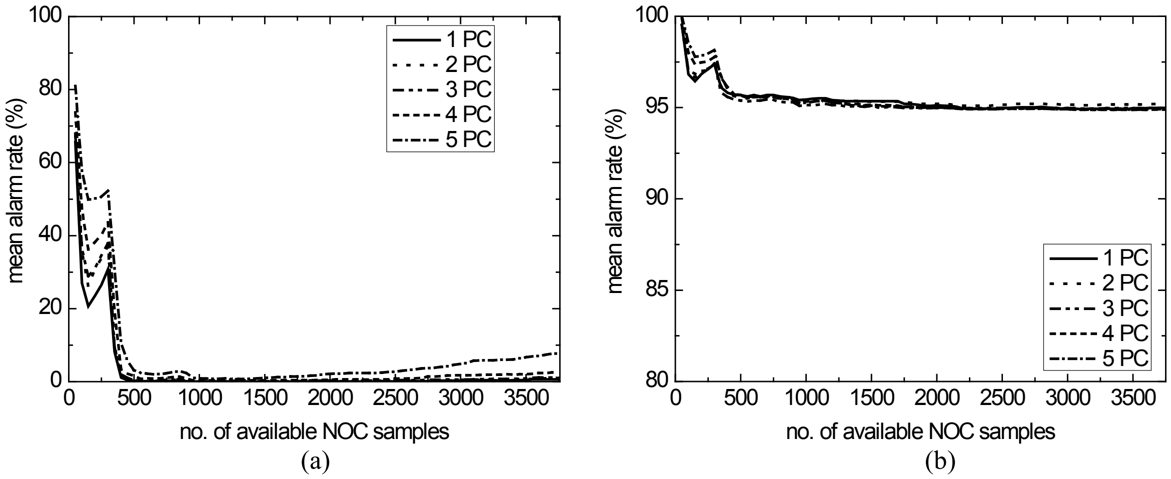
Figure 7.5 indicates that the monitoring performance is very satisfactory. Namely, Figure 7.5a shows that, regardless of the number of retained principal components, a satisfactorily low alarm rate during phase 1 can be obtained if ~300 (or more) normal operating condition samples are available from plant B. On the other hand, during phase 2 (Figure 7.5b) the alarm rate is high even when only few normal operating condition samples are available initially to update the mean and standard deviation of the plant B measurements. The effect of the number of retained PCs seems relatively unimportant in this respect, although using fewer PCs results in a somewhat faster model adaptation.



(a) (b)

Figure 7.5. *Scenario 1: monitoring performance on the plant B real faulty dataset. Effect of the number of available plant B normal operating condition (NOC) samples and of the number of principal components retained in the PCA monitoring model on the alarm rate during (a) phase 1 and (b) phase 2 subsets.*

A similar analysis was carried out also on the 100 artificial realizations of the plant B true fault, and the mean alarm rate plotted in Figure 7.6 confirms the results obtained for the dataset of the real fault.



(a) (b)

Figure 7.6. *Scenario 1: average monitoring performance on the 100 artificial realizations of the plant B faulty dataset. Effect of the number of available plant B normal operating condition (NOC) samples and of the number of principal components retained in the PCA monitoring model on the mean alarm rate during the (a) phase 1 and (b) phase 2 subsets.*

As for the delay in detecting the fault, Figure 7.7 shows that (on average of the whole 100 fault realizations) the fault detection is delayed by at most $\Delta - 1 = 4$ samples, regardless of the number of PCs retained in the model. The apparently small time to detection value obtained when less than (say) ~1000 plant B samples are available is due to the following reason. When only few plant B samples are initially available, the last phase 1 samples are wrongly warned as faulty by the model, and an alarm is generated for each of these samples. Then, when the very first phase 2 samples are projected onto the model, they are detected as faulty and the alarm condition $A_i = 1$ in (7.7) is met even if less than Δ samples have been actually collected in phase 2.

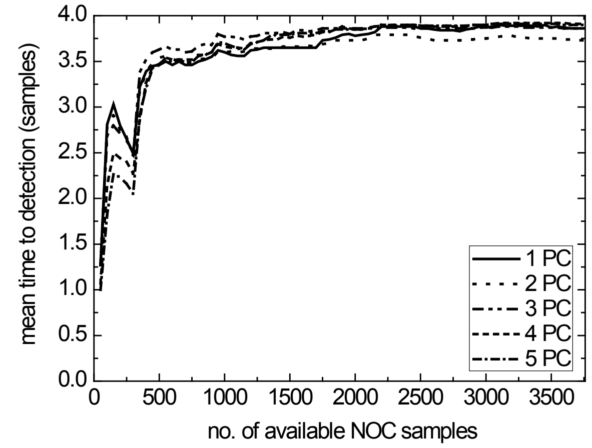


Figure 7.7. *Scenario 1: average monitoring performance on the 100 artificial realizations of the plant B faulty dataset. Effect of the number of available plant B normal operating conditions (NOC) samples and of the number of principal components retained in the PCA monitoring model on the mean time needed to detect the fault.*

7.3.4.2 Results for Scenario 2

An adaptive PCA model has been designed through the procedure described in Section 7.3.2.2 using the entire plant A normal operating condition dataset as well as the incoming data from plant B. The monitoring performance has been evaluated against the plant B real fault dataset as well as its 100 artificial realizations, and similar general results were obtained. Therefore, only results for the case of the artificial realizations are reported. Note that, differently from Scenario 1, in Scenario 2 the available plant B samples are used not only to update the mean and standard deviation of the measured variables, but also to update the monitoring model itself. As in the case of the non-adaptive PCA model, how the alarm rate is affected by the number of plant B NOC samples available initially and by the number of PCs retained in the model was investigated.

Figure 7.8a shows the alarm rate profile during phase 1 of the faulty dataset (normal operating conditions subset). It can be seen that the phase 1 conditions are correctly assessed as normal by the adaptive PCA model, with an acceptably low alarm rate, especially if at least ~300 normal operating condition samples are available initially for model design. Using too many (e.g. 4 or 5) PCs worsens the monitoring results. Figure 8b shows that the faulty data subset (phase 2) is warned as such with a 95% alarm rate, regardless of the number of PCs retained in the model.

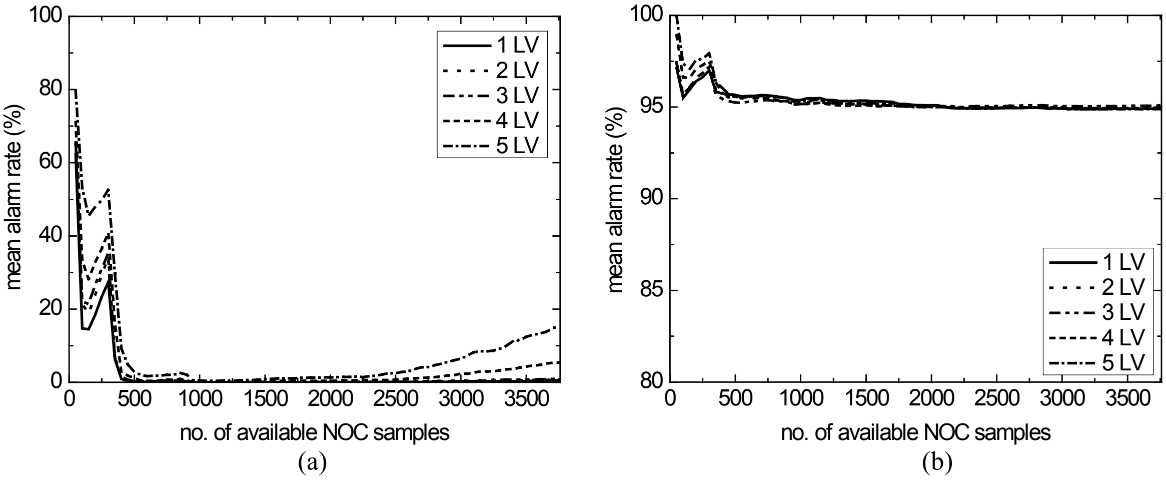


Figure 7.8. *Scenario 2: average monitoring performance on the 100 artificial realizations of the plant B faulty dataset. Effect of the number of available plant B normal operating condition (NOC) samples and of the number of retained principal components on the mean alarm rate during the (a) phase 1 and (b) phase 2 subsets.*

Figure 7.9 indicates that, as in the case of Scenario 1, if enough NOC samples are available from plant B, 4 faulty samples are needed before the fault can be actually warned by the monitoring model.

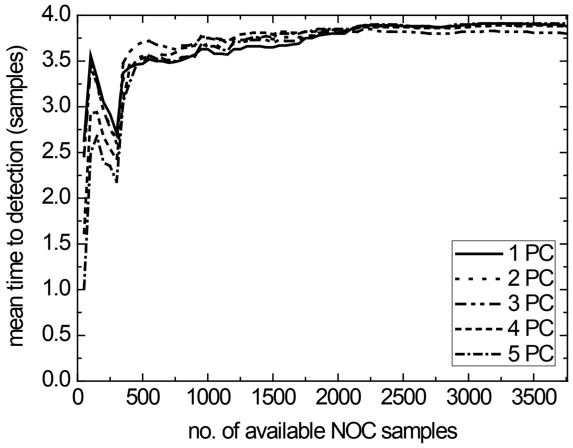


Figure 7.9. *Scenario 2: average monitoring performance on the 100 artificial realizations of the plant B faulty dataset. Effect of the number of available plant B normal operating condition (NOC) samples and of the number of principal components retained in the PCA monitoring model on the mean time needed to detect the fault.*

On the whole, the monitoring performances of the PCA model and of the adaptive PCA model appear very similar.

7.3.4.3 Results for Scenario 3

An adaptive JY-PLS model has been designed through the procedure described in section 7.3.3.1 using the entire plant A normal operating condition dataset as well as the data incoming from plant B. Only those common variables that are controlled by the control system were included in the joint-Y space; this makes the column dimension of the $\mathbf{X}''^{\mathrm{B}}$ matrix be 4. First, results are presented for the monitoring of the artificial realizations of the plant B faulty dataset, then the performance of the monitoring system for the plant B real faulty dataset is considered.

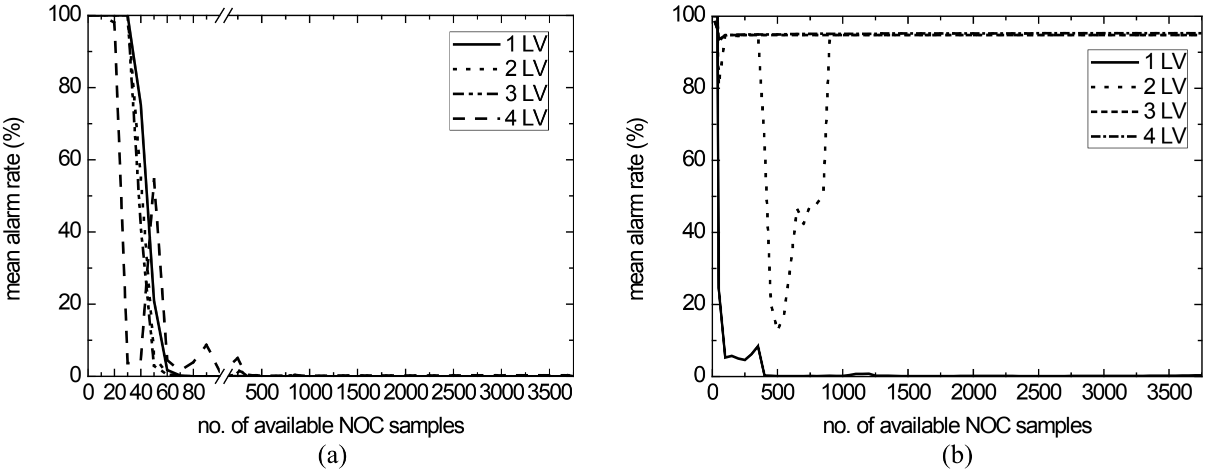


Figure 7.10. *Scenario 3: average monitoring performance on the 100 artificial realizations of the plant B faulty dataset. Effect of the number of available plant B normal operating condition (NOC) samples and of the number of retained latent variables on the mean alarm rate during the (a) phase 1 and (b) phase 2 subsets.*

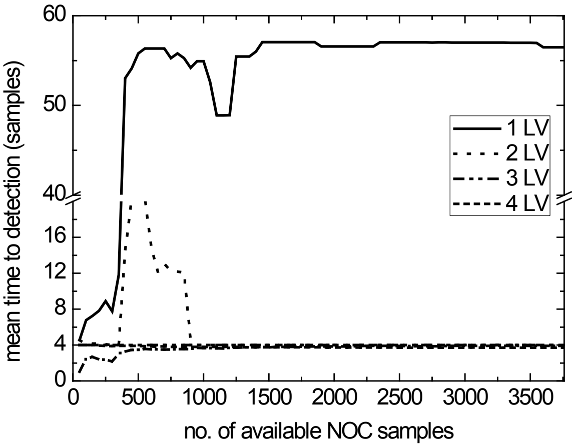


Figure 7.11. *Scenario 3: average monitoring performance on the 100 artificial realizations of the plant B faulty dataset. Effect of the number of available plant B normal operating condition (NOC) samples and of the number of latent variables retained in the JY-PLS monitoring model on the mean time needed to detect the fault.*

Figure 7.10a reports the phase 1 mean AR results for the 100 artificial realizations of the faulty dataset from plant B. The mean AR is very low for any number of latent variables (LVs) retained in the model. On the other hand, Figure 7.10b shows that, in order to achieve an acceptably high mean AR during phase 2, at least three LVs must be retained in the model, regardless of the amount of plant B NOC data initially available to design the model itself. Remarkably, Figure 7.10 also shows that the adaptive JY-PLS model (which makes use of information retrieved from common variables as well as from variables that are not common between the plants) can adapt very quickly to the plant B conditions, especially if at least three LVs are used. It should be reminded that the spaces modeled by the adaptive JY-PLS model and by the PCA models are fundamentally different, and therefore a direct comparison between the optimal number of LVs for the different scenarios is improper. The indication of a minimum optimal number of LVs is confirmed by Figure 7.11, which suggests that retaining 3 or 4 LVs makes the time to fault detection reasonably short.

To further illustrate how the monitoring performance changes as the number K of plant B NOC samples initially available increases, the plant B real faulty sequence was projected onto three different models built on 3 LVs, which differ by the value of K (namely, $K = 20, 40, 60$ samples), and the monitoring performance in each case was analyzed through the relevant control charts (Figure 7.12).

When the model is designed based on a limited initial number of plant B NOC samples (K = 20 samples) and the faulty sequence is projected onto the model (Figure 7.12a-b), the monitoring performance is unsatisfactory because the alarm rate is 100% in phase 1 (see also the averaged values of Figure 7.10a), i.e. alarms are generated even before the fault actually onsets. This is because the variability captured by a model designed on only 20 plant B NOC samples does not fully represent the variability of the incoming plant B samples, even if these samples are used to update the model.

An improvement of the monitoring performance is obtained when K = 40 plant B NOC samples are available initially to design the model (Figure 7.12c-d): the alarm rate in phase 1 decreases to the (still unsatisfactorily high) value of 40% (see also the averaged values of Figure 7.10a), whereas it reaches an appropriate value (94.7%) in phase 2 (see also Figure 7.10b). Note in Figure 7.12c-d that the monitoring performance during phase 1 starts improving after sample #10 has become available from plant B. This highlights the beneficial effect of model adaptation using the incoming plant B NOC samples.

Finally, when more plant B NOC data are available initially to design the model (K = 60 samples, Figures 7.12e-f), the monitoring performance becomes fully satisfactory, with an alarm rate of zero during phase 1, and of 94.7% during phase 2 (also see the averaged values of Figure 7.10). Therefore, it can be concluded that in Scenario 3 satisfactory monitoring performances can be obtained with a much smaller dimension of the plant B NOC database than in the case of Scenarios 1 and 2.

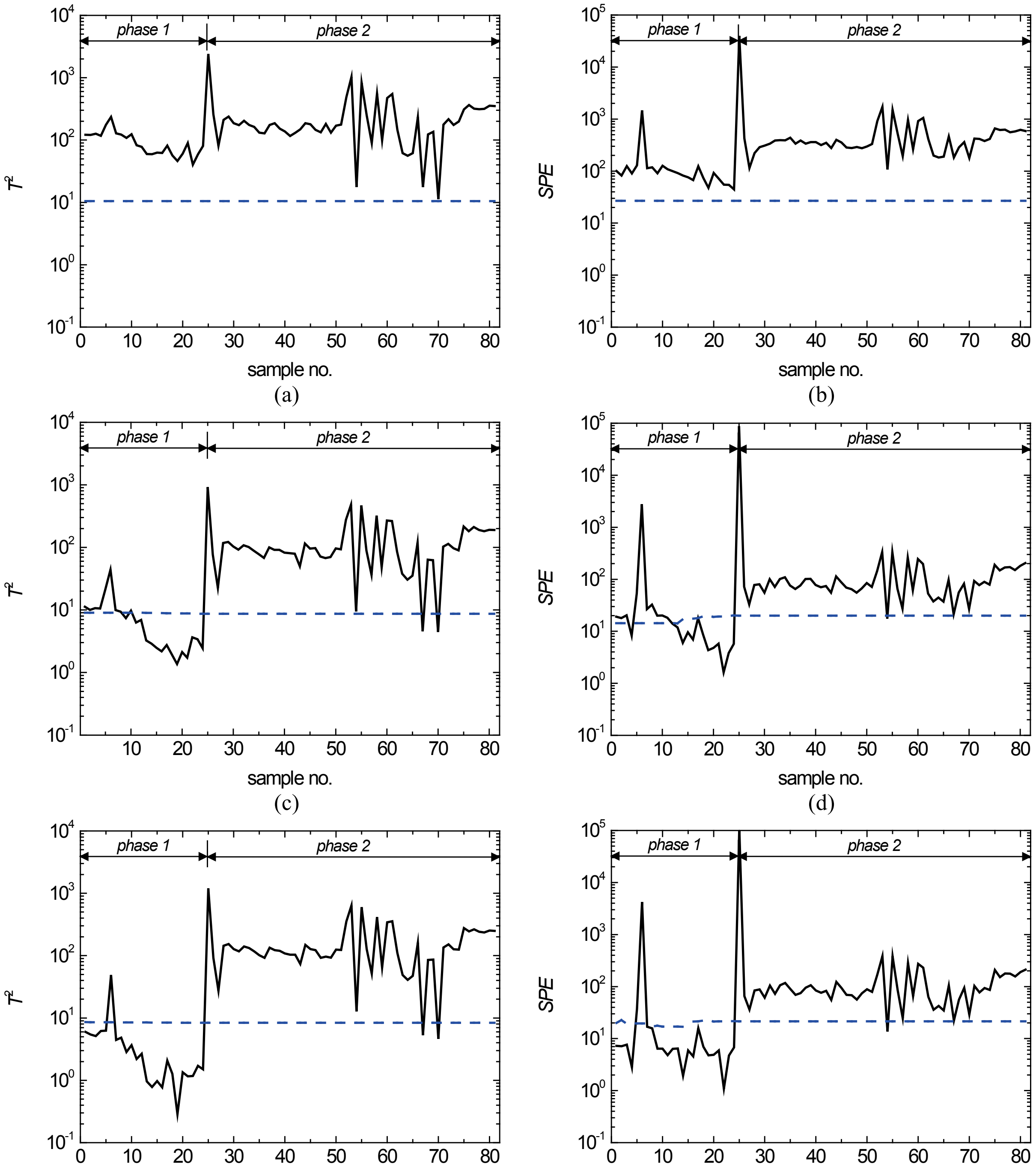


Figure 7.12. *Scenario 3: monitoring performance on the plant B real faulty dataset. Hotelling T^2 and squared prediction error charts for different numbers K of the plant B normal operating condition (NOC) samples available initially: (a) and (b) K = 20 samples, (c) and (d) K = 40 samples, (e) and (f) K = 60 samples. The fault onsets at sample no.25.*

7.4 Transfer based on process data and fundamental knowledge

In general, in two plants dedicated to the same manufacturing process, the fundamental laws describing the physics of the system are expected to be the same, because the underlying physical phenomena driving the process are the same.

As described in Chapter 6, in technology transfer activities between plants (e.g., plant scale-up), it is customary to try identifying combinations of physical variables and/or physical properties (e.g. dimensionless numbers) whose values be as independent as possible from the plant, but dependent on the relevance of the physical phenomena driving the process (Zlokarnik, 2006). The values taken by these plant-independent variables can identify the regime (e.g. heat exchange, fluid flow) at which the plants are operated. Under NOC, two similar plants manufacturing the same product through the same process are expected to be driven by the same driving forces, hence to be characterized by similar values of the relevant plant-independent variables, or similar correlation between their values, or between their values and the values of the other process variables. This feature can be exploited to relate data coming from plant B to the data available from plant A in order to design a monitoring system for plant B.

For the process under investigation, one plant-independent variable can be obtained from a macroscopic steady-state energy balance around the spray-drying chamber (Figure 7.2). This variable summarizes the available knowledge about the manufacturing process and will be used to assist the model transfer exercise.

Following Dobry *et al.* (2009), the energy ΔE^{vap} required to vaporize the solvent in the drying chamber can be calculated as:

$$\Delta E^{\mathrm{vap}} = \dot{M}_{\mathrm{soln}} \cdot (1 - x_{\mathrm{solids}}) \cdot \Delta H^{\mathrm{vap}} \quad , \tag{7.9}$$

where $\dot{M}_{\mathrm{soln}}$ is the solution flowrate entering the system, x_{solids} is the mass fraction of solids in the solution, and ΔH^{vap} is the heat of vaporization. The energy ΔE^{gas} that is lost by the drying gas entering the drying chamber is calculated as:

$$\Delta E^{\mathrm{gas}} = \dot{M}_{\mathrm{gas}} \cdot c_p \cdot \left(T^{\mathrm{IN}} - T^{\mathrm{OUT}}\right) \quad , \tag{7.10}$$

where $\dot{M}_{\mathrm{gas}}$ is the gas flowrate entering the drying chamber, c_p is the gas heat capacity, and T^{IN} and T^{OUT} are the inlet and outlet temperatures of the gas. From an energy balance around the drying chamber, it follows that $\Delta E^{\mathrm{vap}} = \Delta E^{\mathrm{gas}}$. Therefore:

$$\dot{M}_{\mathrm{soln}} \cdot (1 - x_{\mathrm{solids}}) \cdot \Delta H^{\mathrm{vap}} = \dot{M}_{\mathrm{gas}} \cdot c_p \cdot \left(T^{\mathrm{IN}} - T^{\mathrm{OUT}}\right) \quad , \tag{7.11}$$

which, after algebraic manipulation, gives:

$$\frac{\dot{M}_{\mathrm{gas}}}{\dot{M}_{\mathrm{soln}}}\cdot\left(T^{\mathrm{IN}}-T^{\mathrm{OUT}}\right)=\frac{\Delta H^{\mathrm{vap}}}{c_p}\cdot\left(1-x_{\mathrm{solids}}\right) \quad . \tag{7.12}$$

The right-hand side term in Eq.(7.12) contains only physical properties (ΔH^{vap} and c_P) and a variable (x_{solids}) related to the inlet solution, and it is reasonable to assume that the values of these quantities are similar in both plants. Therefore, under normal steady-state conditions, the term in the left-hand side of Eq.(7.12) is expected to take similar values in both plants. This term represents the difference between the inlet and outlet gas temperatures, weighted according to the ratio between the gas and the solution flowrates. We will refer to it as a *weighted temperature difference* (*wtd*). This plant-independent variable can be calculated from the available process measurements in both plants (Table 7.1), and can be used to match similar states reached in both plants. Indeed, *wtd* identifies the thermodynamic design space of the process (Dobry *et al.*, 2009; ICH, 2009): as long as the process moves inside the design space, it can go through different thermodynamic states, characterized by different values of *wtd*, without affecting the operation and hence the quality of the manufactured product. Therefore, assuming that for the NOC of plant A the operation is inside the design space for *wtd*, if the values of *wtd* for the data in plant B match the values of *wtd* of the plant A data, the operation in plant B can be considered acceptable from a thermodynamic point of view. In Figure 7.13, the values of *wtd* calculated for all the samples of $\mathbf{X}^{\mathrm{A}}$ and $\mathbf{X}^{\mathrm{B}}$ are shown. It can be clearly seen that the operation in plant B is represented well by the operation in plant A, according to *wtd*.

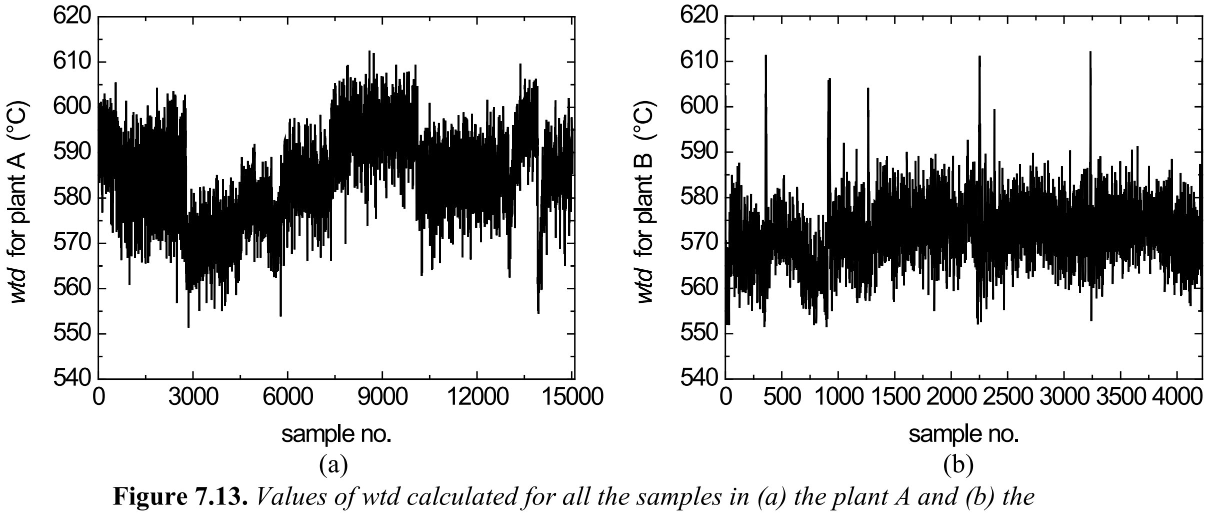


Figure 7.13. *Values of wtd calculated for all the samples in (a) the plant A and (b) the plant B datasets.*

It should be remarked that the identification of an effective plant-independent design space to support the transfer of knowledge from one plant to another one depends not only on the fundamental equations describing the process, but also on the measurement systems available in both plants. For example, *wtd* can be an effective plant-independent variable only if the temperatures and flowrates upon which *wtd* is calculated can be measured in *both* plants, i.e. if they are common variables. Note that in general more than one plant independent variable may be identified, each one describing a different physical phenomenon driving the process.

Next, two different scenarios are proposed to transfer a process monitoring model from plant A to plant B. Both of them use *wtd* as a way to transfer fundamental process knowledge between the plants, but they differ for the way they use the available plant measurements:

- Scenario 4: the plant B monitoring model is a PCA model that uses only measured variables that are common between the plants;
- Scenario 5: the plant B monitoring model is a JY-PLS model using common variables as well as other measured variables.

Differently from the previous scenarios, to evaluate the performance of the proposed model transfer approaches when only a limited number of plant B samples is available, only the first W samples of the whole plant B dataset $\mathbf{X}^{\mathrm{B}}$ are assumed to be available initially (e.g. because the plant was just started up). These samples are organized in matrix $\mathbf{X}_{W}^{\mathrm{B}}$ $[W \times V^{\mathrm{B}}]$.

7.4.1 Scenario 4: model transfer using common process variables only

As disclosed in Section 7.3.2, if the fundamental mechanisms driving the process are assumed to be the same in the two plants, the correlation structure between common variables in one plant is expected to be similar to that in the other plant. Following this rationale, a strategy is proposed to transfer the monitoring system from plant A to plant B based only on the common variables measured in both plants and exploiting the available fundamental knowledge.

The plant B monitoring model is built using PCA on the matrix $\mathbf{X}'^{\mathrm{AB}}$ generated by concatenating the available common variables data from plant B with the data from plant A that (using the available process knowledge) are found to be most similar to plant B data. The similarity between the samples from the two plants is determined by the plant-independent variables (*wtd*, in this case study); details on the similarity concept will be given in the next subsection. The rationale is illustrated in Figure 7.14: *wtd* is calculated for each sample in $\mathbf{X}^{\mathrm{A}}$ and for the W available samples collected in plant B (matrix $\mathbf{X}_{W}^{\mathrm{B}}$). For each value of *wtd* calculated from samples of plant B (collected in column vector $\mathbf{wtd}_{W}^{\mathrm{B}}$), the samples of $\mathbf{X}^{\mathrm{A}}$ having the values of *wtd* (collected in vector $\mathbf{wtd}^{\mathrm{A}}$) most similar to those of plant B are selected to form the matrix $\mathbf{X}'^{\mathrm{A}}_{\mathrm{SUB}}$ $[S \times V']$, in which only common variables are considered. By concatenating $\mathbf{X}'^{\mathrm{A}}_{\mathrm{SUB}}$ with $\mathbf{X}'^{\mathrm{B}}_{W}$ $[W \times V']$ (generated from $\mathbf{X}_{W}^{\mathrm{B}}$ considering only the common variables), matrix $\mathbf{X}'^{\mathrm{AB}}$ is generated:

$$\mathbf{X}'^{\mathrm{AB}} = \begin{bmatrix} \mathbf{X}'^{\mathrm{A}}_{\mathrm{SUB}} \\ \mathbf{X}'^{\mathrm{B}}_{W} \end{bmatrix} \quad . \tag{7.13}$$

The same considerations on data pretreatment reported in Section 7.3.1 are valid in this case. PCA is then applied to $\mathbf{X}'^{\mathrm{AB}}$ in order to build the monitoring model for plant B. The number of PCs used to build the model has been selected automatically, and corresponds to the number of eigenvalues of the matrix $\mathbf{X}'^{\mathrm{AB}^{\mathrm{T}}}\mathbf{X}'^{\mathrm{AB}}$ that are found to be greater than 1 (Chapter 2, Section 2.1.1.2).

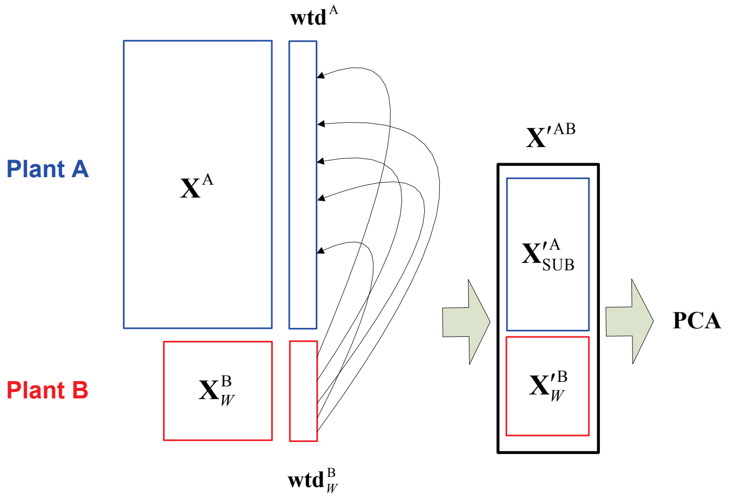

$\mathbf{wtd}^{\mathrm{A}}$
Plant A
$\mathbf{X}^{\mathrm{A}}$
Plant B
$\mathbf{X}_W^{\mathrm{B}}$
$\mathbf{wtd}_W^{\mathrm{B}}$
$\mathbf{X}'^{\mathrm{AB}}$
$\mathbf{X}'^{\mathrm{A}}_{\mathrm{SUB}}$
$\mathbf{X}'^{\mathrm{B}}_W$
PCA

Figure 7.14. *Scenario 4. Schematic of the transfer strategy using common process variables only.*

7.4.1.1 Assessing the similarity between plants

For the construction of the $\mathbf{X}'^{\mathrm{AB}}$ matrix, for each of the W samples available from plant B the most similar samples in plant A are selected. The similarity between samples in the two plants is assessed by comparing the *wtd* values calculated for each of them. For each element wtd_w^{B} of $\mathbf{wtd}_W^{\mathrm{B}}$, the absolute distance between wtd_w^{B} and the i-th element of $\mathbf{wtd}^{\mathrm{A}}$ is calculated as:

$$d_{i,w} = \left| wtd_i^{\mathrm{A}} - wtd_w^{\mathrm{B}} \right| \qquad i = 1,2,\ldots,I^{\mathrm{A}};\quad w = 1,2,\ldots,W \quad . \tag{7.14}$$

The element wtd_i^{A} of $\mathbf{wtd}^{\mathrm{A}}$, for which the smallest value of $d_{i,w}$ is calculated, is selected as the most similar to wtd_w^{B}. Repeating the procedure for all the W elements in $\mathbf{wtd}_W^{\mathrm{B}}$, a vector $\mathbf{wtd}_{\mathrm{MIN}}^{\mathrm{A}}$, formed by the W *wtd* values of the plant A samples considered most similar to those of plant B, is eventually obtained. In $\mathbf{wtd}_{\mathrm{MIN}}^{\mathrm{A}}$, the maximum and minimum *wtd* values are identified, which define the range for the selection of plant A samples: all samples of plant A having *wtd* values falling inside the identified range are selected as the most similar to those in $\mathbf{X}_W^{\mathrm{B}}$, and are used to form the matrix $\mathbf{X}'^{\mathrm{A}}_{\mathrm{SUB}}$ (Figure 7.14).

Note that this method to assess similarity between the plant samples is based exclusively on the plant-independent variable *wtd*. In general, the number and type of plant-independent variables depend on the specific case study. If more than one plant-independent variable is available, the distance calculated using Eq.(7.14) might not be a robust measure of similarity, due to possible correlation between the variables considered. In these cases, it may be more effective to consider correlation-based spectral clustering (Fujiwara *et al.*, 2010 and 2011) or nearest-neighborhood methods (Facco *et al.*, 2010) in the latent space of the plant independent variables (obtained for example from a PCA) rather than in the real variable space.

7.4.2 Scenario 5: model transfer using common process variables as well as other variables

The motivations for considering both common as well as other variables to guide the monitoring model transfer have already been highlighted at the beginning of Section 7.3.3. In Scenario 5 *all* the measured variables in each plant and the available plant-independent variables are considered to transfer knowledge between the plants for monitoring purposes.

In each plant, the plant-independent variables are expected to be highly correlated with the process variables measured in the corresponding samples. Moreover, if the driving forces characterizing the plant operation are similar, it is expected that the plant-independent variables calculated from the plant A and plant B samples share the same correlation structure. This means that plant A and plant B datasets can be studied through the common space generated by the plant-independent variables, following a JY-PLS modeling approach (García-Muñoz *et al.*, 2005).

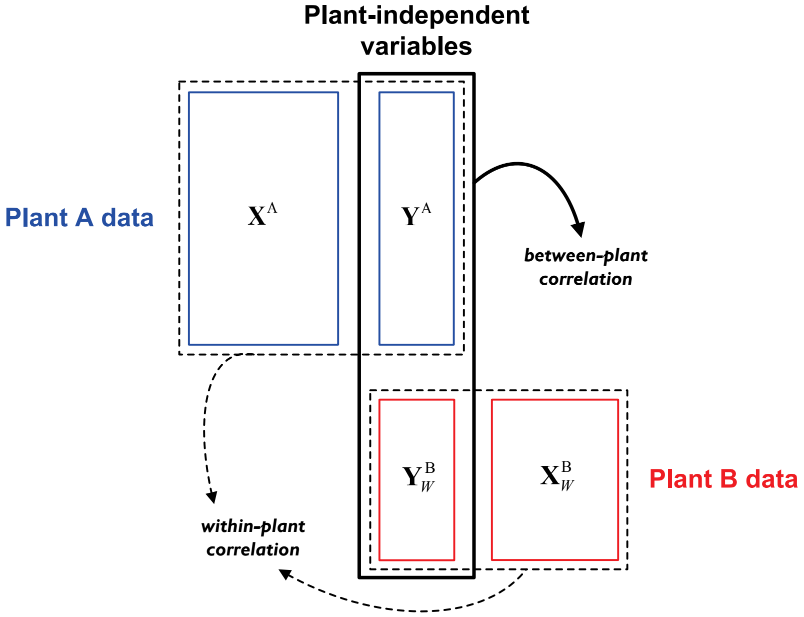


Figure 7.15. *Scenario 5. Schematic of the JY-PLS procedure for the model transfer strategy using common as well as other variables measured in the two plants, together with plant-independent variables.*

Figure 7.15 shows the way in which JY-PLS is used in Scenario 5: the plant-independent variables are calculated for each sample in $\mathbf{X}^A$ and stored in the matrix $\mathbf{Y}^A$, and the same operation is repeated for the W samples available from plant B ($\mathbf{X}_W^{\mathrm{B}}$), generating the matrix $\mathbf{Y}_W^{\mathrm{B}}$ of the plant-independent samples for plant B. In this way, JY-PLS models the within-plant information embedded in each plant dataset jointly with the between-plant information provided by the plant-independent variables.

The parameters of the JY-PLS monitoring model can be found according to what described in Chapter 2 (Section 2.1.3.2). In this case study, the number of LVs used to build the JY-PLS model is selected automatically based on the eigenvalue-greater-than-one rule applied to the correlation matrix of $\mathbf{X}_W^{\mathrm{B}}$ (Chapter 2, Section 2.1.1.2).

Note that to apply the JY-PLS approach it is essential that data in $\mathbf{Y}^{\mathrm{A}}$ and in $\mathbf{Y}_W^{\mathrm{B}}$ share the same correlation structure. This can be assessed for example by building a PCA model on $\mathbf{Y}_W^{\mathrm{B}}$ and verifying that the SPE values for the $\mathbf{Y}^{\mathrm{A}}$ data projected onto the $\mathbf{Y}_W^{\mathrm{B}}$ model are within acceptable limits (García-Muñoz *et al.*, 2005). Nearest-neighborhood methods in the latent space of the joint matrix $\left[\mathbf{Y}^{\mathrm{A}^{\mathrm{T}}}\ \mathbf{Y}_W^{\mathrm{B}^{\mathrm{T}}}\right]^{\mathrm{T}}$ could then be used to select samples in $\mathbf{Y}^{\mathrm{A}}$ most similar to those in $\mathbf{Y}_W^{\mathrm{B}}$, as was described in the case of the transfer strategy considering common variables only (Section 7.4.1). This would generate two datasets ($\mathbf{X}_{\mathrm{SUB}}^{\mathrm{A}}$ and $\mathbf{Y}_{\mathrm{SUB}}^{\mathrm{A}}$) of selected samples, which allow to build local JY-PLS models to improve the performances of the transfer model.

In this case study, $\mathbf{Y}^{\mathrm{A}}$ and $\mathbf{Y}_W^{\mathrm{B}}$ (Figure 7.15) are univariate, because there is only one plant-independent variable (*wtd*). As a consequence, the JY-PLS model may be more affected by the within-plant correlation. For this reason, given that the appropriate data preprocessing is applied (Chapter 2, Section 2.1.3.2), in this case study there is no need to select the plant A samples which are most similar to those available from plant B, but the whole $\mathbf{X}^{\mathrm{A}}$ dataset is used and the model transfer strategy is applied directly as represented in Figure 7.15.

7.4.3 Online monitoring and model adaptation

As long as samples are incoming from plant B, the model transfer procedure needs to be made adaptive for online use (Rännar *et al.*, 1998; Qin, 1998; Li *et al.*, 2000). For Scenario 4 and Scenario 5, a different adaptation strategy has been implemented compared to the Scenario 2 and Scenario 3. At the generic sampling instant j, there are two different reasons why the model may need to be adapted: 1) the j-th sample collected from plant B ($\mathbf{x}_j^{\mathrm{B}}$) is assessed as normal, and therefore a larger plant B database can be used to build the monitoring model (as for Scenario 2 and 3); 2) $\mathbf{x}_j^{\mathrm{B}}$ is assessed as belonging to a set of new plant operating conditions reached recently (e.g. due to fouling, catalyst deactivation), to which the model is required to adapt by changing the plant B database upon which the model is built; this new model will be called a "local" model. If none of these two conditions occur, the monitoring model will not be adapted.

Following this rationale we assume that, at instant j, a monitoring model for plant B is available, which has been built according to the modeling strategies described earlier

(Scenario 1 or Scenario 2) using the plant B NOC samples available until updating instant $(k-1)$ and the relevant plant A samples. Model adaptation is carried out at instant j if any of the following two conditions is met:

- $\mathbf{x}_j^{\mathrm{B}}$ is assessed as normal by the $(k-1)$-th monitoring model;
- $\mathbf{x}_j^{\mathrm{B}}$ is not assessed as normal by the $(k-1)$-th monitoring model, but it is representative of the new NOC defined by a window of the last W plant B samples together with the plant A samples selected on the basis of this window.

Accordingly, two concurrent conditions must be met to warn an alarm for $\mathbf{x}_j^{\mathrm{B}}$:

- the last Φ consecutive samples collected from plant B are found as outliers in the T^2 monitoring chart or in the SPE monitoring chart for the $(k-1)$-th monitoring model;
- the last Φ consecutive samples collected from plant B are found as outliers in the T^2 or in the SPE monitoring charts for the local monitoring model built using a window including only the last W plant B samples plus the plant A samples that can be selected from this window.

A detailed description of the model adaptation mechanism is provided in Appendix D. Note that for both Scenario 4 and Scenario 5, the monitoring of plant B is carried out by calculating the monitored statistics (T^2 and SPE) and their respective confidence limits on the basis of plant B data only. In particular, in Scenario 5 (monitoring through JY-PLS) SPE is calculated only on matrix $\mathbf{X}_W^{\mathrm{B}}$. It would be possible to monitor plant B considering also the SPE on $\mathbf{Y}_W^{\mathrm{B}}$ (similarly to what has been done in Scenario 3). However, it was found that for the case study under investigation this option did not provide any improvement to the monitoring performance.

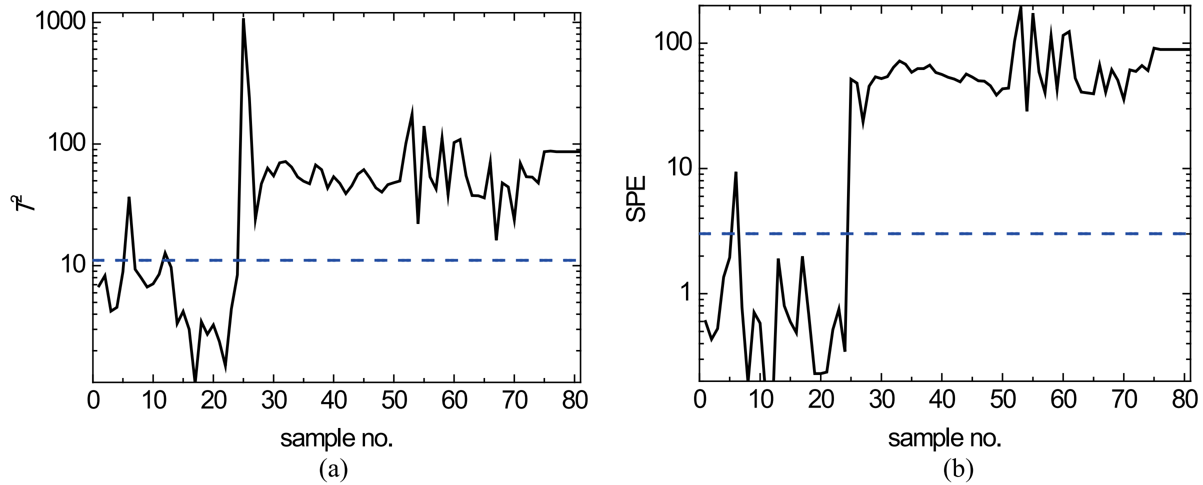
The adaptation procedure requires setting the values of two tuning parameters, namely Φ (i.e. the number of consecutive plant B samples that need to be detected out of the limits to warn a fault) and the window width W (i.e. the number of plant B samples considered for the design of the local model). Note that the local model strategy can be regarded as a form of just-in-time modeling (Cheng and Chiu, 2005; Fujiwara *et al.*, 2010), in which local monitoring models are built around a query point with the most appropriate samples selected according to the plant-independent variables.

7.4.4 Results and discussion

In this section, results on the monitoring performance of the strategies proposed for Scenario 4 and Scenario 5 are reported using both the original plant B faulty dataset and its 100 artificial realizations. For each strategy, the effect of the local model window width W on the monitoring performance is studied. In all presented results, it is assumed that three consecutive samples need to be detected out of the confidence limits of the T^2 or SPE statistics to warn a fault, i.e. $\Phi = 3$ is always used.

The monitoring performance is evaluated in terms of fault detection probability, mean time to detection, and amount of type I and type II errors (Montgomery, 2005b). The fault detection probability is calculated as the percentage of fault realizations in which the fault is detected with respect to the total number of realizations considered (one hundred). The percentages of type I and type II errors averaged over all fault realizations are reported separately for the T^2 and the SPE statistics.

Preliminarily, a limiting condition was studied where the plant B dataset is assumed to be rich enough to allow building a monitoring model using plant B data only. A PCA monitoring model was therefore built using the entire $\mathbf{X}^{\mathrm{B}}$ dataset. Figure 7.16 shows the monitoring charts for the T^2 and SPE statistics when the original faulty dataset is projected onto the model (5 PCs were used). As can be seen, the fault is highlighted very clearly and promptly in both the SPE and the T^2 chart (recall that the fault onsets at sample no. 25).



(a) (b)

Figure 7.16. *Reference monitoring results: (a) T^2 and (b) SPE control charts for the actual faulty dataset projected onto a PCA model built using the entire plant B dataset* $\mathbf{X}^{\mathrm{B}}$ *and no plant A data. The dashed lines correspond to the 95% confidence limits.*

Table 7.3 shows the monitoring performances of the considered model in terms of percentages of type I and type II errors. Recall that type I errors are calculated on the plant B NOC samples available before the onset of the fault (phase 1), whereas type II errors are calculated exclusively on the 57 faulty samples (phase 2).

Table 7.3. *Reference monitoring results: amount of type I and type II errors for the actual faulty dataset projected onto a PCA model built using only the entire plant B dataset* $\mathbf{X}^{\mathrm{B}}$.

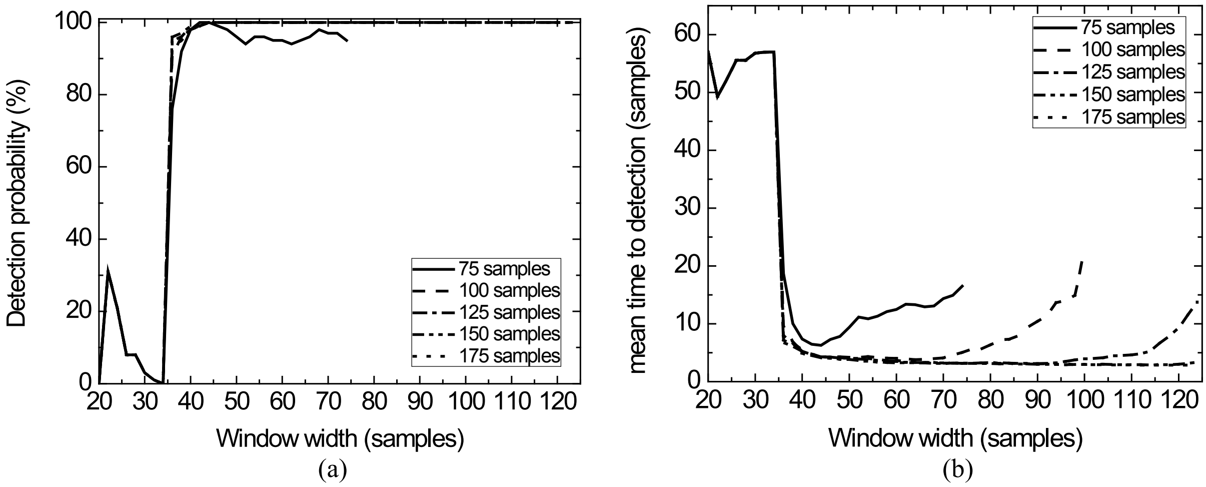
	Type I errors (%)	Type II errors (%)
T^2	8.3	0
SPE	4.2	0

The results shown in Figure 7.16 and Table 7.3 represent a benchmark for the evaluation of the model transfer strategies proposed for Scenario 4 and Scenario 5, as they refer to the best monitoring results that could be achieved using the entire dataset available from plant B (and no plant A data).

7.4.4.1 Results for Scenario 4

The performance of the PCA model transfer strategy was evaluated assuming that the transfer started when *W* samples had become available from plant B, in such a way that the first monitoring model (k = 1) could be built using these samples along with the plant A samples selected at k = 1 (Figure D.1). Incoming plant B data were then projected onto the model as they were collected, and the model was updated whenever appropriate, as discussed in Section 7.4.3. To assess the monitoring performance, the phase 2 samples of the faulty dataset realizations were presented to the monitoring system after a given number of NOC samples had already been collected from plant B and projected onto the model; namely, it was assumed that the fault onset after 75, 100, 125, 150 or 175 plant B NOC samples had already been presented to the monitoring model and assessed according to the procedure indicated in Figure D.1. It was found that the percentage of the total variance of the $\mathbf{X}_k'^{\mathrm{AB}}$ dataset captured by the PCA models varied between 80.3% and 90.3% (minimum and maximum calculated values, respectively) for a number of PCs ranging from 5 to 6.

The fault detection probability and the mean time to detection are shown as a function of the window width *W* in Figure 7.17a and Figure 7.17b, respectively. Recall that *W* corresponds to the number of consecutive in-control plant B samples used for local modeling. Several curves are reported in the figures, each one being parametric in the number of plant B NOC samples presented to the model before the fault onsets. Note that each curve terminates at a window width equal to the number of NOC samples after which the fault is presented to the system (e.g., if the fault onsets following, say, 75 NOC samples, obviously *W* cannot be set larger than 75 samples).



(a) (b)

Figure 7.17. *Scenario 4: (a) fault detection probability and (b) mean time to detection for the artificial realizations of the fault. The curves are parameterized with respect to the number of plant B NOC samples projected onto the model before the onset of the fault. Results have been averaged over all the realizations.*

Figure 7.17a shows that the fault detection probability of the PCA model is excellent (~100%) if windows larger than ~40 samples are used to build the local model and at least ~100 NOC samples have been projected onto the monitoring model before the fault enters the system. On the other hand, the detection performance is slightly inferior when the fault onsets after only 75 NOC samples, due to the fact that the available NOC samples are not yet enough to entirely describe the normal variability of plant B data.

The mean time to detection results reported in Figure 7.17b show that, if the window width for local modeling is selected appropriately, the fault can be detected promptly and the mean time to detection can be made close to the limiting value of $\Phi = 3$ samples. Overall, the mean time to detection results are consistent with the fault detection probability results: likewise the probability to detect a fault is small when a small window width is used (e.g. W < 30 samples), also large mean time to detection values are obtained with small widths. This is due to the fact that when a fault has a small probability to be detected, several incoming faulty samples are needed before the fault can be detected, which delays the detection; this is why the mean time to detection ≅ 55 samples for small window widths (recall that the total number of faulty samples in phase 2 of the faulty datasets is 57). It can also be noted that for large values of W the delay in fault detection tends to increase. This result may be due to the fact that larger window widths include more plant B samples, which account for a larger variability than those included in small windows; this can cause an undue adaptation of the model to the first samples of a faulty sequence and therefore to a delay in the fault detection.

In Figure 7.18, the percentages of type I and type II errors are reported as a function of the window size W for both the T^2 and the SPE monitoring charts, for the same cases presented in Figure 7.17. It can be seen from Figure 7.18a and Figure 7.18b that the proposed transfer methodology is quite performing with respect to type I errors if windows wider than ~40 samples are used for model adaptation, irrespective of the number of plant B NOC samples that have been projected onto the model before the fault onsets. For narrower windows, the high percentages of type I errors suggest that several NOC samples are erroneously projected out of the confidence limits for both statistics (and particularly for SPE).

Figure 7.18c shows that the percentage of type II errors in the T^2 monitoring chart is significant even for large window widths and large numbers of plant B NOC samples presented to the model. In the case of SPE (Figure 7.18d), the percentage of type II errors is very small if window widths larger than ~40 samples are used and more than ~100 NOC plant B samples have been presented to the model before the fault onsets. For smaller window widths, higher percentages of type II errors are obtained, due to the fact that narrow windows are prone to make the model adapt to faulty data, preventing the correct detection of the fault (as is also highlighted in Figure 7.17a). With respect to type II errors, the number of plant B NOC samples has a larger impact on the results than in the case of type I errors; despite this, the fault detection performance is satisfactory, as was shown in Figure 7.17.

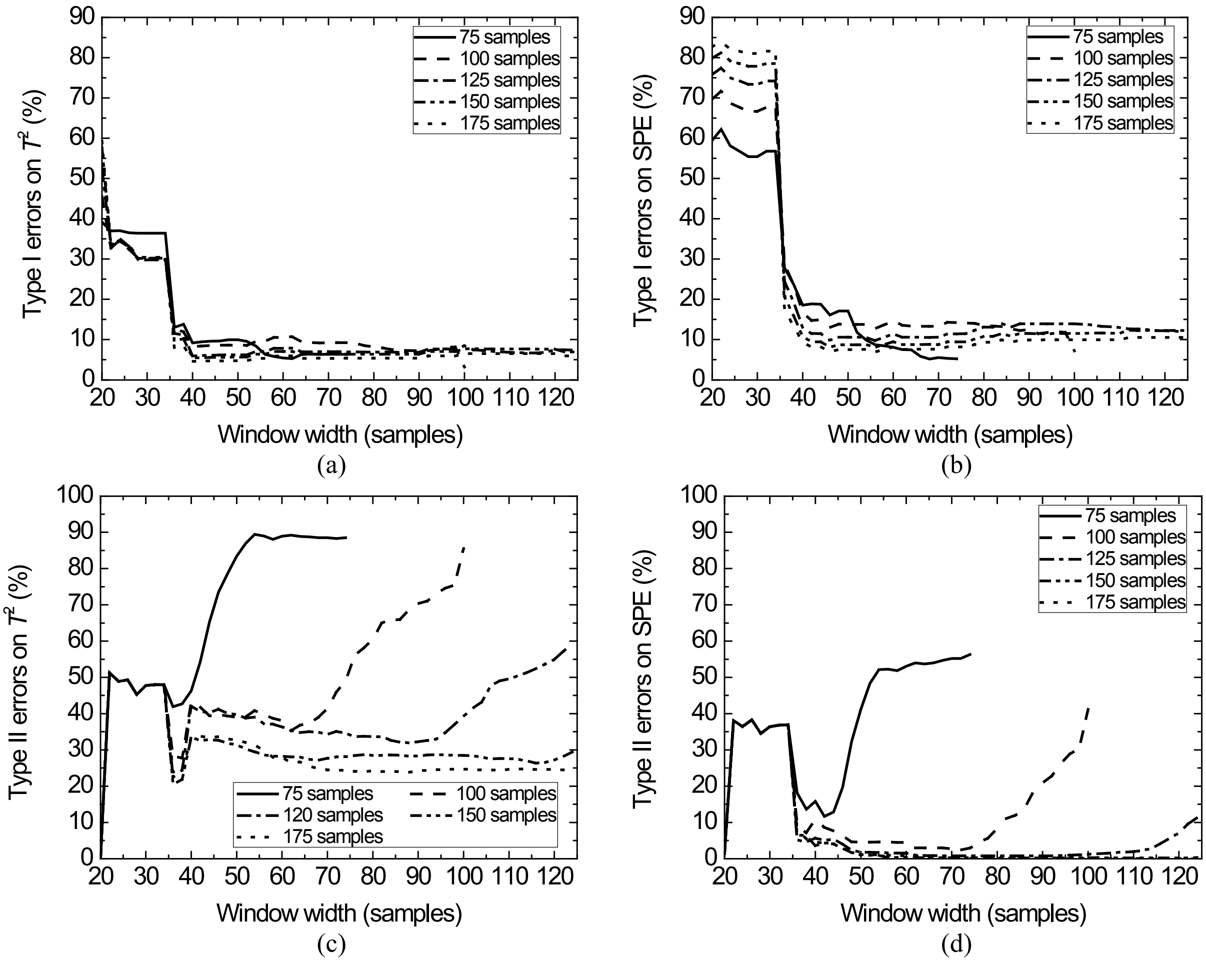


Figure 7.18. *Scenario 4: type I and type II error results for the artificial realizations of the fault. (a) Type I error percentage in the T^2 monitoring chart; (b) Type I error percentage in the SPE monitoring chart; (c) Type II error percentage in the T^2 monitoring chart; (d) Type II error percentage in the SPE monitoring chart. The curves are parameterized with respect to the number of plant B NOC samples projected onto the model before the onset of the fault. Results have been averaged over all the realizations.*

As discussed in Section 7.4.1, the transfer of the monitoring model is based on the selection of the plant A samples that are more similar to those available from plant B on the basis of the values of the plant-independent variable (*wtd*). A study was therefore carried out to understand the impact of the plant A data selection method on the monitoring performance.

The plant A data selection procedure uses a range of *wtd* values identified from the plant B samples, but the limits of this range could be tightened or widened to increase or reduce the number *S* of samples selected. A weight can be used to tighten or widen the range limits: negative values of the weight imply that the limits for *wtd* selection are tightened (hence fewer plant A samples are selected), while positive values imply that they are widened (hence more plant A samples are selected). Note that the results presented in Figure 7.17 and Figure 7.18 were obtained using zero weight.

The effect of the weight (hence of the number of selected plant A samples) on the performance of the monitoring model was studied. The faulty samples (phase 2) were presented to the model after 100 NOC samples had already been collected from plant B and projected onto the model. Results are reported in Figure 7.19 in terms of fault detection probability and fractional number of selected plant A samples, averaged over all fault realizations. The fractional number of selected plant A samples is the ratio between the number *S* of samples selected and the window width *W*. The reported results are parametric in the local model window width *W*, and five different evenly spaced values of *W* were considered, namely 20, 38, 56, 74, and 92 samples.

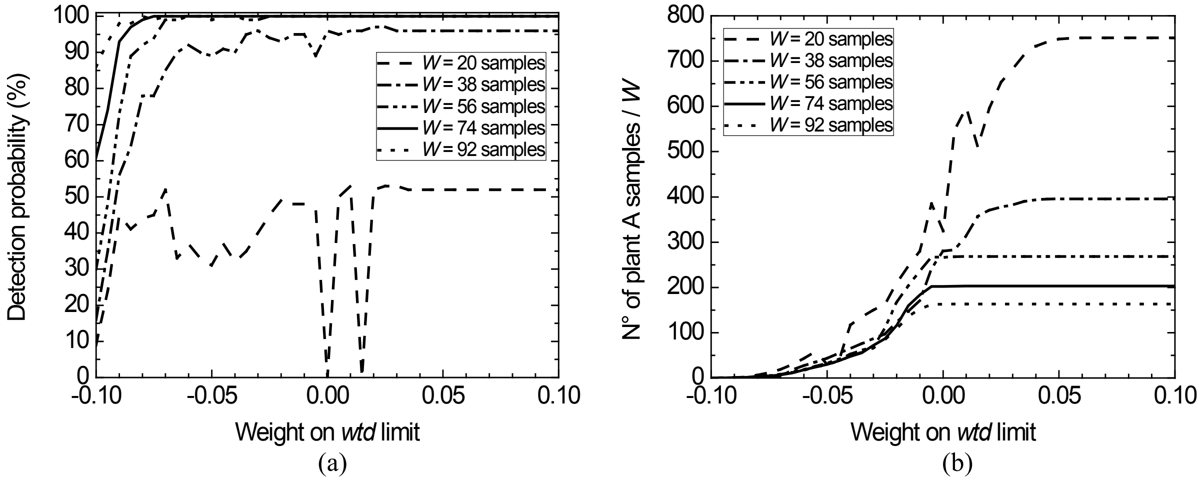


Figure 7.19. *Scenario 4, plant A data selection: effect of the weight on the wtd range limits on (a) the fault detection probability, and (b) the fractional number of selected plant A data.*

Figure 7.19a illustrates that to achieve a satisfactory fault detection probability, sufficiently wide windows of plant B samples ($W \geq 38$ samples) should be used. Loosely speaking, this means that in order to achieve a good detection probability, the monitoring model should keep sufficient memory of past plant B data to assess whether or not adaptation is needed. For any value of *W*, the fault detection probability increases as the weight increases, i.e. as more data from plant A are included in the PCA model. However, note that increasing the weight above a threshold value (which depends on *W*) make *all* plant A be selected. This occurrence is visible in Figure 7.19b because each curve steadies at a constant value. Values of the weight smaller than about –0.08 imply that no samples are selected from plant A, and this significantly decreases the fault detection probability. Also note that even when very small weight values are used (e.g., –0.06), the number of selected plant A samples is not zero (approximately, it ranges from 20 to 45 times the selected window width). It can be concluded that if no plant A data were selected, the resulting adaptive PCA monitoring system based on plant B data only would have lower monitoring performance. This indeed confirms that the transfer of knowledge from plant A to plant B is useful to monitor plant B when not enough NOC data are available from it. Furthermore, good monitoring performance can be achieved even without using all the data available from plant A.

7.4.4.2 Results for Scenario 5

The same procedure adopted to evaluate the performance of the PCA model transfer approach was used also to evaluate the adaptive JY-PLS approach. To build the joint-Y matrix, only the *wtd* values were used, as shown in Figure 7.15. It was found that the models capture a percentage of the total variance of $\mathbf{X}_k^{\mathrm{B}}$ (see Figure 7.15) varying between 39.1% and 90.0%, and a percentage varying between 75.1% and 97.5% of the total $\mathbf{Y}_k^{\mathrm{B}}$ variance, with a number of selected LVs ranging from 3 to 6.

In Figure 7.20, the effect of the local model window size and of the number of plant B NOC data presented to the model before the fault onsets is reported in terms of fault detection probability (Figure 7.20a) and mean time to detection (Figure 7.20b). Good monitoring performance is achieved both in the probability to detect the fault and in the mean time to detection when more than 100 NOC samples from plant B have been collected prior to the onset of the fault. However, even if enough NOC samples are available, good detection probabilities are reached only when *W* is wider than ~90 samples, i.e. for window widths larger than in Scenario 4. This seems to suggest that in this case, if narrow window widths are used, the JY-PLS adaptive approach is more prone to adapt to the faulty data compared to the PCA approach.

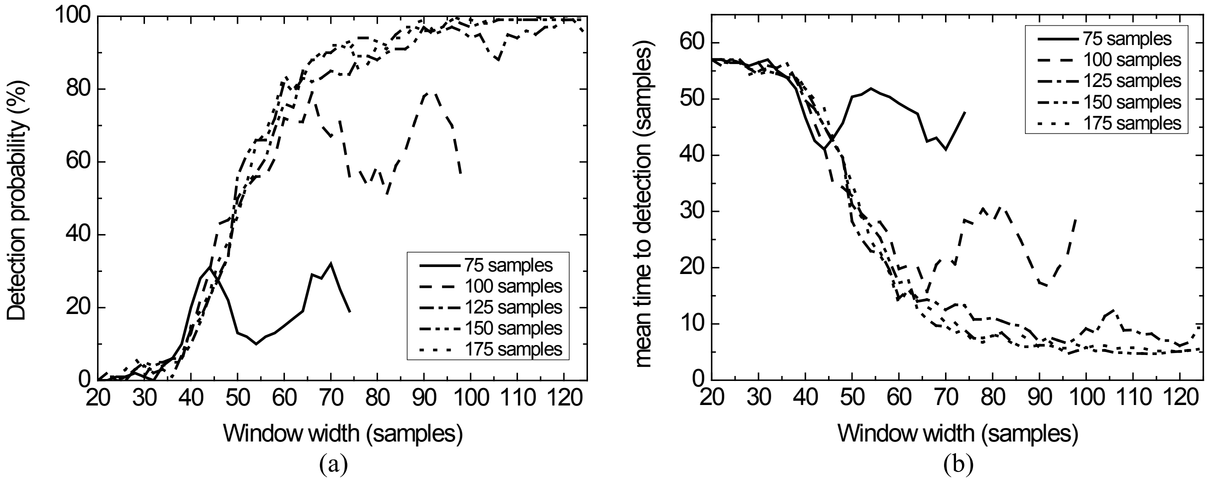
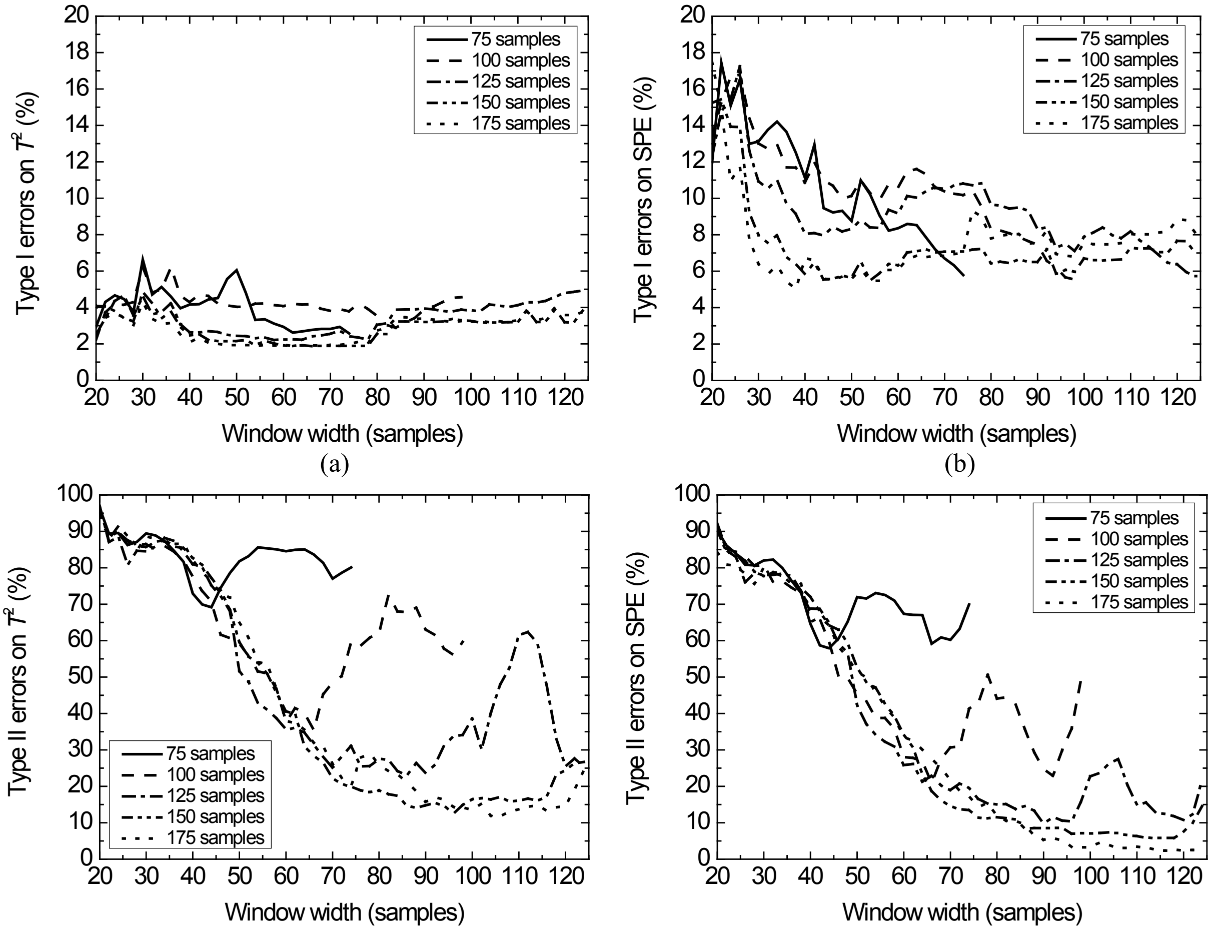


Figure 7.20. *Scenario 5: (a) fault detection probability and (b) mean time to detection for the artificial realizations of the fault. The curves are parameterized with respect to the number of plant B NOC samples projected onto the model before the onset of the fault. Results have been averaged over all the realizations.*

The mean time to detection plot in Figure 7.20b confirms the results provided by the fault detection probability plot. In this case, the smallest values reached for the mean time to detection (~5 samples) are slightly greater than those obtained in Scenario 4, confirming that the JY-PLS approach slightly tends to adapt to the faulty data.

The percentages of type I and type II errors for both the T^2 and the SPE statistics are reported in Figure 7.21. The percentage of type I errors for T^2 (Figure 7.21a) is small (~2 to 5 %), even for small window widths and independently of the number of NOC samples from plant

B. The percentage of type I errors for SPE (Figure 7.21b) is more affected by the selection of the window width, but the smallest percentage reached (<10%) for appropriate values of W are lower than in Scenario 4, even if they are larger than the reference value in Table 7.3 (4.2%).



(a) (b)

(c) (d)

Figure 7.21. *Scenario 5: type I and type II error results for the artificial realizations of the fault. (a) Type I error percentage in the T^2 monitoring chart; (b) Type I error percentage in the SPE monitoring chart; (c) Type II error percentage in the T^2 monitoring chart; (d) Type II error percentage in the SPE monitoring chart. The curves are parameterized with respect to the number of plant B NOC samples projected onto the model before the onset of the fault. Results have been averaged over all the realizations.*

The percentage of type II errors in the T^2 and SPE monitoring charts (Figure 7.21c and 7.21d) is very large (>80%) for small window widths, independently of the number of plant B NOC data available. This means that for small window sizes, high percentages of truly faulty data are projected inside the confidence limits, which confirms that the JY-PLS approach is prone to adapt to the faulty data. Overall, in Scenario 5 a smaller number of type II errors is obtained in the T^2 chart for an appropriate choice of W compared to Scenario 4, whereas a slightly larger number of type II errors in the SPE chart is obtained.

The JY-PLS procedure for the transfer of knowledge between different plants in Scenario 5 seems therefore to require more plant B NOC data to obtain an effective monitoring model than the PCA approach in Scenario 4 (which uses only common variables), and consequently larger window sizes for model adaptation are required. This is due to the fact that the joint correlation between plants and within each plant is more difficult to be captured when also variables that are not common are present, because of the larger variability introduced by the uncommon variables. However, both techniques seem to be very effective in the transfer of knowledge between plants for monitoring purposes. Deeper comparison between the techniques seems improper, due to the differences in the spaces modeled by PCA and by JY-PLS.

7.5 A comparison between scenarios

In order to have a better touch of the relative performances of the monitoring model transfer methods proposed in the previous sections, the presented approaches have been compared on a common basis. Namely, Scenario 2, Scenario 3, Scenario 4 and Scenario 5 have been used to address the model transfer problem, by using the same sets of data, model adaptation mechanism and fault detection criterion, as well as the same diagnostics for the monitoring performance evaluation (Scenario 1 is not considered since it is not adaptive).

All the scenarios have been implemented assuming that W samples were available initially from plant B and following the online monitoring and model adaptation method described in Section 7.4.3 (formerly used for Scenario 4 and Scenario 5). Furthermore, for all the scenarios it was assumed that the fault onset after 75, 100, 125, 150 or 175 plant B NOC samples had already been presented to the monitoring model and assessed according to the procedure indicated in Figure D.1. Monitoring performances have been compared on the basis of the mean alarm rate (AR) and mean time to detection, defined in Section 7.3.4, and averaged over the 100 artificial fault realizations. $\Delta = 5$ is considered as criterion to warn an alarm.

Table 7.4 summarizes the results for the scenario comparison. For each scenario, the minimum window width W is reported, together with the minimum number of initially available plant B NOC samples after which the mean alarm rate profile in phase 1 is steadily below a 5% threshold, while at the same time the mean alarm rate profile in phase 2 is steadily above a 90% threshold[‡]. The corresponding value of the mean time to detection is reported as well.

[‡] Note that the 5% and 95% thresholds for the alarm rate have been used for illustrative purposes. They have nothing to do with confidence limits.

Table 7.4. *Comparison between the performances of the monitoring model transfer scenarios in terms of minimum window width and number of plant B NOC samples required to achieve good monitoring performances. The corresponding mean time to detection is reported as well.*

Scenario	Window width W [samples]	Plant B NOC samples [samples]	Mean time to detection [samples]
2	52	100	4.2
3	68	125	2.7
4	40	125	4.4
5	68	100	5.1

From the analysis of the results in Table 7.4 some general considerations can be drawn.

- Overall, all scenarios show satisfactory monitoring performances when at least 100 samples from plant B are available, and considering at least 70 past samples in the window to update the model each time the adaptation mechanism requires it. The mean time to detection values conform to the criterion selected to highlight the fault, except for Scenario 3, where on average the fault is detected earlier (for the same reason described in Section 7.3.4.1).
- The PCA-based methods requires in general smaller windows than the JY-PLS-based methods to achieve good performances. Otherwise stated, the JY-PLS approaches are more prone to adapt to the data (and therefore also to faulty data) if not enough samples are available in the window. This is mainly due to the differences in the latent space modeled by PCA (which captures the actual correlation structure between variables), and JY-PLS. Furthermore, only common variables are considered in the PCA cases, which may require less samples to fully observe the data latent structures.
- Considering the PCA-based approaches, Scenario 2 requires less NOC samples from plant B than Scenario 4, even if the required window width is larger. Therefore, in this case study, the PCA-based transfer approach does not seem to benefit from the plant A samples selection mechanism based on *wdt*.
- Considering the JY-PLS-based approaches, Scenario 5 shows very satisfactory performances compared to Scenario 3. In this case the use of the plant-independent variable *wdt* looks advantageous. However, it must be noted that the two scenarios are characterized by a completely different matrix structure (as can be seen from Figure 7.3 and Figure 7.15). In Scenario 3, variables are divided in common response variables between the plants and “other” variables, proper of each plant. Furthermore, the process is monitored only on the joint space of the common response variables. This may worsen the monitoring performances, which depend on the observability of the fault in the joint space. Differently, in Scenario 5 the process is monitored on the latent space estimated from the plant B available data ($\mathbf{X}_W^{\mathrm{B}}$), which includes all the plant B variables. In this way, and thanks to *wdt*, the full plant B covariance structure is captured better.

The results confirm that the PCA-based approaches are preferable to support the transfer, as they are very effective in detecting possible faults even if limited plant B NOC samples are available. However, they can be used only with common variables and they can be reliable only if the fault leaves a signature on their correlation structure. In any other case, JY-PLS is better, especially if plant-independent variables are used to match different plant datasets.

7.6 Extension to batch processes

The approaches described earlier have been conceived with the aim of transferring information between different plants to monitor the steady state operation of a process. This implicitly requires that the monitored process is continuous. However, the general framework proposed in Figure 7.1 can be extended to consider the transfer between *batch* processes. In this Section, a study on the extension of part of the framework to batch processes is reported. Although results are still preliminary, it was nevertheless decided to make them available to the reader to complete the analysis of the monitoring model transfer issue.

7.6.1 Batch fermentation process and available data

The case study under investigation is a simulated fed-batch fermentation process for the production of penicillin. A detailed description of the process can be found in Birol *et al.* (2002) and in Çinar *et al.* (2003). The operation goes through two operating stages: the first stage is a batch phase for biomass growth, consuming oxygen and the initial substrate; the second stage is the fed-batch production of penicillin in the absence of substrate. The penicillin is produced in a well-mixed reactor where substrate and air are fed in a controlled environment. A control system keeps the reactor temperature and pH at desired values. A batch is considered terminated when the penicillin concentration attains the assigned target (1.1 g/L for plant A and 0.74 g/L for plant B).

Data on two different plants were obtained using the PenSim[§] simulator, which solves a detailed mechanistic model of differential-algebraic equations describing the biological behavior of the process. Two plants were simulated, which differ for scale, instrumentation, and control system. Plant A is a smaller plant with a culture volume of 105 L, whereas plant B has an average culture volume of 195 L. Table 7.5 reports the process variables that are obtained as simulation (i.e. measured) outputs. Not that some variables are measured in both plants, while other are measured in only one of the two plants. Table 7.5 reports the indication of the plant in which the variable is assumed to be measured, together with the classification between common variables (symbol †) and other variables.

[§] http://simulator.iit.edu/web/pensim/index.html

Table 7.5. *Variables measured in the simulated process for the production of penicillin. Variables that are considered common between the plants are marked by* †.

Var. no.	Measured variable		Plant A	Plant B
1	Aeration rate (L/h)	†	✓	✓
2	Agitation power (W)	†	✓	✓
3	Substrate feed rate (L/h)	†	✓	✓
4	Substrate feed temperature (K)	†	✓	✓
5	Glucose concentration (g/L)			✓
6	Concentration of dissolved oxygen (mmol/L)			✓
7	Biomass concentration (g/L)			✓
8	Penicillin concentration (g/L)	†	✓	✓
9	Culture volume (L)	†	✓	✓
10	Carbon dioxide concentration (mmol/L)	†	✓	✓
11	Fermentor pH (-)		✓	✓
12	Fermentor temperature (K)		✓	✓
13	Generated heat (kcal)			✓
14	Acid flowrate (L/h)	†	✓	✓
15	Base flowrate (L/h)	†	✓	✓
16	Coolant/heating flow rate (L/h)		✓	

Whereas the fermentor temperature control system is assumed to be the same in the two plants, the pH control system is an on-off controller in Plant A and a proportional-integral-derivative (PID) controller in Plant B, with the same settings as indicated by Birol *et al.* (2002). The coupling of controlled and manipulated variables for both controllers is the same in the two plants. Namely, the reactor temperature is controlled by manipulating the heating/cooling water flowrate in the reactor jacket, while pH is controlled by adjusting the concentrated acid/base flowrate entering the reactor.

Finally, different initial conditions are used to simulate the two plants in terms of biomass and carbon dioxide availability, aeration rate, and substrate feed rate and inlet temperature. Details on the simulations are provided in Appendix E.

One hundred simulations have been carried out for both plant A and plant B, obtaining the time trajectories of the measured outputs. For a given plant, the trajectories differ due to noise and to different initial conditions. This causes the batch length to vary from batch to batch for a given plant. The length of plant A batches ranges between 250 h and 320 h, and the length of plant B batches ranges between 200 h and 315 h. Variables trajectories have therefore been synchronized before proceeding with any analysis. The indicator variable approach (Nomikos and MacGregor, 1995b) has been adopted to synchronize time trajectories. Through this method, variable trajectories are reported as a function of the indicator variable, which is an index of the percentage of batch completion. As a result each variable trajectory is not reported as a function of time, where batches have different length, but as a function of the percent of batch completion, where batches have the same length. In particular, considering the dynamic of the process, variable trajectories have been resampled in K = 200 aligned samples, corresponding to increasing percentages of batch completion. Further details on the indicator variables and on the synchronization procedure are provided in Appendix E.

Two datasets of NOC batches for plant A and plant B have been obtained.

- $\underline{\mathbf{X}}^{\mathrm{A}}$ $[100 \times 12 \times 200]$ includes 100 plant A NOC batches, for which 12 variables are reported, each with 200 samples (after alignment);
- $\underline{\mathbf{X}}^{\mathrm{B}}$ $[100 \times 15 \times 200]$ includes 100 plant B NOC batches, for which 15 variables are reported each with 200 samples (after alignment).

In addition to NOC batches, 6 faulty batches have been simulated for Plant B. The following faults were considered: *i*) anomalies of the aeration systems (fault #2, fault #3 and fault #4); *ii*) anomalies of the substrate feeding systems (fault #5); *iii*) anomalies of the agitation system (fault #1); *iv*) anomalies on sensors (fault #6). The fault characterization is summarized in Table 3. Note that some faults are sustained (e.g., fault #1), some are very mild (e.g., fault #5). Furthermore, fault #6 is a sensor fault which affects only variable 7. Namely, the relevant batch is a normal batch, but a step has been appended to variable 7 only, once the simulation results have been obtained, to simulate a sensor damage. In this way, the fault does not affect the other monitored variables.

Table 7.6. *Faulty batches of Plant B: process variable affected by the anomaly, amplitude and type of anomaly.*

Fault no.	Process variable affected by fault	Type of anomaly	Fault amplitude
#1	2	step	-15%
#2	1	step	-25%
#3	1	ramp	-0.5
#4	1	ramp	-1.0
#5	3	ramp	-0.001
#6	7	step	-70%

Whatever a fault, it onsets 100 h after the start of a batch and is protracted until the end of the batch. Therefore, the first 100 h of operation always refer to normal conditions, and will be referred to as phase 1 of the batch in the following. On the other hand, phase 2 of a batch will refer to the fraction of the batch after the occurrence of the fault.

7.6.2 Transfer methodology

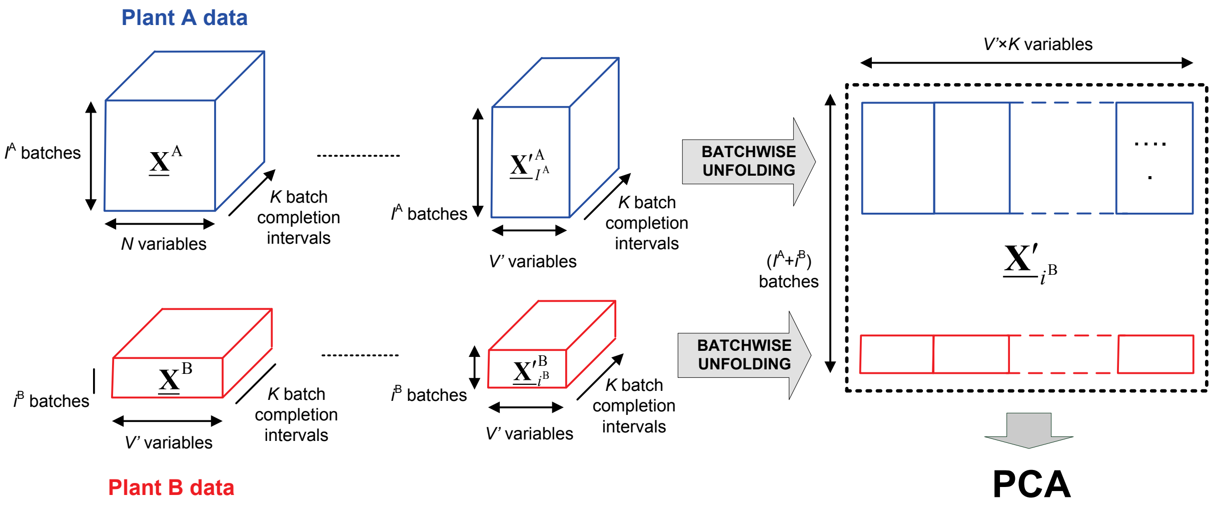
As in the spray-drying case-study, here the objective is to exploit the data available from plant A, where an extended experimental campaign has already been completed, to monitor plant B, where the production is assumed to have just been transferred.

Transfer scenario 2 and transfer Scenario 3 of the general framework of Figure 7.1 are referred to in this case study; therefore, only common variables (Scenario 2) or both common as well as other variables (Scenario 3) are considered for monitoring, with no first-principle information. Set $v' = \{1, 2, 3, 4, 8, 9, 10, 14, 15\}$ identifies the V' common variables between plants (Table 7.5). The remaining variables are assigned to either plant. Namely, $v''^{\mathrm{A}} = \{11, 12, 16\}$ are the V''^{A} variables measured only in plant A, whereas $v''^{\mathrm{B}} = \{5, 6, 7, 13\}$ are the V''^{B} variables measured only in plant B. Note that, despite being measured in both plants, controlled variables are not assigned to the set v' of common variables, due to the mentioned differences in the control systems.

Given the peculiar nature of the considered datasets, which include variable trajectories, the multiway versions of PCA (MPCA; Nomikos and MacGregor, 1994) and of JY-PLS (MJY-PLS) are used.

7.6.2.1 Scenario 2: process monitoring in plant B using MPCA

In this case the aim is to exploit the information embedded in common variables only to transfer information useful to monitor plant B. It is assumed that a number I^{A} of completed batches is available from plant A, while i^{B} batches have been completed in plant B, with $i^{\mathrm{B}} << I^{\mathrm{A}}$. The approach is the same described in Section 7.3.2.2, with the difference that the model is multiway, and the rationale is summarized in Figure 7.22.



Plant A data
I^A batches
$\underline{\mathbf{X}}^{\mathrm{A}}$
N variables
K batch completion intervals
I^A batches
$\underline{\mathbf{X}}'^{\mathrm{A}}_{I^{\mathrm{A}}}$
V' variables
K batch completion intervals
BATCHWISE UNFOLDING
V'×K variables
(I^A+i^B) batches
$\underline{\mathbf{X}}'_{i^{\mathrm{B}}}$
i^B batches
$\underline{\mathbf{X}}^{\mathrm{B}}$
V' variables
K batch completion intervals
i^B batches
$\underline{\mathbf{X}}'^{\mathrm{B}}_{i^{\mathrm{B}}}$
V' variables
K batch completion intervals
BATCHWISE UNFOLDING
Plant B data
PCA

Figure 7.22. *Schematic of the multiway PCA applied in Scenario 2 of the proposed framework for the transfer of monitoring models between batch processes.*

The three-way datasets $\underline{\mathbf{X}}'^{\mathrm{A}}_{I^{\mathrm{A}}}$ $\left[I^{\mathrm{A}} \times V' \times K\right]$ and $\underline{\mathbf{X}}'^{\mathrm{B}}_{i^{\mathrm{B}}}$ $\left[i^{\mathrm{B}} \times V' \times K\right]$ are considered, which are subsets of the simulated databases $\underline{\mathbf{X}}^{\mathrm{A}}$ and $\underline{\mathbf{X}}^{\mathrm{B}}$; they include I^{A} and i^{B} batches (respectively), and consider only common variables. First, batch-wise unfolding (Nomikos and MacGregor, 1994) is applied to the datasets to generate the bi-dimensional matrices $\underline{\mathbf{X}}'^{\mathrm{A}}_{I^{\mathrm{A}}}$ $\left[I^{\mathrm{A}} \times \left(V' \cdot K\right)\right]$ and $\underline{\mathbf{X}}'^{\mathrm{B}}_{i^{\mathrm{B}}}$ $\left[i^{\mathrm{B}} \times \left(V' \cdot K\right)\right]$, where $K = 200$. The data in each matrix are auto-scaled according to the mean and the standard deviation values of the plant in which they are collected. A PCA model is then built on matrix $\mathbf{X}'_{i^{\mathrm{B}}}$ $\left[\left(I^{\mathrm{A}}+i^{\mathrm{B}}\right)\times\left(V'\cdot K\right)\right]$ built according to Eq.(7.2).

The same steps described in Section 7.3.2.2 apply for monitoring of the $\left(i^{\mathrm{B}}+1\right)$-th batch in plant B at each aligned sample *k*. Recall that in the continuous case (Section 7.3.2.2) the model is adapted each time a new set of measurements is available from plant B (i.e., adaptation is done at *k*). In the batch case, the situation is different: the model is updated after completion of a batch using *all* the data from that batch.

Considering that *K* is the total number of indicator variable intervals representing the batch duration, vector $\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1}$ $\left[\left(V'\cdot K\right)\times 1\right]$ includes the trajectories of the V' variables for the entire history of the batch:

$$\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1}=\left[x'^{\mathrm{B}}_{i^{\mathrm{B}}+1,v'\cdot k}\ x'^{\mathrm{B}}_{i^{\mathrm{B}}+1,(v'+1)\cdot k}\cdots x'^{\mathrm{B}}_{i^{\mathrm{B}}+1,v'\cdot(k+1)}\ x'^{\mathrm{B}}_{i^{\mathrm{B}}+1,(v'+1)\cdot(k+1)}\cdots x'^{\mathrm{B}}_{i^{\mathrm{B}}+1,V'\cdot K}\right]^{\mathrm{T}} \quad , \tag{7.15}$$

with $v'=1,...,V'$ and $k=1,...,K$. The online monitoring of the $\left(i^{\mathrm{B}}+1\right)$-th batch at sample *k* is performed by projecting vector $\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$ $\left[\left(V'\cdot K\right)\times 1\right]$ onto the space of the model PCs. This vector includes the unfolded trajectories of all the V' variables until aligned sample *k*. Since at the *k*–th interval measurements for the $\left(i^{\mathrm{B}}+1\right)$-th batch are available only from the beginning of the batch up to the aligned sample *k* itself, the elements of $\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$ for samples from $k+1$ to *K* are missing. In order to allow the batch monitoring, the missing data are filled by assuming that, at interval *k*, the deviation of each variable from the mean trajectory remains unchanged for the rest of the batch duration and is protracted until the end of the batch (Nomikos and MacGregor, 1995). Other missing data approaches can be used as an alternative (Arteaga and Ferrer, 2002; García-Muñoz *et al.*, 2004).

7.6.2.2 Scenario 3: process monitoring in plant B using MJY-PLS

Multiway JY-PLS (MJ-YPLS) allows to use common variables as well as “other” variables (i.e. variables that are measured only in one plant). Furthermore, as in MPCA, it allows to consider also the dynamic characteristics of the batch process within each plant.

The approach is the same described in Section 7.3.3.1 and represented in Figure 7.3. In this case, the joint-Y space of the model is formed by the common variable space. Therefore, with reference to Figure 7.3, matrix $\mathbf{X}'_{i^{\mathrm{B}}}$ built as in the MPCA case (Figure 7.22) corresponds to the joint-Y space. The two JY-PLS model regressor spaces are represented by the datasets of “other” variables, measured only in one of the plants (analogously to $\mathbf{X}''^{\mathrm{A}}$ and $\mathbf{X}''^{\mathrm{B}}_{k}$ in Figure 7.3). For plant A, matrix $\mathbf{X}''^{\mathrm{A}}_{I^{\mathrm{A}}}$ $\left[I^{\mathrm{A}}\times\left(V''^{\mathrm{A}}\cdot K\right)\right]$ is considered, which includes the I^{A} batches available from plant A database $\underline{\mathbf{X}}^{\mathrm{A}}$ and the V''^{A} variables considered only in plant A. Similarly, matrix $\mathbf{X}''^{\mathrm{B}}_{i^{\mathrm{B}}}$ $\left[i^{\mathrm{B}}\times\left(V''^{\mathrm{B}}\cdot K\right)\right]$ is considered for plant B, which includes the i^{B} batches available from plant B and the V''^{B} variables measured only in plant B.

The datasets are batch-wise unfolded (Nomikos and MacGregor, 1994) to generate the corresponding bi-dimensional matrices $\mathbf{X}''^{\mathrm{A}}_{I^{\mathrm{A}}}$ and $\mathbf{X}''^{\mathrm{B}}_{i^{\mathrm{B}}}$. Before applying any analysis, the data in each matrix are pretreated according to the mean and the standard deviation values of the plant in which they are collected.

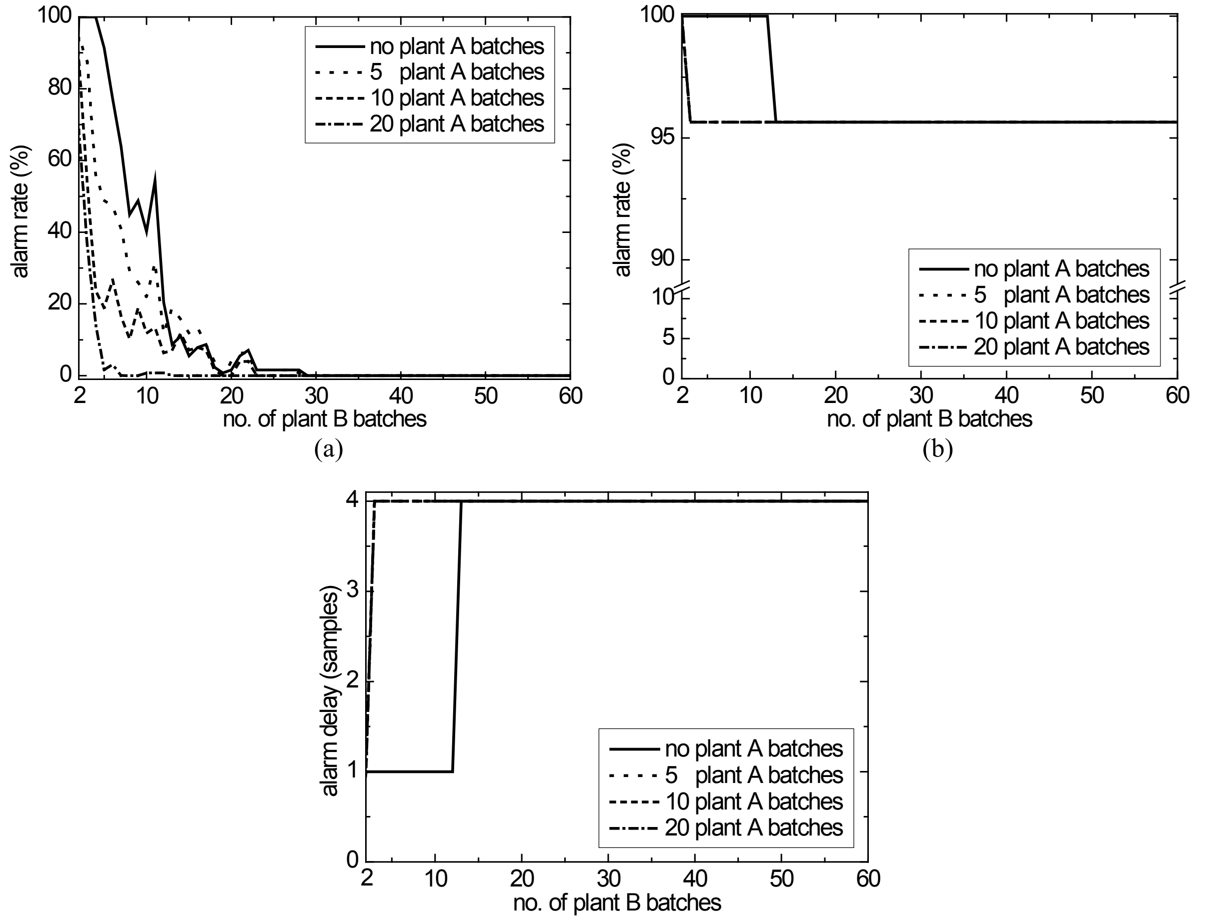
The same procedure described in Section 7.3.3.1 applies in this case for monitoring. For each aligned sample *k*, measurements available from the running $\left(i^{\mathrm{B}}+1\right)$-th batch are organized in vectors $\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$ and $\mathbf{x}''^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$, depending on whether the variables are common or not between the plants. The missing values in $\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$ and $\mathbf{x}''^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$ for instants from $k+1$ to K are filled according to the same procedure described in the previous section. Vector $\mathbf{x}''^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$, which includes the variables measure only in plant B, is projected and reconstructed through the model and used to predict the common variables vector $\hat{\mathbf{x}}'^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}$ (Eqs.(7.3)-(7.4)). In this case the monitoring is performed both in the space of the common variables (the joint-**Y** space, as done in Scenario 3) and in the space of the variables measured only in plant B (as done in Scenario 5). As a consequence, in addition to the Hotelling's T_k^2 statistic, two different squared prediction error statistics are calculated for each aligned sample *k* ($\mathrm{SPE}_k^{\mathbf{x}'^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}}$ and $\mathrm{SPE}_k^{\mathbf{x}''^{\mathrm{B}}_{i^{\mathrm{B}}+1,k}}$), and the relevant monitoring charts are therefore considered.

7.6.3 Results and discussion

The results on the performance of the monitoring models in the transfer are reported using the same indices described in Section 7.3.4, namely the alarm rate in phase 1 and phase 2 of the fault and the time to detection. The same criterion for fault detection is used as well. Namely, an alarm is warned by the monitoring model at interval *k* when $\Delta-1=4$ out of the last $\Delta=5$ consecutive samples from plant B lie outside the 95% confidence limit on one of the monitoring charts.

7.6.3.1 Results for Scenario 2 (fault #1)

An MPCA model to monitor the operation of plant B was built on 2 principal components, selected through cross-validation (Chapter 2, Section 2.1.1.2). Results in terms of alarm rate and of time to detection of the fault are presented in Figure 7.23; they refer to the case where fault #1 of Table 7.6 is presented to the model. The aim of the analysis is also to explore the influence of the number of batches available from plant A and plant B on the monitoring of plant B. For this reason, Figure 7.23 shows the performance of the transfer methodology assuming that an assigned number of normal plant B batches has already been completed and used to build the MPCA model. The results are parametric in the number of normal batches available from plant A.



(a)

(b)

(c)

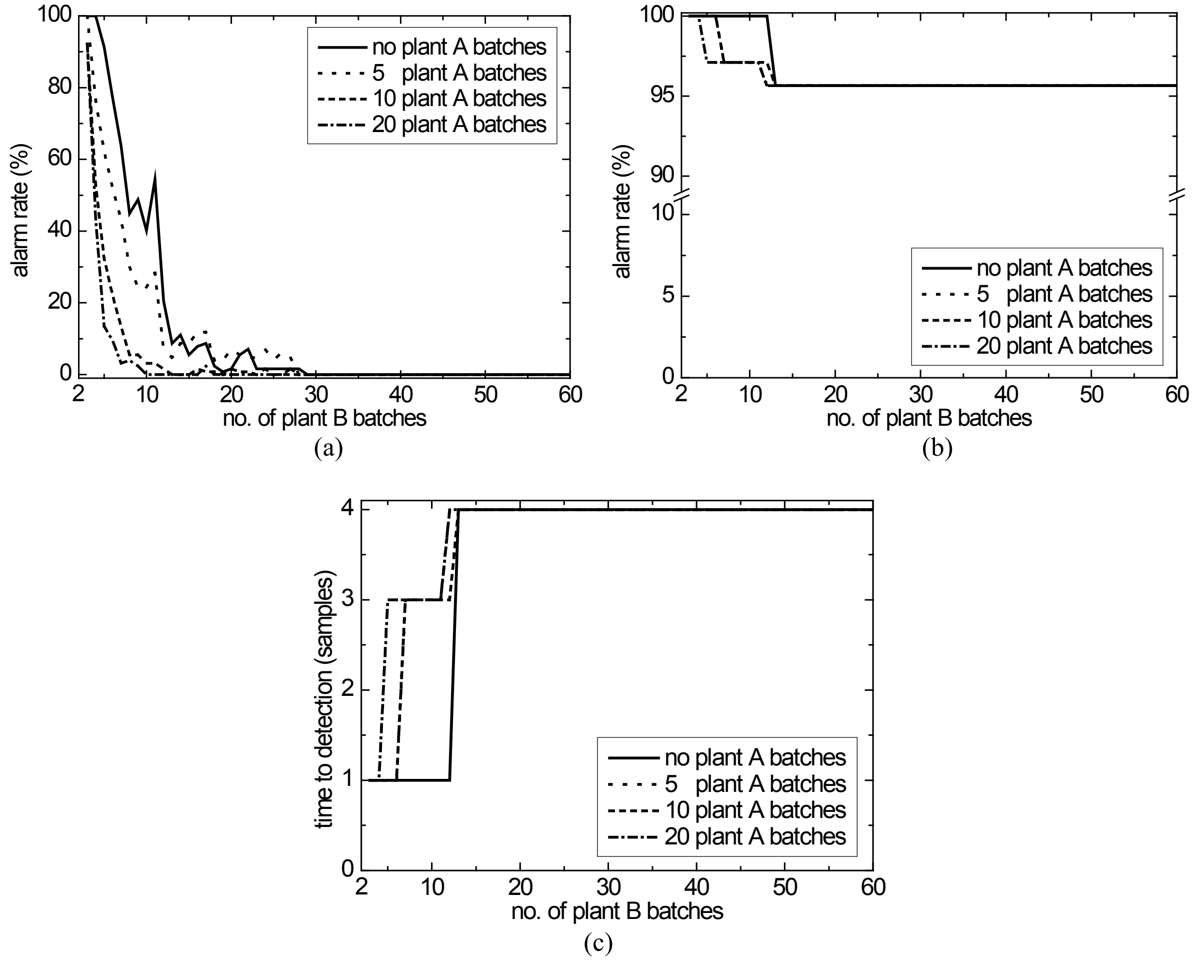
Figure 7.23. *Scenario 2: monitoring performance on a plant B batch affected by fault #1. Effect of the number of available plant A and plant B normal batches on (a) the alarm rate during phase 1, (b) the alarm rate during phase 2, and (c) the time to detection.*

Figures 7.23a and 7.23b show that the availability of some normal batches from plant A is highly beneficial to the monitoring of plant B if the proposed transfer methodology is employed. If no datasets pertaining to normal plant A batches are available (solid lines), the monitoring system requires several plant B batches to provide a satisfactory performance. In fact, about 13 plant B batches (i.e. about 3300 hours of operation) are needed to have a sufficiently low alarm rate during phase 1. Including even few batches from Plant A improves the monitoring performances. The more plant A batches are available, the fewer plant B batches are needed: twenty normal batches from plant A decrease to 4 the number of required normal plant B batches in order to obtain a good monitoring performance, with a time saving of about 2300 hours on experimentation.

The time to detection (Figure 7.23c) shows that when some batches from Plant A are available for the transfer ($I^{\mathrm{A}} > 13$), the delay to warn the alarm asymptotically reaches the expected value of $\Delta - 1 = 4$ samples. Note that, when very few batches from Plant B are completed, the time to detection is small because false alarms are warned during phase 1 of the fault.

7.6.3.2 Results for Scenario 3 (fault #1)

An MJY-PLS model to monitor the operation of plant B was built using 2 LVs, selected through cross-validation (Chapter 2, Section 2.1.3.2). Figure 7.24 reports the alarm rate calculated in phase 1 ad phase 2 and the time to detection for fault #1 of Table 7.6. Recall that the AR is generated after analysis of three monitoring charts: the Hotelling T^2 chart, the SPE chart in the space of the variables measured only in plant B, and the SPE chart in the joint-Y space.



(a)

(b)

(c)

Figure 7.24. *Scenario 3: monitoring performance on a plant B batch affected by fault #1. Effect of the number of available plant A and plant B normal batches on (a) the alarma rate during phase 1, (b) the alarm rate during phase 2, and (c) the time to detection.*

Similar consideration on the value of using plant A batches can be drawn in Scenario 3 as were done in Scenario 2 (Figure 7.24a,b,c). In fact, the use of normal batches from plant A improves the monitoring performance. The performance of the monitoring model transfer especially improves in phase 1 (Figure 7.24a) as more batches are available from plant A.

However, the MJY-PLS model seems to require slightly more normal batches from plant B to obtain the same performance achieved with MPCA for an assigned number of batches from Plant A. For example, assuming that 20 batches from plant A are available, a good monitoring performance is achieved when at least 6 normal batches from Plant B are completed and included into the model through the proposed updating procedure. This can be explained considering that the inclusion in the joint-Y model of variables that are measured only in one plant adds an additional source of variability to be captured by the model. This variability is different from the variability in the common variables considered in the MPCA model of Scenario 3.

Despite the (slight) superiority of the monitoring performance in transfer Scenario 2, transfer Scenario 3 can solve one of the major drawbacks of the use of MPCA in Scenario 2. In fact, if a fault manifests itself only on variables that are not common and the common variables are not correlated with the faulty variables, an MPCA model cannot detect the fault. Instead, transfer Scenario 3 can efficiently detect the effect of the fault in the space of the variables that are measured only in plant B, by exploiting the MJY-PLS model.

To clarify this issue, consider fault #6, which is a sensor fault affecting only variable 7 (measured only in Plant B) and does not leave any signature on other (possibly correlated) variables. Clearly, the fault cannot be detected by an MPCA monitoring model: Figure 7.25a shows that the alarm rate in phase 2 is unsatisfactorily small even when several normal batches from plant A and plant B are used. This is because the variable through which the fault is visible is not included in the model and does not affect any other measurement included in the model. On the other hand, an MJY-PLS model (transfer Scenario 3, Figure 7.25b) is appropriate to monitor Plant B: few normal batches from Plant A together with about 10 normal batches from plant B are able to ensure the expected monitoring performance.

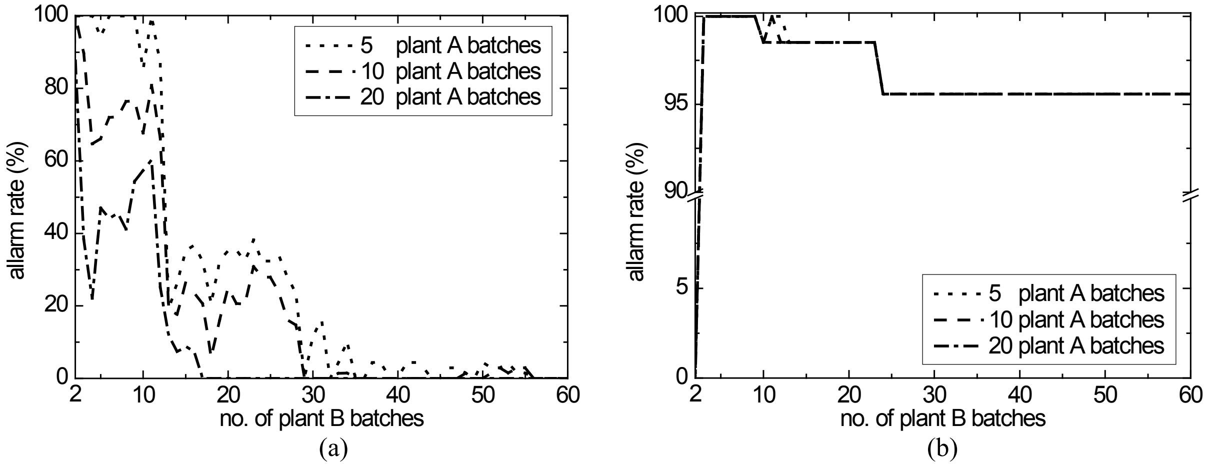


Figure 7.25. *Alarm rate during phase 2 for the monitoring of fault #6 through: (a) transfer Scenario 2; (b) transfer Scenario 3.*

Note that the selection of variables to be included in the model and the partition between “common” and “other” variables can deeply affect the effectiveness of the transfer and the performance of the monitoring model. For this reason, some form of engineering knowledge on the system under investigation is always needed to guide the selection and the partition of variables.

7.6.3.3 Scenario 2 and Scenario 3 results for other faults

The results reported in the previous subsections can be extended to the other faults considered in this study (Table 7.6). Table 7.7 shows the summary of the performances of the monitoring transfer in transfer Scenarios 2 and 3, in terms of alarm rates (AR) in phase 1 and phase 2, and time to detection (TD). To support the comparison, in both transfer scenarios the monitoring models models were built on 2 LVs, and it was assumed that 20 normal batches were available from Plant A and 10 normal batches were available from Plant B. Note that faults #1 and #2, which are step-like, are detected very well: during phase 1 (normal operating conditions) a very small rate of alarms is warned, while the alarm rate in phase 2 is close to 100%.

Table 7.7. *Comparison between the monitoring performance of transfer Scenarios 2 and 3 on several faults affecting batch plant B (alarm rates AR and times to detection are calculated for 20 batches available from Plant A and 10 batches available from Plant B).*

	MPCA transfer (Scenario 2)			**MJ-YPLS transfer (Scenario 3)**		
Fault no.	**AR Phase 1 (%)**	**AR Phase 2 (%)**	**TD (samples)**	**AR Phase 1 (%)**	**AR Phase 2 (%)**	**TD (samples)**
1	0.7	95.7	4	0	97.1	3
2	0	97.1	3	0	95.7	4
3	0	23.1	54	3.1	15.9	59
4	14.1	76.8	1	14.1	24.6	53
5	5.5	0	∞	3.9	1.4	64
6	39.1	57.4	30	52.3	98.5	2

With respect to the ramp-like faults #3, #4, and #5, the ramp determines larger delays in fault detection (high TD). This is due to the nature of the faults and to the fact that these faults are much less pronounced. Generally speaking, if the ramp slope is small, higher detection delays are expected. Large values of TD are generally coupled to lower alarms rates during phase 2, while the rate of alarms in phase 1 are acceptably small.

It is important to highlight that in all cases the transfer is useful and improves the monitoring performance with respect to a monitoring model built on data of Plant B only. Furthermore, the transfer through Scenario 2 (MPCA) seems to be faster to adapt to the plant B process conditions than the one through Scenario 3 (MJY-PLS).

7.7 Conclusions

In this Chapter the issue of transferring a process monitoring model from a reference plant (plant A), where a large amount of normal operating conditions data is available, to a target plant (plant B), where a limited amount of process data have been collected (e.g., because the plant operation has just been started) has been addressed. This would respond to the need of having the operation in plant B monitored as quickly as possible.

A framework has been proposed according to which different model transfer scenarios can be considered depending on *i*) the available information (process data only or process data as well as process knowledge in the form of conservation laws), *ii*) the source where the data available for model design come from (plant A only or plant A as well as plant B), and *iii*) the nature of the measured process variables that are used for model design (only variables that are common between the plants or common variables as well as all other variables). Following this framework, five latent variable approaches to transfer a process monitoring model from plant A to plant B were proposed and illustrated in detail. Three of them were based on process data only, while the last two combined process data with fundamental knowledge which can be derived for example from conservation laws.

The proposed model transfer procedures were tested on a benchmark problem related to the scale-up of the monitoring model for an industrial spray-drying process, where plant A is a pilot unit and plant B is a commercial production unit.

Approaches based on process data were applied with the aim of studying the performances of the transfer in monitoring plant B, for a different number of NOC samples available from plant B and a different number of LVs used to build the model. The simplest proposed model transfer approach (Scenario 1) was a PCA one, where only plant A data related to common variables were used to monitor the plant B operation. The monitoring performances were fully satisfactory (the fault was detected soon after its appearance, and the numbers of false alarms and undetected faults were limited). Making the PCA model adaptive (Scenario 2) by incorporation of measurements incoming from plant B did not improve the monitoring performance significantly. Both PCA approaches seemed to work better when only few LVs were retained into the monitoring model.

A JY-PLS approach was used in Scenario 3 in order to exploit not only measurements of process variables that were common between plants, but also all the other variables measured in each plant. This approach allowed to analyze the within-plant correlation (in each plant) jointly with the between-plant correlation, therefore exploiting the information provided by all the available measurement sensors. The monitoring performances were shown to be very good, with the further advantage that this modeling approach required fewer plant B normal operating condition data to design an effective monitoring model than the PCA approaches.

The last two scenarios of the framework based the transfer on the combination of LVMs with fundamental engineering knowledge derived from a simple energy balance in terms of a new physically-meaningful variable that could be considered plant-independent. In these approaches, process monitoring was achieved using adaptive LVMs, namely PCA (Scenario 4, if only the common variables that were measured in both plants were used to build the monitoring model) or JY-PLS (Scenario 5, if both common as well as all other available measurements were used). The definition of a plant-independent variable was shown to be essential for the model transfer, because it helped in quantifying one driving force acting on both plants, allowing to match similar samples from different plants (Scenario 4) or to relate different plant data through the joint space generated by it (Scenario 5). Results showed that in both the scenarios robust and prompt fault detection performances were achieved, even when very few data were available from plant B.

All the proposed scenarios were also compared on a common basis, using them as adaptive models for monitoring online the process in plant B. Results showed that in the considered case-study, the PCA-based approaches performed better than the JY-PLS approaches in terms of initially required plant B data to achieve satisfactory fault detection performances. Moreover, the use of the plant-independent variables was shown to improve the performances especially in the JY-PLS-based scenarios.

The proposed framework was applied preliminarily for the transfer of the monitoring model in a batch process, dealing with the production of penicillin. In particular Scenario 2 and Scenario 3 were applied, using multiway PCA and JY-PLS in order to account for the batch dynamic features. The obtained preliminary results confirmed that considering data from batches run on a similar plant improved the monitoring performances of the model in the target plant, even if a very limited number of batches (5÷10) was available from the target plant. Further research has to be carried out in future to better understand the effect of the partition of variables in common and not common and to extend also Scenario 3 and Scenario 4, considering fundamental knowledge, to batch process monitoring.

Finally, some issues deserve special attention in any of the proposed approaches. The first one refers to the approaches in which the space of common variables is exploited to address the transfer. Because the fundamental driving forces of the process are the same in both plants, it is reasonable to assume that the process leaves nearly the same signature (i.e. correlation structure) on similar variables in each plant. In these cases, it is therefore assumed that the correlation structure between the common variables remains essentially the same between the plants. This assumption may be verified *a priori* by engineering judgment, evaluating the similarity between plants of different sites/scales. Furthermore the model diagnostics of the adaptive procedures (T^2 and SPE) provide a metric to check the validity of the assumption itself during process monitoring.

The second issue is related to data scaling. Although the correlation structure of common variables is nearly the same in both plants, the actual variability of each measured variable may be not the same in both plants (e.g. because of sensor type and location, process layout and actual operating conditions). Therefore, in order for the proposed approaches to work effectively, each variable should be autoscaled using the values of mean and standard deviation related to the plant in which the variable is measured.

A third issue concerns the scenarios which use plant-independent variables to relate different plant data. In the case study analyzed in this Chapter only one plant-independent variable representative of one of the different physical phenomena driving the process was used. It should be noted that using additional plant-independent variables (which might be more representative of the physics of the systems) may improve the transfer performance.

Finally, note that since for the adaptive procedures, the monitoring performance depends on the tuning of some parameters (adaptation window size *W*, number of initially available plant B NOC samples), for the implementation of these tools, *a-priori* engineering knowledge on the system under study is needed to assign first-trial values for them.

Despite the above-mentioned issues, the proposed model transfer strategies can provide a particularly valuable contribution to the practical implementation of QbD methodologies and continuous quality assurance programs in pharmaceutical product manufacturing, where limited data are usually available if new productions are being started.

Conclusions and future perspectives

The introduction of the Quality-by-Design (QbD) philosophy in the pharmaceutical industry has opened the route towards the adoption of systematic tools in pharmaceutical development and manufacturing, with the aim of modernizing and improving the way pharmaceutical products are designed and produced. The aim of the QbD founding paradigms is to provide tools to improve the understanding pharmaceutical scientists have on their products and processes, in order to ensure a robust and tight control on the quality (in terms of physical properties, but especially of efficacy and safety) of the final drug products.

From an engineering perspective, QbD can be seen as the attempt of introducing modeling principles in pharmaceutical development and manufacturing. This offers tremendous opportunities to the pharmaceutical industry, which can benefit from tools and methodologies that other industries have already experienced. At the same time, pharmaceutical productions are characterized by specific features, such as the product complexity, the low volume multi-product (and mainly batch) productions and, above all, the regulatory oversight, that require dedicated tools to address the issues that may arise in such a diversified production environment. For this reason, there is the need to conceive methodologies that are suitable to fit the peculiar characteristics of the pharmaceutical industry, but, at the same time, general enough to be applied in a wide range of situations.

Under this perspective, this Dissertation has proposed to use latent variable models (LVMs) to assist the practical implementation of QbD paradigms in pharmaceutical development and manufacturing. LVMs offer the important advantage that they can be efficiently used to analyze datasets of highly correlated variables that may come from developmental experiments, materials characterization, process operating conditions or historical products.

In particular, this Dissertation has proposed some *general* methodologies based on LVMs to support researchers in the achievement of the three milestones on which the QbD initiative is founded: process understanding, product and process design, and process monitoring and control. Table 1 summarizes the main achievements of the Dissertation, with the indication of the considered application, data origin and reference to related papers that have been published, or are in press or in preparation.

With respect to **process understanding**, in Chapter 3 a general strategy was proposed to apply LVMs in the development of continuous manufacturing systems. An industrial continuous tablet manufacturing line on a pilot scale was used as a test bed for the analysis. A general procedure based on three main steps was proposed: *i*) a data management step, *ii*) an exploratory analysis step, and *iii*) a comprehensive analysis step.

Table 1. *Summary of the main achievements of this Dissertation, with the indication of the considered application, of the data origin and of the relevant references.*

Chapter	Main achievement	Application	Data origin	Reference
Chapter 3	General procedure for the application of LVMs to support the development of continuous processes	Continuous tablet manufacturing line	Industrial	Tomba, E., M. De Martin, P. Facco, J. Robertson, S. Zomer, F. Bezzo and M. Barolo. General approach to aid the development of continuous pharmaceutical processes using multivariate statistical modeling – An industrial case study. *Int. J. Pharm.*, in press. DOI: 10.1016/j.ijpharm.2013.01.018.
Chapter 4	General framework to aid the design and manufacturing of new products through LVM inversion Procedure to design the target quality profile for a new product	High shear wet-granulation process	Industrial	Tomba, E., M. Barolo and S. García-Muñoz (2012). General framework for latent variable model inversion for the design and manufacturing of new products. *Ind. Eng. Chem. Res.*, **51**, 12886-12900. Tomba, E., S. García-Muñoz, P. Facco, F. Bezzo and M. Barolo (2012). A general framework for latent variable model inversion to support product and process design. *Computer Aided Chemical Engineering 30*, (I.D.L. Bogle and M. Fairweather, Eds.), Elsevier, Amsterdam (The Netherlands), p.512-516. Tomba, E., P. Facco, F. Bezzo and S. García-Muñoz (2012). Exploiting historical databases to design the target quality profile for a new product. Submitted to *Ind. Eng. Chem. Res.*.
Chapter 5	LVM inversion for *in-silico* product formulation design	Formulation for an API	Industrial	Tomba E., M. Barolo and S. García-Muñoz. *In-silico* product formulation design through latent variable model inversion. *In preparation*.
Chapter 6	LVM inversion to support product transfer between different manufacturing plants Experimental demonstration of the existence of the null space	Nanoparticle precipitation	Laboratory	Tomba E., N. Meneghetti, P. Facco, T. Zelenkova, D.L. Marchisio, A.A. Barresi, F. Bezzo and M. Barolo. Product transfer between different plants through latent variable model inversion. Submitted to *AIChE J.*. Meneghetti, N., E, Tomba, P. Facco, F. Lince, D.L. Marchisio, A.A. Barresi, F. Bezzo and M. Barolo. Supporting the transfer of products between different equipment through latent variable model inversion. *Computer Aided Chemical Engineering*, in press.
Chapter 7	General framework to transfer process monitoring models between different plants	Spray-drying process	Industrial	Tomba, E., P. Facco, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Combining fundamental knowledge and latent variable techniques to transfer process monitoring models between plants. *Chemom. Intell. Lab. Syst.*, **116**, 67-77. Facco, P., E. Tomba, F. Bezzo, S. García-Muñoz and M. Barolo (2012). Transfer of process monitoring models between different plants using latent variable techniques. *Ind. Eng. Chem. Res.*, **51**, 7327-7339.
		Penicillin fermentation process	Simulated	Facco, P., M. Largoni, E. Tomba, F. Bezzo and M. Barolo. Transfer of process monitoring models between plants: batch systems. *In preparation*.

It was shown how the parameters of the LVMs could be interpreted from first principles, identifying the main driving forces acting on the system and ranking them according to their importance. This can be useful to support risk assessment in providing the rationale for a robust control strategy and to guide further experimentation from the early development phases. It was found that the route chosen to reduce the size of the active pharmaceutical ingredient (API) particles prior to granulation and the point at which the API was formulated were the most important driving forces. From the comprehensive analysis, it was shown how multi-block LVMs can help in identifying the most critical units in the process and the most critical variables within them. Furthermore, it was shown that these tools can be useful in identifying paths along which a process moves, depending on the selected process settings, thus providing a system to ensure that the operation follows the desired path.

The effectiveness of the use of LVMs built on historical datasets in **product and process design** was demonstrated in Chapter 4, Chapter 5 and Chapter 6. In Chapter 4, a general framework to use latent variable regression model (LVRM) inversion to support the design of new products and of their manufacturing conditions was proposed. The aim of the proposed framework was to provide a general tool in which the most appropriate problem can be solved depending on the objectives and the constraints an user may have in the product/process design activity.

Since multiple solutions may be obtained from the inversion, four possible optimization approaches were identified. The objective of the inversion was to estimate the best input conditions (in terms of raw material properties and process parameters) to achieve a desired quality for the output product. The framework was successfully applied to an industrial particle engineering problem for the design of the raw material properties in a high-shear wet granulation process, to obtain granules with specified quality characteristics.

The null space concept, namely the space of the LVRM inversion solutions that correspond to the same desired set of output variables, was investigated. The null space definition has been shown to have many common features with the definition of the design space of a process given by the regulatory Agencies' guidelines, and it has been demonstrated to be a useful tool for its preliminary identification. In order to have a measure of the reliability of the model inversion solution and of the null space solutions, a strategy based on a jackknife procedure was proposed to estimate their uncertainties.

Some possible solutions were also presented to address specific issues related to LVRM inversion. In particular, a new statistic (P^2) was introduced to select the number of LVs in a model suitable for inversion, in order to adequately describe the regressor dataset. Furthermore, since due to the model mismatch it is not guaranteed that model inversion provides a solution satisfying the desired properties for the final product, a strategy was proposed to exploit the historical data covariance structure in the selection of new desirable product properties most suitable for model inversion. The proposed approaches exploited the model parameters and the constraints provided by the user for the product quality in order to estimate new product quality profiles for which the model mismatch was minimum. This helped the model inversion in achieving the desired product properties, since they could be assigned as hard constraints in the optimization problem.

Chapter 5 provided an application of the framework proposed in Chapter 4 to address a pharmaceutical product formulation problem, in which the objective was to estimate the best excipient type and amount in order to obtain a blend of suitable properties for direct compression. The framework of Chapter 4 was then extended to consider constraints for the material selection and to account for the specific objectives a formulation problem may have (e.g., API dose maximization, minimization of the final tablet weight). This changed the model inversion problem into a mixed-integer nonlinear programming problem, for which a user-friendly interface was developed to allow formulators specifying the objectives and the constraints the formulation problem may have. The proposed methodology was tested on an industrial case study to design the formulation for a proprietary API. The *in-silico* designed formulations were then prepared and experimentally tested, providing results in agreement with the model predictions.

In Chapter 6 an additional application of the proposed general LVRM inversion framework was considered. The case study dealt with a product transfer problem, in which the objective was to obtain nanoparticles of desired mean size through a solvent displacement process in a target device. The methodology exploited the historical data available from a source device and from the target one, but obtained under a slightly different experimental setup. A joint-Y PLS (JY-PLS) model was first used to relate data coming from different sources (plants and experimental setups). The JY-PLS model was then included in the inversion framework to determine the process conditions in the target plant that ensure the manufacturing of a product with desired particle size. Validation experiments confirmed the results obtained by simulation. The experiments also allowed to experimentally validate the null space concept, by showing that different combinations of process conditions ensured to manufacture products with the same desired mean particle size.

The final section of the Dissertation proposed an application of LVMs for **process monitoring and statistical process control** of pharmaceutical operations. In particular, in Chapter 7 the problem of the transfer of models for process monitoring was studied. In this case, the problem was to ensure that the operation in a target plant was under statistical control since the first instants from its start-up, by exploiting the knowledge available (in terms of data) from other plants. A general framework was proposed on the use of LVMs to cope with this kind of problems. The framework identified five different scenarios, depending on the type of the available information to support the transfer (only process data or both process data and fundamental knowledge), on the sources of the available data (only from the reference plant or both from the reference and the target plant) and on the process variables used for the model design (only common variables or both common and other variables). PCA and JY-PLS were used to model jointly the data available from the different plants, depending on whether only common variables (in the PCA case) or both common and other variables (in the JY-PLS case), were used to build the model.

The proposed framework was tested on a model transfer problem related to the monitoring of an industrial spray-drying process, where the reference plant was a pilot unit while the target plant a commercial production unit. The monitoring performances were shown to be satisfactory for all the proposed scenarios. In particular, it was demonstrated that the transfer of information from the reference plant improved the monitoring performances of the model.

The proposed methodologies were also extended, in a preliminary study, for batch process monitoring, considering a simulated fermentation process for penicillin production, in which plants of different scale and technology were simulated. Results on the monitoring performances demonstrated again that considering for the model design also data from batches carried out in the reference plant made the model more efficient in the detection of the simulated faults than in the case only data from the target plant were considered.

In summary, it has been shown how LVMs can be employed in any phase of the product and process design activities under a QbD framework, from risk assessment, to the identification of the design space and of the control strategy for a process.

The **most important contribution** of the LVM approaches proposed in this Dissertation for pharmaceutical product and process design and manufacturing is that they can optimally leverage historical data. These data may come from designed experiments or already developed products, and can guide the industrialization of a pharmaceutical product, starting from its design, going through the design of its manufacturing process, and up to the technology transfer from small-scale plants to large-scale plants for its mass production. The proposed tools allow guiding the experimentation from the first stages of the development of a new product, in order to streamline its design and reduce the time between the discovery of the API and the launch of the final product on the market. This can be crucial in the pharmaceutical industry, which is commonly characterized by long times between the approval of the active ingredient and the final drug mass production. Under the same perspective, burdensome activities, like the technology transfer between different plants, can be sped up, and the risk linked to possible drawbacks in the start-up of a large-scale production can be controlled better. This offers undeniable advantages in terms of time and resources saving, but also in a regulatory perspective, since the manufacturing process can be demonstrated to be understood and controlled since its conception. Therefore, thanks to their transparency and scientifically-sound basis, these tools can completely fit in the regulatory framework.

On the other hand, there are some limitations in the use of LVMs. Being them data-based models, they can only be reliable in the range of the historical data used to build the model. This can be a limit, since using the model for design can give sub-optimal solutions, being not ensured that other possible designs out of the historical data range are possible. Similarly, in monitoring applications the model can detect as faults normal operating conditions of the process that were not considered in the historical dataset. To this purpose, the model itself could be used to expand the knowledge space of the historical data, for example by allowing slight extrapolations in model inversion to suggest new experiments to perform.

It must be also considered that all the methodologies used in this Dissertation were mainly linear. This can be a limit, especially for the model performances, if they are used to describe systems with a nonlinear behavior. In this case, appropriate data transformations or the nonlinear versions of the LVMs should be considered.

The studies carried out in this Dissertation have opened further perspectives and issues, that could be considered in future research. One of the most interesting **areas open to further investigation** is the use of LVMs to guide product transfer between different plants. As shown in this Dissertation, this could be particularly useful to accelerate and improve process scale-up activities. Strategies should be conceived by combining advanced LVM tools with classical scale-up approaches, to improve the capabilities of LVMs in handling correlated variables with available mechanistic knowledge (e.g., dimensionless analysis). In Chapter 7 of this Dissertation, a preliminary study has been presented in the case of model transfer. The methodology could be further developed and applied to scale-up products (as in the case study proposed in Chapter 6) or to guide the design of the process operating conditions for large-scale plants, based on small-scale plant data.

Further research is needed to show how to use LVMs in the systematic identification of the design space of a process. In this Dissertation, it has been demonstrated that there is a link between the regulatory design space definition and the LVM null space concept. Strategies to properly define the design space limits (e.g., in the latent space of the model) and to communicate them are still missing. To this end, some feasibility analysis studies (Swaney and Grossman, 1985) have already been proposed for pharmaceutical processes, using data-based models (Boukouvala *et al.*, 2010; Boukouvala and Ierapetritou, 2012). It would be needed to extend these studies to LVMs, in order to show how the latent space of the model can be used to identify a design space and how it is affected by uncertainties.

Additional studies are needed to demonstrate how advanced process control tools (e.g., model predictive control) could be applied to increase the efficiency and the robustness of pharmaceutical manufacturing processes. In the field of LVMs, some studies have already been carried out by Flores-Cerillo and MacGregor (2003, 2004, 2005) for different industrial applications. These contributions could be extended and adapted for pharmaceutical process control.

From a modeling point of view, a LVM methodology to jointly analyze the complex data structures the pharmaceutical development and manufacturing environments generate is still missing. This method should consider in a whole modeling framework raw materials databases, formulation databases, databases of process conditions referred to different unit operations and databases of product quality (preferably both on physical but also on efficacy and safety properties). This methodology could therefore describe different plant configurations and unit operations, from the API synthesis to the final product enhancement operations, and be used as a tool to support product and process design (e.g., through model inversion).

Finally, it would be desirable that some benchmark problems be prepared for the scientific and industrial communities, where the practical implementation of the QbD paradigms could be addressed by using LVMs as well as other process systems engineering tools This would have the potential both to boost academic research on this topic and to make industrial practitioners more confident on the usefulness of these tools for their needs.

Appendix A

On the interpretation of the latent variable model parameters

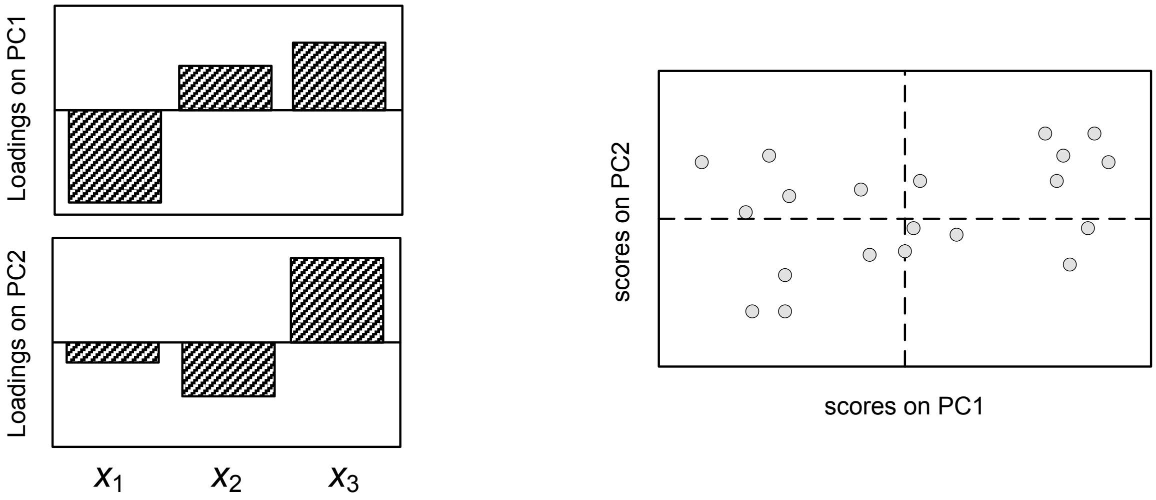
This Appendix reports some details on the interpretation of the parameters of a latent variable model (LVM). In particular, some indications are provided on how to interpret the loading and score diagrams in order to get information from the data. The interpretation of the loading plots of the case study considered in Chapter 4 is used as an example.

A.1 Interpretation of the score and loading plots

PCA and PLS models (Chapter 2) are usually built on multivariate datasets to gain understanding on the system the data have been generated from. This can be achieved by analyzing the correlation between variables and the similarities between samples. The advantage in using LVMs to this purpose is due to the fact that the model structure is transparent and allows to interpret the correlation structure in a straightforward way. From the analysis of the model parameters, an interpretation on the mechanisms acting on the system can be drawn.

For the purpose of a practical application of PCA and PLS, the analysis of the scores and of the loadings of the model is crucial. In general, this is done by considering plots of the scores and of the loadings, which can be reported in several ways. According to common practice (which is adhered to in this Dissertation), the scores are reported as scatter plots, in which the scores on a PC (or on a LV indifferently) are reported versus the scores on another PC. This is usually done for the scores on the first LVs found by the model, because they explain most part of the variability in the data. Bi-dimensional plots are usually used as they are easier to visualize than three-dimensional ones. Figure A.1b reports an example of a score plot.

Loadings are usually reported as bar plots or as scatter plots. In the first case (which is the way used in this Dissertation) a bar plot of the loadings of the original variables on each PC is reported, as in Figure A.1a. In the second case, as in score plots, the loadings of the variables on a PC are plotted versus the loadings of the same variables on a different PC. This is a useful way to summarize in a single plot more exhaustive information on variable correlation. In general, loading plots are useful for two important reasons: *i*) understanding which are the variables related to the data variability and which are not; *ii*) understanding if there are correlations among the variables. Recalling the meaning of loadings in PCA and weights in PLS (Chapter 2, Section 2.1.1 and Section 2.1.2), a measured variable which shows a high loading or weight has a significant importance on the related PC/LV, thus being responsible of a significant part of the variability in the data. Therefore, loadings in PCA and weights in PLS help in identifying the "most important" variables for the system under study, and to rank them by importance order. If this information is combined with physical knowledge on the system, one can obtain additional physical insights on the system under investigation, by understanding which are the driving forces linked to physical phenomena that drive the system. When two variables have similar loadings on a PC, they are said to be correlated. If the loading absolute values are similar but the values are opposite, they are said to be inversely related (or anti-correlated). This means that it is expected that, considering the data used to build the model, an increase in one variable results in a decrease of the other variable. Figure A.1a gives an example of this occurrence. For example, in the top bar plot it can be clearly seen that variable x_1 and variable x_3 are the most significant variables on this direction, and they are inversely related as their loadings are opposite. Differently, on the second bar plot (which refers to PC2), x_3 has the highest loading and looks inversely related to x_1 and x_2, which have a lower importance. Note that the PCA loadings and the PLS weights on each PC/LV are independent. Therefore, the information obtained from the analysis of one latent component is not contrasting with the other ones, but it simply provides a different type of information (namely, it identifies a different driving force for the process).



(a) (b)

Figure A.1. *Example of (a) loading bar plots and (b) score plot for a model with 2 PCs.*

Score plots as the one reported in Figure A.1b are useful to identify similarities between samples. This means that samples with similar characteristics fall in the same region of the score plot. Moreover, the pattern observed in a score plot reflects the correlation structure identified by the variable loadings. For example, in Figure A.1b three main clusters can be observed along PC1. Samples are therefore grouped according to their similarities or differences in the variables that have the highest loading on PC1. By analyzing the loading plot, one can identify which these variables are (x_1 and x_3 in this case). Therefore, samples having a high positive score on PC1 will have higher x_3 values and lower x_1 values on average, because x_3 has a positive loading on PC1 whereas x_1 has a negative one. The situation is opposite in the case of samples with negative PC1 scores. A similar analysis can be done also for the other PCs.

Finally, note that in the PLS case it is more useful to analyze jointly the model weights with the loadings of $\mathbf{Y}$ ($\mathbf{Q}$ loadings). This analysis allows to identify cross-correlations among variables (i.e., how the regressors are related with the responses), which is of particular interest considering that PLS is a regression model built to predict the responses from the inputs.

A.1.1 Interpretation of the loading plots for the wet granulation case study

This Section reports the interpretation of the $\mathbf{Q}$ loadings plots reported in Figure 4.7 (Chapter 4) for the PLS model built to relate the raw materials characteristics with the granule properties (Section 4.3.1, high-shear wet granulation case study,).

LV1, which accounts for ~32 % of the total variance in $\mathbf{Y}$ ($R^2\mathbf{Y}$, Table 4.4), is mainly driven by the particle size distribution (PSD) variables, namely D[3,2] and D90/D10 (top plot of Figure 4.7). In particular, D[3,2] and D90/D10 are opposite, meaning that granulated materials in the database with high PSD (high D[3,2]) have usually narrower PSD (low D90/D10), compared to the mean of the data. This affects the percentage of oversize granules, which is directly correlated with D[3,2] and the difference in compactability compared to the raw material (ΔCompactability). Namely, it is expected that products with high PSD mean, have larger percentages of oversize granules compared to materials with lower PSD mean, and this seems to slightly affect the compactability difference of the granulated product with the raw material, which is expected to be lower.

LV2, which explains ~23 % of the total variance of the data, is affected mainly by the moisture content loss upon drying (LOD) of the materials after the granulation (second plot of Figure 4.7). From the analysis of the variable loadings it can be argued that materials having a higher loss of water upon drying are in general those whose granules are less grown and result in products with lower compactability properties compared to the raw materials (ΔCompactability).

LV3 (third plot of Figure 4.7) explains ~8 % of the total variance of the data and is mainly driven by the flow properties of the products (ΔFlodex). In particular it appears that products which are more flowable than the corresponding raw materials are also those which have had higher growth ratio and LOD. This is somehow expected since one of the objectives of granulation is that of increasing the flow properties of the processed materials, by enlarging their size. Moreover, moisture (inversely related to LOD) can act as a binding, making the granules more cohesive and thus less flowable (Emery *et al.*, 2009).

LV4 seems to be less significant than the other ones in explaining the systematic variability of **Y**, as reported in Table 4.4 ($R^2\mathbf{Y} = 1.58\%$). This can be also noticed from the low value of the loadings in the last plot of Figure 4.7.

Appendix B

Algorithmic notes

In this Appendix some notes are provided on the main algorithms implemented to estimate the parameters of the latent variable models used in this Dissertation and described in Chapter 2.

B.1 Principal component analysis (PCA)

As shown in Chapter 2 (Section 2.1.1), given a dataset $\mathbf{X}$ $[I \times N]$ of I samples and N variables, the parameters of a PCA model can be found through the eigenvector decomposition of matrix $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$. In this Dissertation, this has mostly been done using singular value decomposition (SVD; Meyer, 2000) or the nonlinear iterative partial least squares algorithm (NIPALS; Wold, 1966).
The first method requires the estimation of $\mathbf{C}$ and then would compute all the PCs of the system (as many as the variables in $\mathbf{X}$) at once.

$$\mathbf{C} = \mathbf{USV}^{\mathrm{T}} \quad . \tag{B.1}$$

In Eq.(B.1), $\mathbf{V} = \mathbf{U}$ and they include the eigenvectors of $\mathbf{C}$, namely corresponding to the PCA loading matrix $\mathbf{P}$. $\mathbf{S}$ is the $[N \times N]$ diagonal matrix of the singular values, which coincides with the eigenvalues of $\mathbf{C}$. The calculation of the $\mathbf{C}$ matrix requires however that there are no missing data in the $\mathbf{X}$ dataset. Given that real datasets are usually characterized by the presence of missing data, the NIPALS algorithm is usually preferred.
The algorithm computes the scores and loadings of each PC in an iterative way, starting from PC1 and extracting each PC one at a time. As with SVD, PCs are found and ordered according to the amount of variance of the original dataset they capture. Starting from PC1, for each PC the algorithm calculates the scores and loadings vectors $\mathbf{t}$ and $\mathbf{p}$ from the $\mathbf{X}$ matrix. The outer product $\mathbf{tp}$ is then subtracted from $\mathbf{X}$ to give the residual matrix $\mathbf{E}$:

$$\mathbf{E} = \mathbf{X} - \mathbf{tp}^{\mathrm{T}} \tag{B.2}$$

$\mathbf{E}$ is then used at the next iteration to extract the scores and loadings vectors for PC2. The algorithm can be summarized in the following steps (Geladi and Kowalski, 1986):

1. consider a row vector $\mathbf{x}_i$ from $\mathbf{X}$ and set $\mathbf{t} = \mathbf{x}_i$.

2. Calculate $\mathbf{p}^{\mathrm{T}}$:

$$\mathbf{p}^{\mathrm{T}} = \frac{\mathbf{t}^{\mathrm{T}}\mathbf{X}}{\mathbf{t}^{\mathrm{T}}\mathbf{t}} \quad ; \tag{B.3}$$

3. Normalize $\mathbf{p}^{\mathrm{T}}$ to unit length.
4. Calculate $\mathbf{t}$:

$$\mathbf{t} = \frac{\mathbf{X}\mathbf{p}}{\mathbf{p}^{\mathrm{T}}\mathbf{p}} \quad ; \qquad \text{(B.4}$$

5. compare $\mathbf{t}$ used in step 2. with $\mathbf{t}$ calculated in step 4. If they are the same (their difference is less than an assigned tolerance), stop (the method has converged), else restart from step 2., with the last calculated $\mathbf{t}$.
6. If converged, calculate $\mathbf{E}$ according to Eq.(B.2), and go back to step 1, by setting $\mathbf{X} = \mathbf{E}$ to calculate the next PC.

The algorithm iterates until the *A* PCs selected to build the PCA model have been determined. It is demonstrated that the parameters provided by the NIPALS algorithm are the same as the eigenvector solution problem of Eq.(2.4) (Chapter 2, Section 2.1.1) (Geladi and Kowalski, 1986). Furthermore, it can feasibly handle datasets with missing data.

Other approaches have been used to calculate the PCA loadings and scores based on optimization frameworks for parameter estimation. In these cases, the PCA loadings are found in order to minimize the sum of squared errors between the data matrix $\mathbf{X}$ and the reconstructed matrix using the PCA model, which for one PC can be set as in Eq.(B.5):

$$\begin{gathered} \min \frac{1}{2}\left\| \mathbf{X} - \mathbf{t}\mathbf{p}^{\mathrm{T}} \right\|_F^2 \\ \text{subject to} \quad \mathbf{p}^{\mathrm{T}}\mathbf{p} = 1 \end{gathered} \quad , \tag{B.5}$$

being $\|\cdot\|_F$ the Frobenius norm. To determine the PCs after the first through Eq.(B.5), the optimization problem should be applied in an iterative way, subject to the additional constraint that the loadings on different PCs are orthonormal. Otherwise, the minimization problem can be formulated in order to estimate simultaneously the *A* PCs selected to build the model:

$$
\begin{aligned}
&\min \frac{1}{2}\sum_{i=1}^{I}\sum_{n=1}^{N}\left(x_{i,n} - \sum_{a=1}^{A} t_{i,a}p_{n,a} \right)^2 \\
&s.t \;\; \sum_{n=1}^{N} p_{n,a}p_{n,a'} = \delta_{a,a'} \quad a,a' = 1,...,A \\
&\qquad \sum_{i=1}^{I} t_{i,l}t_{i,l'} = 0 \quad l' < l \quad l,l' = 1,...,A \\
&\qquad \sum_{i=1}^{I} t_{i,l} = 0 \quad l = 1,...,A
\end{aligned}
\tag{B.6}
$$

In the optimization problem of Eq.(B.6) $x_{i,n}$, $t_{i,a}$ and $p_{n,a}$ are the elements of respectively matrix $\mathbf{X}$, $\mathbf{T}$ and $\mathbf{P}$, while $\delta_{a,a'}$ is the Kronecker delta. The constraint set in Eq.(B.6) forces the loadings to be orthonormal, and the scores to be orthogonal and to have zero mean. The optimization can be solved through appropriate nonlinear programming problem algorithms, and allows to handle datasets with high percentages of missing data (López-Negrete de la Fuente *et al.*, 2011).

B.2 Projection to latent structures (PLS)

Projection to latent structures (PLS) includes a class of algorithms that attempts to summarize the variation in a regressor matrix $\mathbf{X}$ that is in some way predictive of a corresponding matrix $\mathbf{Y}$ of response variables (Chapter 2, Section 2.1.2) (MacGregor *et al.*, 1994). One of the most common algorithms to estimate the PLS model parameters is NIPALS (Wold, 1966; Wold *et al.*, 1983), whose steps are summarized below and illustrated in Figure B.1 (MacGregor *et al.*, 1994).

1. Set $\mathbf{u}$ equal to any column of $\mathbf{Y}$.
2. Regress columns of $\mathbf{X}$ on $\mathbf{u}$ to get weights $\mathbf{w}$:

$$\mathbf{w}^{\mathrm{T}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{X}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}} \quad . \tag{B.7}$$

3. Normalize $\mathbf{w}$ to unit length.
4. Calculate the scores $\mathbf{t}$:

$$\mathbf{t} = \frac{\mathbf{X}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}} \quad . \tag{B.8}$$

5. Regress the columns of $\mathbf{Y}$ on $\mathbf{t}$:

$$\mathbf{q}^{\mathrm{T}} = \frac{\mathbf{t}^{\mathrm{T}}\mathbf{Y}}{\mathbf{t}^{\mathrm{T}}\mathbf{t}} \quad . \tag{B.9}$$

6. Calculate the new score vector for $\mathbf{Y}$:

$$\mathbf{u} = \frac{\mathbf{Yq}}{\mathbf{q}^{\mathrm{T}}\mathbf{q}} \quad . \tag{B.10}$$

7. Check convergence of $\mathbf{u}$: if yes go to step 8; if no go to step 2.
8. Calculate $\mathbf{X}$ matrix loadings, by regressing columns of $\mathbf{X}$ on $\mathbf{t}$:

$$\mathbf{p}^{\mathrm{T}} = \frac{\mathbf{t}^{\mathrm{T}}\mathbf{X}}{\mathbf{t}^{\mathrm{T}}\mathbf{t}} \quad . \tag{B.11}$$

9. Calculate residual matrices $\mathbf{E}$ and $\mathbf{F}$:

$$\mathbf{E} = \mathbf{X} - \mathbf{tp}^{\mathrm{T}} \tag{B.12}$$

$$\mathbf{F} = \mathbf{Y} - \mathbf{tq}^{\mathrm{T}} \tag{B.13}$$

10. To calculate the next set of latent vectors, restart from step 1, replacing $\mathbf{X}$ and $\mathbf{Y}$ by $\mathbf{E}$ and $\mathbf{F}$ respectively.

Eq.(B.7) and Eq.(B.9) allow to change place, allowing each dataset latent space to get information about the other. As in PCA, an important property is that the scores $\mathbf{t}$ calculated on each LV are orthogonal to one another. Various interpretation of the PLS algorithm and its properties are discussed by Höskuldsson (1988).

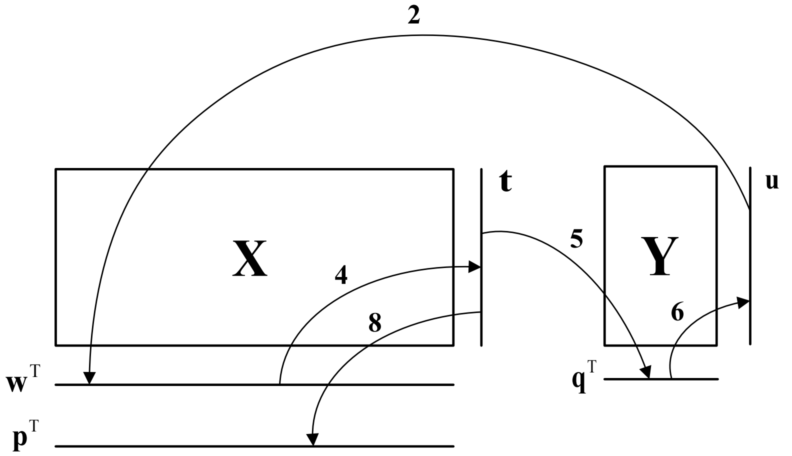


Figure B.1. *Schematic of a PLS algorithm iteration (adapted from MacGregor et al., 1994).*

B.2.1 Multi-block PLS

As shown in Chapter 2 (Section 2.1.3.1), only the case of two regressor blocks $\mathbf{X}_\mathrm{A}$ and $\mathbf{X}_\mathrm{B}$ and a single response variables block $\mathbf{Y}$ is considered. The iteration sequence of the algorithm is reported below and represented in Figure B.2 (MacGregor *et al.*, 1994):

1. Set $\mathbf{u}$ equal to any column of $\mathbf{Y}$.
2. Perform part of a PLS round on each of the blocks $\mathbf{X}_\mathrm{A}$ and $\mathbf{X}_\mathrm{B}$ to get $\mathbf{w}_\mathrm{A}$, $\mathbf{t}_\mathrm{A}$ and $\mathbf{w}_\mathrm{B}$, $\mathbf{t}_\mathrm{B}$ as in steps 2 to 4 of the NIPALS algorithm described above for the PLS case (Eqs.(B.7)-(B.8)).
3. Collect all the score vectors $\mathbf{t}_\mathrm{A}$ and $\mathbf{t}_\mathrm{B}$ in the consensus (superblock) matrix $\mathbf{T}_\mathrm{MB}$.
4. Perform one round of PLS with $\mathbf{T}_\mathrm{MB}$ as $\mathbf{X}$ (steps 2 to 6 in the above-described NIPALS algorithm for PLS) to get a super-weights vector $\mathbf{w}_\mathrm{S}$ and a super-scores vector $\mathbf{t}_\mathrm{S}$, as well as a loading vector $\mathbf{q}$ and a new score vector $\mathbf{u}$ for the $\mathbf{Y}$ matrix.
5. Return to step 2 and iterate until convergence of $\mathbf{u}$.
6. Compute the loadings for each block:

$$\mathbf{p}_\mathrm{A}^\mathrm{T} = \frac{\mathbf{t}_\mathrm{S}^\mathrm{T}\mathbf{X}_\mathrm{A}}{\mathbf{t}_\mathrm{S}^\mathrm{T}\mathbf{t}_\mathrm{S}} \quad . \tag{B.14}$$

$$\mathbf{p}_\mathrm{B}^\mathrm{T} = \frac{\mathbf{t}_\mathrm{S}^\mathrm{T}\mathbf{X}_\mathrm{B}}{\mathbf{t}_\mathrm{S}^\mathrm{T}\mathbf{t}_\mathrm{S}} \quad . \tag{B.15}$$

7. Compute the residual matrices for each block:

$$\mathbf{E}_\mathrm{A} = \mathbf{X}_\mathrm{A} - \mathbf{t}_\mathrm{S}\mathbf{p}_\mathrm{A}^\mathrm{T} \tag{B.16}$$

$$\mathbf{E}_\mathrm{B} = \mathbf{X}_\mathrm{B} - \mathbf{t}_\mathrm{S}\mathbf{p}_\mathrm{B}^\mathrm{T} \tag{B.17}$$

$$\mathbf{F} = \mathbf{Y} - \mathbf{t}_\mathrm{S}\mathbf{q}^\mathrm{T} \tag{B.18}$$

8. Calculate the next set of latent vectors by replacing $\mathbf{X}_\mathrm{A}$, $\mathbf{X}_\mathrm{B}$ and $\mathbf{Y}$ by their residual matrices $\mathbf{E}_\mathrm{A}$, $\mathbf{E}_\mathrm{B}$ and $\mathbf{F}$ and repeating from step 1.

The algorithm implemented in this way allows to obtain orthogonal super scores $\mathbf{t}_\mathrm{S}$ on each LV, while the block scores $\mathbf{t}_\mathrm{A}$ (or $\mathbf{t}_\mathrm{B}$) on different LVs are slightly correlated. Alternatively, one could formulate the algorithm to yield orthogonal block scores but nonorthogonal super scores. This can be achieved by performing the deflation step in the above-mentioned algorithm (steps 6-7) with the block scores $\mathbf{t}_\mathrm{A}$ and $\mathbf{t}_\mathrm{B}$ instead of the super scores $\mathbf{t}_\mathrm{S}$. However, it has been showed that by using the block score deflation method, some of the information in the $\mathbf{X}$ datasets may be lost in the deflation step, possibly leading to poor performances for the model (Westerhuis *et al.*, 1998). Furthremore, as already stated in Chapter 2 (Section 2.1.3.1), the multi-block PLS algorithm parameters can be obtained from the standard PLS method (Figure B.1), if the appropriate data pretreatment is applied to the block matrices.

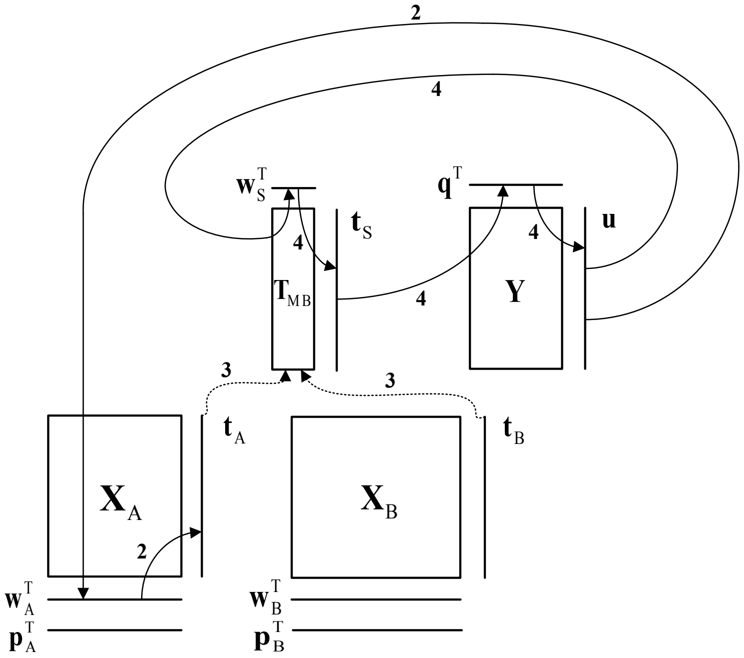


Figure B.2. *Schematic of a multi-block PLS algorithm iteration (adapted from MacGregor et al., 1994).*

B.2.2 Joint-Y PLS

As in the PLS and in the multi-block PLS cases, a modified version of the NIPALS algorithm has also been proposed for Joint-Y PLS (JY-PLS). The algorithm steps for an iteration are summarized below and represented in Figure B.3 in the case two regressor (i.e. sites) datasets $\mathbf{X}_A$ and $\mathbf{X}_B$ and two response variables datasets $\mathbf{Y}_A$ and $\mathbf{Y}_B$, forming the joint-**Y** matrix $\mathbf{Y}_J$, are considered (García-Muñoz, 2004):

1. Initialize $\mathbf{u}_A$ and $\mathbf{u}_B$ with the first column of $\mathbf{Y}_A$ and $\mathbf{Y}_B$.
2. Regress $\mathbf{X}_A$ and $\mathbf{X}_B$ onto $\mathbf{u}_A$ and $\mathbf{u}_B$ to compute $\mathbf{w}_A$ and $\mathbf{w}_B$:

$$\mathbf{w}_A^T = \frac{\mathbf{u}_A^T \mathbf{X}_A}{\mathbf{u}_A^T \mathbf{u}_A} \tag{B.19}$$

$$\mathbf{w}_B^T = \frac{\mathbf{u}_B^T \mathbf{X}_B}{\mathbf{u}_B^T \mathbf{u}_B} \quad . \tag{B.20}$$

3. Normalize $\mathbf{w}_A$ and $\mathbf{w}_B$ as:

$$\left\| \begin{matrix} \mathbf{w}_A \\ \mathbf{w}_B \end{matrix} \right\| = 1 \quad . \tag{B.21}$$

4. Regress $\mathbf{X}_\mathrm{A}$ and $\mathbf{X}_\mathrm{B}$ onto $\mathbf{w}_\mathrm{A}$ and $\mathbf{w}_\mathrm{B}$ to obtain $\mathbf{t}_\mathrm{A}$ and $\mathbf{t}_\mathrm{B}$:

$$\mathbf{t}_\mathrm{A} = \frac{\mathbf{X}_\mathrm{A}\mathbf{w}_\mathrm{A}}{\mathbf{w}_\mathrm{A}^\mathrm{T}\mathbf{w}_\mathrm{A}} \tag{B.22}$$

$$\mathbf{t}_\mathrm{B} = \frac{\mathbf{X}_\mathrm{B}\mathbf{w}_\mathrm{B}}{\mathbf{w}_\mathrm{B}^\mathrm{T}\mathbf{w}_\mathrm{B}} \quad . \tag{B.23}$$

5. Regress the joint-**Y** matrix onto the joint scores to obtain the joint loadings ($\mathbf{q}_\mathrm{J}$):

$$\mathbf{q}_\mathrm{J}^\mathrm{T} = \begin{bmatrix} \mathbf{t}_\mathrm{A} \\ \mathbf{t}_\mathrm{B} \end{bmatrix}^\mathrm{T} \begin{bmatrix} \mathbf{Y}_\mathrm{A} \\ \mathbf{Y}_\mathrm{B} \end{bmatrix} \left(\begin{bmatrix} \mathbf{t}_\mathrm{A} \\ \mathbf{t}_\mathrm{B} \end{bmatrix}^\mathrm{T} \begin{bmatrix} \mathbf{t}_\mathrm{A} \\ \mathbf{t}_\mathrm{B} \end{bmatrix} \right)^{-1} \quad . \tag{B.24}$$

6. Regress $\mathbf{Y}_\mathrm{A}$ and $\mathbf{Y}_\mathrm{B}$ onto $\mathbf{q}_\mathrm{J}$ to re-compute $\mathbf{u}_\mathrm{A}$ and $\mathbf{u}_\mathrm{B}$:

$$\begin{bmatrix} \mathbf{u}_\mathrm{A} \\ \mathbf{u}_\mathrm{B} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_\mathrm{A} \\ \mathbf{Y}_\mathrm{B} \end{bmatrix} \mathbf{q}_\mathrm{J} \left(\mathbf{q}_\mathrm{J}^\mathrm{T}\mathbf{q}_\mathrm{J} \right)^{-1} \quad . \tag{B.25}$$

7. Check convergence with respect to the orginal values of $\mathbf{u}_\mathrm{A}$ and $\mathbf{u}_\mathrm{B}$. If convergence fails, go back to step 2 with the new values of $\mathbf{u}_\mathrm{A}$ and $\mathbf{u}_\mathrm{B}$, otherwise go to step 8.
8. Calculate $\mathbf{p}_\mathrm{A}$ and $\mathbf{p}_\mathrm{B}$ for deflation:

$$\mathbf{p}_\mathrm{A}^\mathrm{T} = \frac{\mathbf{t}_\mathrm{A}^\mathrm{T}\mathbf{X}_\mathrm{A}}{\mathbf{t}_\mathrm{A}^\mathrm{T}\mathbf{t}_\mathrm{A}} \tag{B.26}$$

$$\mathbf{p}_\mathrm{B}^\mathrm{T} = \frac{\mathbf{t}_\mathrm{B}^\mathrm{T}\mathbf{X}_\mathrm{B}}{\mathbf{t}_\mathrm{B}^\mathrm{T}\mathbf{t}_\mathrm{B}} \quad . \tag{B.27}$$

9. Deflate $\mathbf{X}_\mathrm{A}$ and $\mathbf{X}_\mathrm{B}$, $\mathbf{Y}_\mathrm{A}$ and $\mathbf{Y}_\mathrm{B}$ and estimate the next component, going back to step 1.

$$\mathbf{X}_\mathrm{A} = \mathbf{X}_\mathrm{A} - \mathbf{t}_\mathrm{A}\mathbf{p}_\mathrm{A}^\mathrm{T} \tag{B.28}$$

$$\mathbf{X}_\mathrm{B} = \mathbf{X}_\mathrm{B} - \mathbf{t}_\mathrm{B}\mathbf{p}_\mathrm{B}^\mathrm{T} \tag{B.29}$$

$$\mathbf{Y}_\mathrm{A} = \mathbf{Y}_\mathrm{A} - \mathbf{t}_\mathrm{A}\mathbf{q}_\mathrm{J}^\mathrm{T} \tag{B.30}$$

$$\mathbf{Y}_\mathrm{B} = \mathbf{Y}_\mathrm{B} - \mathbf{t}_\mathrm{B}\mathbf{q}_\mathrm{J}^\mathrm{T} \quad . \tag{B.31}$$

The algorithm above is identical to the one previously described for the PLS model parameter estimation, with the exception of step 5, in which the joint loadings $\mathbf{q}_\mathrm{J}$ are computed with the joint score vector. This step allows to achieve the main goal of the methodology, which is to find the common plant defined by both $\mathbf{Y}_\mathrm{A}$ and $\mathbf{Y}_\mathrm{B}$, since both matrices are projected simultaneously onto the score space (García-Muñoz, 2004).

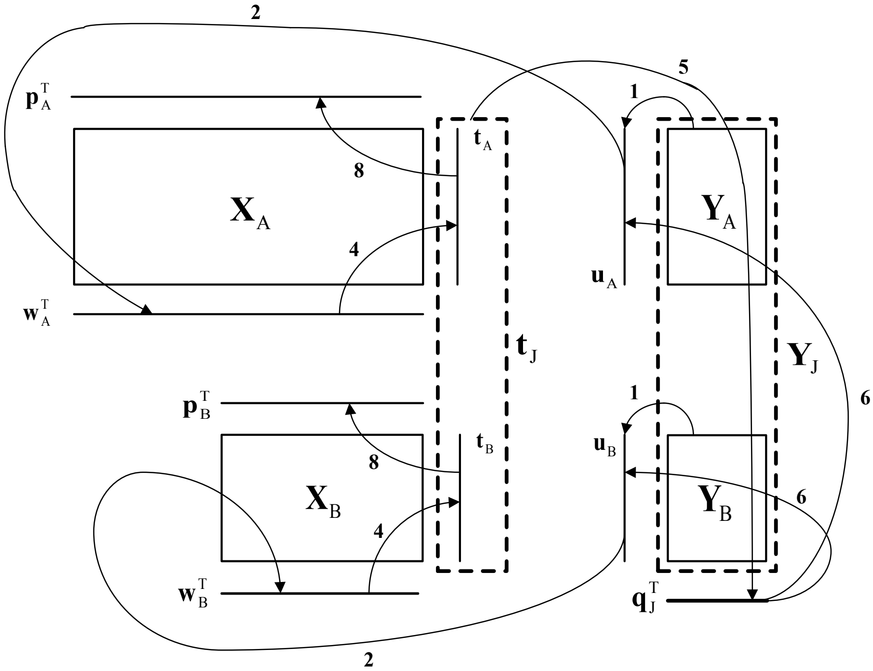


Figure B.3. *Schematic of a Joint-Y PLS algorithm iteration (adapted from García-Muñoz, 2004).*

Appendix C

Model inversion improvements

This Appendix reports some notes on the latent variable regression model (LVRM) inversion framework presented in Chapter 4. In particular, a procedure is described to estimate uncertainties in the calculation of the null space and of the model inversion solution (Chapter4, Section 4.3.2). In the second part, a discussion is provided on the model to be used in the reconstruction of the desired product attribute profile for model inversion (Chapter4, Section 4.4.1).

C.1 Calculation of the confidence limits for the null space and the optimization solution

To calculate the confidence limits to account for the uncertainties in the estimation of the null space, a jackknife approach was applied (Duschesne and MacGregor, 2001). Let us define $\mathbf{X}$ and $\mathbf{Y}$ as the regressor and the response variable datasets, $\mathbf{y}^{\mathrm{DES}}$ as the desired product property set in which each property is fixed by the user, and $\mathbf{t}$ as the vector of the projections of the model inversion solution $\hat{\mathbf{x}}^{\mathrm{NEW}}$. At the generic iteration *i*, the procedure goes through the following steps:

1. Remove the *i*-th row from both $\mathbf{X}$ and $\mathbf{Y}$, generating the new matrices $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$.
2. Estimate the PLS model between $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$, keeping the same number *A* of LVs as in the PLS model built between $\mathbf{X}$ and $\mathbf{Y}$.
3. Estimate the null space points following the procedure described by Jaeckle and MacGregor (2000a).
4. Invert the PLS model to estimate the projections $\mathbf{t}^{(i)}$ corresponding to the new input conditions $\hat{\mathbf{x}}^{\mathrm{NEW}(i)}$.

Repeating this procedure for all the *I* samples contained in the original datasets gives in output a three-dimensional array $\underline{\mathbf{T}}^{\mathrm{NULL}}$ $[L \times I \times A]$, being *L* the number of the null space considered points, and a matrix $\mathbf{T}^{\mathrm{NEW}}$ $[I \times A]$ of the solutions from the *I* different model inversions. For the *l*-th point $\mathbf{t}_l^{\mathrm{NULL}}$ $[A \times 1]$ of the null space the 95% confidence ellipsoid is formed by the points with coordinates $\boldsymbol{\tau}^{\mathrm{NULL}}$ $[A \times 1]$ which satisfy the ellipsoidal equation (Mardia *et al.*, 1979):

$$\left(\boldsymbol{\tau}^{\mathrm{NULL}}-\mathbf{t}_l^{\mathrm{NULL}}\right)^{\mathrm{T}}\mathbf{C}_l^{-1}\left(\boldsymbol{\tau}^{\mathrm{NULL}}-\mathbf{t}_l^{\mathrm{NULL}}\right)=f \quad , \tag{C.1}$$

where $\mathbf{C}_l$ is the $[A \times A]$ covariance matrix calculated from matrix $\mathbf{T}_l^{\mathrm{NULL}}$ $[I \times A]$ of the I different jackknife estimations of the l-th null space point, while c can be approximated as:

$$f=\frac{A\cdot\left(I^2-1\right)}{I\cdot\left(I-A\right)}\cdot F_{A,I-A,\alpha} \quad ; \tag{C.2}$$

$F_{A,I-A,\alpha}$ is the critical value of the F distribution with A and $I-A$ degrees of freedom at significance level α ($\alpha=0.05$ for the 95% confidence limits).

To calculate the projections of the confidence ellipsoid in the bi-dimensional score diagrams it is considered that $A=2$. The confidence ellipse for the l-th null space point in the null space can then be found solving the second order equation derived from (C.1):

$$\left(\tau_1^{\mathrm{NULL}}-t_{1,l}^{\mathrm{NULL}}\right)^2\cdot c_{11}+\left(\tau_1^{\mathrm{NULL}}-t_{1,l}^{\mathrm{NULL}}\right)\cdot\left(\tau_2^{\mathrm{NULL}}-t_{2,l}^{\mathrm{NULL}}\right)\cdot\left(c_{12}+c_{21}\right)+\left(\tau_2^{\mathrm{NULL}}-t_{2,l}^{\mathrm{NULL}}\right)^2\cdot c_{22}=f \tag{C.3}$$

being τ_1^{NULL} and τ_2^{NULL} the elements of $\boldsymbol{\tau}^{\mathrm{NULL}}$ (namely the coordinates of the ellipse points), $t_{1,l}^{\mathrm{NULL}}$ and $t_{2,l}^{\mathrm{NULL}}$ the elements of $\mathbf{t}_l^{\mathrm{NULL}}$, while c_{11}, c_{12}, c_{21} and c_{22} the elements of the covariance matrix $\mathbf{C}_l$.

The null space confidence limits represented by the red lines in Figure 4.3 (Chapter 4, Section 4.3.2) are then formed by joining the points of the L 95% confidence ellipses at maximum distance from each null space point $\mathbf{t}_l^{\mathrm{NULL}}$.

The described procedure is applied also to estimate the 95% confidence ellipse for the optimization solution $\hat{\mathbf{t}}$ (blue dot ellipses in Figure 4.3 diagrams). In this case, the equation in (C.1) is modified accordingly:

$$\left(\boldsymbol{\tau}-\hat{\mathbf{t}}\right)^{\mathrm{T}}\mathbf{C}^{-1}\left(\boldsymbol{\tau}-\hat{\mathbf{t}}\right)=f \quad , \tag{C.4}$$

where $\mathbf{t}_l^{\mathrm{NULL}}$ is substituted with the optimal solution vector $\hat{\mathbf{t}}$, while the covariance matrix $\mathbf{C}$ is estimated from the jackknife replications of the inversion solution included in $\mathbf{T}^{\mathrm{NEW}}$.

C.2 On the reconstruction of $\mathbf{y}^{\mathrm{DES}}$ through the model on Y

The LVRM inversion relies on the assumption that $\mathbf{y}^{\mathrm{DES}}$ adheres to the covariance structure of the response matrix $\mathbf{Y}$. In the case equality constraints are specified for some of the elements in $\mathbf{y}^{\mathrm{DES}}$, this assumption has to be checked, for example by considering the PCA model on $\mathbf{Y}$, and comparing $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ calculated through it with the SPE of the historical samples in $\mathbf{Y}$ (García-Muñoz *et al.*, 2006). Indeed, the covariance structure described by the

PCA model on $\mathbf{Y}$ is not the same as the one described by the $\mathbf{Q}$ loadings of the PLS model between $\mathbf{X}$ and $\mathbf{Y}$ (Chapter 2, Section 2.1.2). This means that if $\mathbf{y}^{\mathrm{DES}}$ belongs to the space of the PCA model on $\mathbf{Y}$ (i.e. $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}^{\mathrm{PCA}} = 0$ according to the PCA model), it is not ensured that $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}^{\mathrm{PLS}} = 0$ according to the PLS model.

If $\mathbf{y}^{\mathrm{DES}}$ is coplanar to the subspace mapped by the PCA model on $\mathbf{Y}$ (or it is reconstructed through it), then the new product quality $\hat{\mathbf{y}}^{\mathrm{NEW}}$ corresponding to the model inversion solution $\mathbf{x}^{\mathrm{NEW}}$ will differ from $\mathbf{y}^{\mathrm{DES}}$ proportionally to the differences between $\mathbf{P}$ (loadings from PCA on $\mathbf{Y}$) and $\mathbf{Q}$. These differences are fundamentally driven by the correlation between the latent spaces of $\mathbf{X}$ and $\mathbf{Y}$. In the best case, the greatest and significant, directions of variability in $\mathbf{Y}$ will also be explained by $\mathbf{X}$ resulting in a $\mathbf{Q}$ matrix that is a rotated version of the PCA $\mathbf{P}$ loadings. In such a case, reconstructing $\mathbf{y}^{\mathrm{DES}}$ (or assessing its correlation) using the PCA $\mathbf{P}$ or the $\mathbf{Q}$ loadings should make no difference. One way the practitioner can determine this similarity (or lack of thereof) is by comparing the total residual sum of squares from the PCA model built on $\mathbf{Y}$ with the residuals for the $\mathbf{Y}$ space in the PLS model and performing a canonical correlation analysis (Mardia *et al.*, 1979) between $\mathbf{P}$ and $\mathbf{Q}$.

If $\mathbf{y}^{\mathrm{DES}}$ belongs to the sub-space of the PLS model described by $\mathbf{Q}$ (or it is reconstructed through it), than $\mathbf{y}^{\mathrm{DES}} = \mathbf{Qt}$ can be set as an hard constraint in the model inversion problems, and the soft constraints on $\mathbf{y}^{\mathrm{DES}}$ in the optimization problems disappear. This is the strategy followed in the approaches presented in Section 4.4 of Chapter 4.

Considering the differences between the two modeling techniques, one could argue that it is best to choose (or reconstruct) $\mathbf{y}^{\mathrm{DES}}$ using the PCA $\mathbf{P}$ loadings to ensure that the complete correlation structure from the historical data is accounted for; one could also argue that given the ultimate objective (estimate $\mathbf{x}^{\mathrm{NEW}}$) it only makes sense to consider the covariance in $\mathbf{Y}$ that is correlated with $\mathbf{X}$.

Regardless of the methodology, the reconstruction of $\mathbf{y}^{\mathrm{DES}}$ allows to discard the (possible) uncertainties in the quality variables, which are not easily identifiable and are not handled in the presented model inversion framework.

Appendix D

Monitoring model transfer adaptive mechanisms

This Appendix reports the details on the implementation of the adaptive monitoring transfer scenarios of the general framework described in Chapter 7 (Figure 7.1). In the first part, the description of the monitoring chart design for the adaptive JY-PLS model (transfer Scenario 3 in Figure 7.1) is provided. In the second part, the modelling strategy and the adaptation mechanism implemented for transfer Scenario 4 and transfer Scenario 5 of the proposed framework are described.

D.1 Monitoring chart design and interrogation procedures for the adaptive JY-PLS model

To build the adaptive JY-PLS model (Chapter 7, Section 7.3.3.1) control charts, the confidence limits of T^2 and SPE are calculated from the data for plant B:

$$\mathbf{T}_k^* = \mathbf{T}_k^{\mathrm{B}} \tag{D.1}$$

$$\mathbf{E}_k^* = \mathbf{E}_{\mathrm{B},k} \quad , \tag{D.2}$$

where $\mathbf{T}_k^{\mathrm{B}}$ is the $[(k-1)\times A]$ matrix of the scores of the $(k-1)$ samples available from plant B, while $\mathbf{E}_k^{\mathrm{B}}$ is the matrix of the reconstruction errors of $\mathbf{Y}_k''^{\mathrm{B}}$ (Figure 7.3, Section 7.3.3.1). The limits of the Hotelling's T^2 and of the SPE statistics are calculated at each updating instant k using the available $(k-1)$ samples from plant B, according to the same equations described in Chapter 2 (Section 2.1.4)[Ä].

The monitoring of plant B operating conditions is carried out by interrogating the JY-PLS model with the data incoming from this plant. The incoming data are projected onto the joint monitoring charts utilizing the procedure described in the following.

[Ä] As an alternative, the control limits for T^2 and *SPE* can be defined using both plant A data and plant B data, with $\mathbf{T}_k^* = \begin{bmatrix} \mathbf{T}_{\mathrm{A},k} \\ \mathbf{T}_{\mathrm{B},k} \end{bmatrix}$ and $\mathbf{E}_k^* = \mathbf{E}_k^{\mathrm{J}}$.

The Hotelling's T_k^2 statistic for the incoming plant B data at updating instant k can be calculated as:

$$T_k^2 = \hat{\mathbf{t}}_k^{\mathrm{B}^\mathrm{T}} \boldsymbol{\Lambda}^{-1} \hat{\mathbf{t}}_k^{\mathrm{B}} \quad , \tag{D.3}$$

where:

$$\boldsymbol{\Lambda} = \mathrm{diag}([\lambda_1, \lambda_2, ..., \lambda_A]) \quad , \tag{D.4}$$

and $\lambda_1, \lambda_2, ..., \lambda_A$ are the elements of:

$$\boldsymbol{\lambda} = \frac{\mathrm{diag}\left(\mathbf{T}^{*^\mathrm{T}} \mathbf{T}^*\right)}{n-1} \quad . \tag{D.5}$$

Similarly, the SPE statistic for the incoming data is determined by:

$$\mathrm{SPE}_k = \sum_{v=1}^{V''} e_{k,v}^{\mathrm{B}\ 2} \quad , \tag{D.6}$$

where is:

$$e_{k,v}^{\mathrm{B}} = y_{k,v}''^{\mathrm{B}} - \hat{y}_{k,v}''^{\mathrm{B}} \quad . \tag{D.7}$$

It is worth remarking that the values of both T_k^2 and SPE$_k$ depend not only on the incoming plant B data, but also on the whole set of available plant A data. In fact, T_k^2 is determined by the projection of $\mathbf{x}_k''^{\mathrm{B}}$ through $\mathbf{W}_k^{*\mathrm{B}}$ (Eq.(7.3) in Section 7.3.3.1), which are the loadings describing the directions of maximum variability in the space of the regressor variables mostly correlated to the joint space of the common response variables through $\mathbf{Q}_{\mathrm{J},k}$. Furthermore, the value of SPE$_k$ is obtained as the difference between the actual value and the predicted value of the response variables (Eqs.(D.6)-(D.7)), where the predicted value depends on the joint space of common response variables (Eq.(7.4)) and on the correlation within a single plant (Eq.(7.3)).

D.2 Adaptation mechanism for Scenario 4 and Scenario 5 of the monitoring model transfer framework

A flow chart of the monitoring strategy and model adaptation mechanism is shown in Figure D.1.

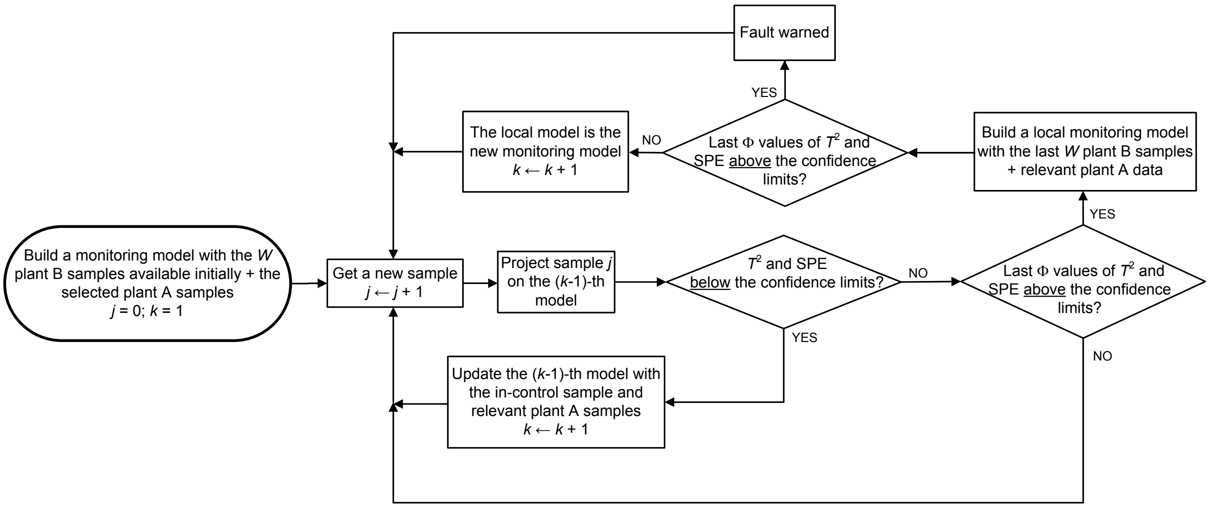

Fault warned
YES
The local model is the new monitoring model $k \leftarrow k + 1$
NO
Last Φ values of T^2 and SPE above the confidence limits?
Build a local monitoring model with the last W plant B samples + relevant plant A data
YES
Build a monitoring model with the W plant B samples available initially + the selected plant A samples $j = 0$; $k = 1$
Get a new sample $j \leftarrow j + 1$
Project sample j on the (k-1)-th model
T^2 and SPE below the confidence limits?
NO
Last Φ values of T^2 and SPE above the confidence limits?
YES
NO
Update the (k-1)-th model with the in-control sample and relevant plant A samples $k \leftarrow k + 1$

Figure D.1. *Simplified flow chart of the online monitoring strategy and model adaptation mechanism for Scenario 4 and Scenario 5 of the proposed framework for the transfer of monitoring models.*

The main steps for any new incoming plant B sample $\mathbf{x}_j^{\mathrm{B}}$ are the following:

1. preprocess[†] $\mathbf{x}_j^{\mathrm{B}}$ on the current values of mean and standard deviation of the plant B samples available so far, and calculate the plant independent variable *wtd*;
2. project $\mathbf{x}_j^{\mathrm{B}}$ onto the space of the monitoring model designed at updating instant $(k-1)$;
3. calculate the T^2 and SPE statistics for $\mathbf{x}_j^{\mathrm{B}}$, and compare them to the confidence limits in the relevant control chart (Chapter 2, Section 2.1.4). If both of them are below the confidence limits, update the monitoring model using $\mathbf{x}_j^{\mathrm{B}}$ and the relevant selected plant A samples, and go to step 1, otherwise go to step 4;
4. if the last Φ samples are found as outliers in the T^2 or in the SPE monitoring charts, go to step 5, otherwise go to step 1;
5. build a local monitoring model for plant B based on the last W samples from plant B plus the plant A samples that can be selected from these plant B samples;
6. if the last Φ samples are found as outliers in the T^2 or SPE monitoring charts of the local model, a fault is detected, an alarm is warned and the procedure goes back to step 1[‡]; otherwise the local monitoring model becomes the new plant B monitoring model and the procedure goes back to step 1.

[†] For the PCA model, the preprocessing operation is autoscaling. For the JY-PLS model, the preprocessing operations are those reported by García-Muñoz *et al.* (2005) (Chapter 2, Section 2.1.3.2).

[‡] To avoid adaptation to the fault, the possibility to build a local model is inhibited until a new NOC sample if found. This is not shown in Figure D.1 for clarity.

Appendix E

Details on the penicillin fermentation process

This Appendix reports the details on the penicillin fermentation process analyzed in Chapter 7 (Section 7.6). First, the description of the two simulated plants and of the inputs assigned to the simulator is provided. Then, details are given on the strategy used to synchronize the variable trajectories of different batches.

E.1 Process simulations

Data were obtained using the PenSim[*] simulator, which solves a detailed mechanistic model of differential-algebraic equations describing the biological behavior of the fermentation process. Overall, one hundred normal batches were simulated both for plant A and for plant B.

The initial values assigned to the variables in order to carry out a simulation are listed in Table E.1. The same table reports the average input values assigned for the simulations of plant A and plant B, and the maximum (common-cause) variability across batches due to input changes. The variability was generated by assigning a random Gaussian noise to the average input values. As can be seen from Table E.1, some initial values were maintained the same in both plants (substrate concentration; concentration of dissolved oxygen; pH; reactor temperature; pH setpoint; reactor temperature setpoint). Other initial values were assigned different average values in the different plants (biomass concentration; culture volume; carbon dioxide concentration; aeration rate; agitation power; substrate federate; substrate inlet temperature). Fluctuations were also introduced to some input variables (aeration rate; agitation power; substrate federate; pH setpoint; reactor temperature setpoint) throughout the process as pseudo-random binary signals (Birol *et al.*, 2002).

A batch is terminated when the penicillin concentration attains an assigned target, that is 1.1 g/L for plant A, and 0.74 g/L for plant B. Therefore, the actual batch length is not set *a priori*, and it can significantly change from batch to batch depending on the actual values of the inputs. Typical time trajectories of the penicillin concentration for some batches in plant A

[*] http://simulator.iit.edu/web/pensim/index.html

and plant B are shown in Figure E.1. Across the entire set of simulated batches, the length of plant A batches ranges between 250 h and 320 h, and the length of plant B batches ranges between 200 h and 315 h. The sampling time for the process measured variables (including the initial penicillin concentration shown below) was set to 0.5 h.

Table E.1. *Initial values of the variables for the simulation of the fed-batch process for the production of penicillin: average value and maximum variability for Plant A and Plant B.*

	Plant A		**Plant B**	
Variables	**Average**	**Maximum variability**	**Average**	**Maximum variability**
Substrate concentration (g/L)	15	±2.5	15	±2.5
Concentration of dissolved oxygen (mmol/L)	1.16	0	1.16	0
Biomass concentration (g/L)	0.05	±0.025	0.15	±0.025
Initial penicillin concentration (g/L)	0	0	0	0
Culture volume (L)	105	±2.5	195	±2.5
Carbon dioxide concentration (mmol/L)	0.65	±0.0625	0.85	±0.0625
pH	5	±0.25	5	±0.25
Reactor temperature (K)	299	±0.25	299	±0.25
Initial generated heat (kcal)	0	0	0	0
Aeration rate (L/h)	4	±0.25	8	±0.25
Agitation power (W)	20	0	40	0
Substrate feed rate (L/h)	0.037	±0.00125	0.042	±0.00125
Substrate inlet temperature (K)	296.5	±0.125	297.5	±0.125
pH setpoint	5	0	5	0
Reactor temperature setpoint (K)	298	0	298	0

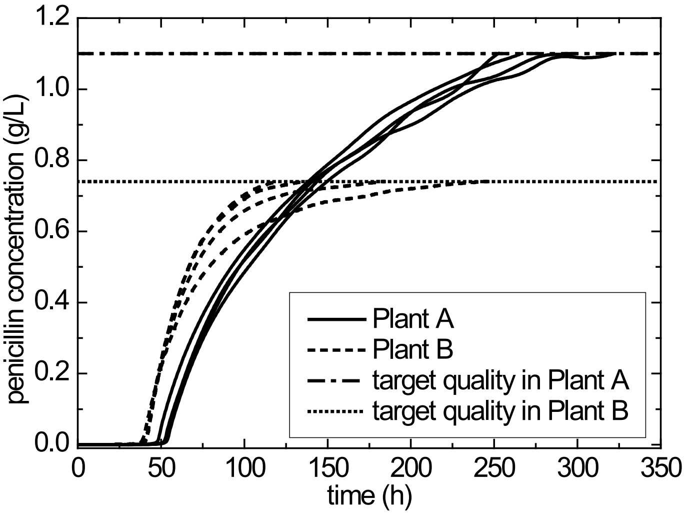


Figure E.1. *Typical time profiles of the fermentation product quality (penicillin concentration) for Plant A and Plant B.*

E.2 Variable trajectories alignment

As mentioned earlier, the batch length changes within a plant due to (normal) variations in the input conditions. Therefore, the measured variable trajectories need to be synchronized before proceeding with a multivariate statistical analysis.

The indicator variable approach (Nomikos and MacGregor, 1995b) was used to synchronize the time trajectories. According to this approach, a monotonically-changing measured variable can be selected as the indicator variable and used to align the batches, being this variable an index of the percentage of batch completion. For the process under investigation, two different indicator variables were selected, one for each operating stage (García-Muñoz *et al.*, 2003): in the first (batch) stage, the substrate concentration was used as the indicator variable, whereas in the second (fed-batch) stage the penicillin concentration was used. These variables are appropriate indicator variables not only because they are monotonic, but also because they have fixed initial and final values. Namely, the substrate concentration during stage one was scanned by evenly partitioning its time trajectory into 25 levels, from the initial value of 12 g/L to the final value of 0.56 g/L (i.e., the end of stage 1); the penicillin concentration was scanned by evenly partitioning its time trajectory into 175 levels, from the initial value of 0.18 g/L to the target value of 1.1 g/L in Plant A, and to the target value of 0.74 g/L in Plant B. As a result, after synchronization the batches are not monitored in the domain of time, but in the domain of the percent of batch completion.

References

Aamir, E., Z.K. Nagy and C.D. Rielly (2010). Optimal seed recipe design for crystal size distribution control for batch cooling crystallization processes. *Chem. Eng. Sci.*, **65**, 3602-3614.

Aboud, L. and H. Scott (2003). New prescription for drug makers: update the plants. *Wall Street Journal*, September 12, 2003.

Abraham, A., C. Grosan and S. Tigan (2007). Ensemble of hybrid neural network learning approaches for designing pharmaceutical drugs. *Neural Comp. Appl.*, **16**, 307-316.

Adam, S., D. Suzzi, C. Radeke and J.G. Khinast (2011). An integrated Quality-by-Design (QbD) approach towards design space definition of a blending unit operation by discrete element method (DEM) simulation. *Eur. J. Pharm. Sci.*, **42**, 106-115.

Agatonovic-Kustrin, S. and R. Beresford (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.*, **22**, 717-722.

Akkisetty, P.K., U. Lee, G.V. Reklaitis and V. Venkatasubramanian (2010). Population balance model-based hybrid neural network for a pharmaceutical milling process. *J. Pharm. Innov.*, **5**, 161-168.

Alexis, F., E. Pridgen, L.K. Molnar and O.C. Farokhzad (2008). Factor affecting the clearance and biodistribution of polymeric nanoparticles. *Mol. Pharm.*, **5**, 505-515.

am Ende, D., K.S. Bronk, J. Mustakis, G. O'Connor, C.L. Santa Maria, R. Nosal and T.J.N. Watson (2007). API Quality by Design example from the Torcetrapib manufacturing process. *J. Pharm. Innov.*, **2**, 71-86.

Andemichael, Y., J. Chen, J.S. Clawson, W. Dai, A. Diederich, S.V. Downing, A.J. Freyer, P. Liu, L.M. Oh, D.B. Patience, S. Sharpe, J. Sisko, J. Tsui, F.G. Vogt, J. Wang, L. Wernersbach, E.C. Webb, J. Wertman ans L. Zhou (2009). Process development for a novel pleuromutilin-derived antibiotic. *Org. Proc. Res. Dev.*, **13**, 729-738.

Andersson, M., A. Ringberg and C. Gustafsson (2007). Multivariate methods in tablet formulation suitable for early drug development: predictive models from a screening design of several linked responses. *Chemom. Intell. Lab. Syst.*, **87**, 125-130.

Arteaga, F. and A. Ferrer (2002). Dealing with missing data in MSPC: several methods, different interpretations, some examples. *J. Chemom.*, **16**, 408-418.

Barmpalexis, P., K. Kachrimanis, A. Tsakonas and E. Georgarakis (2011). Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. *Chemom. Intell. Lab. Syst.*, **107**, 75-82.

Bergman, R., M.E. Johansson, T. Lundstedt, E. Seifert and J. Åberg (1998). Optimization of a granulation and tabletting process by sequential design and multivariate analysis. *Chemom. Intell. Lab. Syst.*, **44**, 271-286.

Bernardo, C.A., A.M. Cunha and M.J. Oliveira (1996). The recycling of thermoplastics: prediction of the properties of virgin and reprocessed polyolefins. *Polym. Eng. Sci.*, **36**, 511-519.

Biegler, L.T. (2010). *Nonlinear programming: concepts, algorithms and applications to chemical processes*. SIAM, Philadelphia, PA (U.S.A.).

Bilgili, E. and B. Scarlett (2005). Population balance modeling of non-linear effects in milling processes. *Powder Tech.*, **153**, 59-71.

Birol, G., C. Ündey, and Çinar (2002). A modular simulation package for fed-batch fermentation: penicillin production. *Computers Chem. Eng.*, **26**, 1553-1565.

Blanco, M., M. Alcala, J.M. Gonzales and E. Torras (2006). Near infrared spectroscopy in the study of polymorphic transformation. *Anal. Chim. Acta*, **567**, 262-268.

Borosy, A.P. (1999). Quantitative composition-property modeling of rubber mixtures by utilizing artificial neural networks. *Chemom. Intell. Lab. Syst.*, **47**, 227-238.

Boukouvala, F., F.J. Muzzio and M. Ierapetritou (2010). Design space of pharmaceutical processes using data-driven-based methods. *J. Pharm. Innov.*, **5**, 119-137.

Boukouvala, F., A. Dubey, A. Vanarase, R. Ramachandran, F.J. Muzzio and M. Ierapetritou (2011). Computational approaches for studying the granular dynamics of continuous blending proecsses, 2 – Population balance and data-based method. *Macromol. Mat. Eng.*, **297**, 9-19.

Boukouvala, F. and M. Ierapetritou (2012). Feasibility analysis of black-box processes using an adaptive sampling Kriging-based method. *Computers Chem. Eng.*, **36**, 358-368.

Boukouvala, F., V. Niotis, R. Ramachandran, F.J. Muzzio and M.G. Ierapetritou (2012). An integrated approach for dynamic flowsheet modeling and sensitivity analysis of a continuous tablet manufacturing process. *Computers Chem. Eng.*, **42**, 30-47.

Burggraeve, A., T. Van Den Kerkhof, M. Hellings, J.P. Remon, C. Vervaet and T. De Beer (2011). Batch statistical process control of a fluid bed granulation process using in-line spatial filter velocimetry and product temperature measurements. *Europ. J. Pharm. Sci.*, **42**, 584-592.

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

Burnham, A.J., J.F. MacGregor and R. Viveros (1999a). Latent variable multivariate regression modeling. *Chemom. Intel. Lab. Sys.*, **48**, 167-180.

Burnham, A.J., J.F. MacGregor and R. Viveros (1999b). A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *J. Chemom.*, **13**, 49-65.

Burt, J.L., A.D. Braem, A. Ramirez, B. Mudryk, L. Rossano and S. Tummala (2011). Model-guided design space development for a drug substance manufacturing process. *J. Pharm. Innov.*, **6**, 181-192.

Campisi, B., D. Chicco, D. Vojnovic and R. Phan-Tan-Luu (1998). Experimental design for a pharmaceutical formulation: optimisation and robustness. *J. Pharm. Biomed. Anal.*, **18**, 57-65.

Chalus, P., Y. Roggo, S. Walter and M. Ulmschneider (2005). Near-infrared determination of active substance content in intact low-dosage tablets. *Talanta*, **66**, 1294-1302.

Chen, Z., D. Lovett and J. Morris (2011). Process analytical technologies and real time process control a review of some spectroscopic issues and challenges. *J. Process Control*, **21**, 1467-1482.

Cheng, C. and M.S. Chiu (2005). Nonlinear process monitoring using JITL-PCA. *Chemom. Intell. Lab. Syst.*, **76**, 1-13.

Chew, W. and P. Sharratt (2010). Trends in process analytical technology. *Anal. Methods*, **2**, 1412-1438.

Chiang, L.H. and L.F. Colegrove (2007). Industrial implementation of on-line multivariate quality control. *Chemom. Intell. Lab. Syst.*, **88**, 143-153.

Chong, I.G., and C.H. Jun (2005). Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.*, **78**, 103-112.

Çinar, A., S.J. Parulekar, C. Ündey, and G. Birol (2003). *Batch fermentation modeling, monitoring, and control*. Marcel Dekker, Inc., New York (U.S.A.).

Conlin, A.K., E.B. Martin and A.J. Morris (2000). Confidence limits for contribution plots. *J. Chemom.*, **14**, 725-736.

Dayal, B.S. and J.F. MacGregor (1997). Recursive exponentially weighted PLS and its application to adaptive control and prediction. *J. Process Control*, **7**, 169-179.

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.*, **18**, 251-263.

Di Pasquale, N., D.L. Marchisio and A.A Barresi. (2012) Model validation for precipitation in solvent-displacement processes. *Chem. Eng. Sci.*, **84**, 671-683.

Dobry, D.E., D.M. Settell, J.M. Baumann, R.J. Ray, L.J. Graham and R.A. Beyerinck (2009). A model-based methodology for spray-drying process development. *J. Pharm. Innov.*, **4**, 133-142.

Doyle III, F.J., C.A. Harrison and T.J. Crowley (2003). Hybrid model-based appraoch to batch-to-batch control of particle size distribution in emulsion polymerization. *Computers Chem. Eng.*, **27**, 1153-1163.

Dubey, A., A. Sarkar, M. Ierapetritou, C.R. Wassgren and F.J. Muzzio (2011). Computational approaches for studying the granular dynamics of continuous blending processes, 1 – DEM based methods. *Macromol. Mat. Eng.*, **296**, 290-307.

Duchesne, C. and J.F. MacGregor (2001). Jackknife and bootstrap methods in the identification of dynamic models. *J. Process Control*, **11**, 553-564.

Duchesne, C. and J.F. MacGregor (2004). Establishing multivariate specification regions for incoming materials. *J. Quality Techn.*, **36**, 78-94.

Dumarey, M., H. Wikström, M. Fransson, A. Sparén, P. Tajarobi, M. Josefon and J. Trygg (2011). Combining experimental design and orthogonal projections to latent structures to study the influence of microcrystalline cellulose properties on roll compaction. *Int. J. Pharm.*, **416**, 110-119.

Emery, E., J. Oliver, T. Pugsley, J. Sharma and J. Zhou (2009). Flowability of moist pharmaceutical powders. *Powder Techn.*, **189**, 409-415.

Eriksson, L., E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström and S. Wold (2006). *Multi- and megavariate data analysis. Part I. Basic principles and applications*. Umetrics AB, Umeå (Sweden).

Facco, P., F. Bezzo and M. Barolo (2010). Nearest-Neighbor method for the automatic maintenance of multivariate statistical soft sensors in batch processing. *Ind. Eng. Chem. Res.*, **49**, 2336-2347.

FDA (2004a). Innovation or stagnation. Challenge and opportunity on the critical path to new medical products. *U.S. Department of Health and Human Services*. *U.S. Food and Drug Administration*.

FDA (2004b). Pharmaceutical CGMPs for the 21st century – A risk based approach. Final report. *U.S. Department of Health and Human Services*. *U.S. Food and Drug Administration*.

FDA (2004c). Guidance for industry. PAT – A framework for innovative pharmaceutical development, manufacturing and quality assurance. *Center for Drug Evaluation and Research*, *U.S. Food and Drug Administration*, Rockwille (MD), USA.

Feudale, R.N., N.A. Woody, H. Tan, A.J. Myles, S.D. Brown and J. Ferré (2002). Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst.*, **64**, 181-192.

Flores-Cerillo, J. and J.F. MacGregor (2004). Control of batch product quality by trajectory manipulation using LV models. *J. Process Control*, **14**, 539-553.

Flores-Cerillo, J. and J.F. MacGregor (2005). Latent variable MPC for trajectory tracking in batch processes. *J. Process Control*, **15**, 651-663.

Fujiwara, K., M. Kano and S. Hasebe (2010). Development of correlation-based clustering method and its application to software sensing. *Chemom. Intell. Lab. Syst.*, **101**, 130-138.

Fujiwara, K., M. Kano and S. Hasebe (2011). Correlation-based spectral clustering for flexible process monitoring. *J. Process Control*, **21**, 1438-1448.

Gabrielsson, J., N.-O. Lindberg and T. Lundstedt (2002). Multivariate methods in pharmaceutical applications. *J. Chemom.*, **16**, 141-160.

Gabrielsson, J., N.-O. Lindberg, M. Pålsson, F. Nicklasson, M. Sjöström and T. Lundstedt (2003). Multivariate methods in the development of a new tablet formulation. *Drug. Dev. Ind. Pharm.*, **29**, 1053-1075.

Gabrielsson, J., N.-O. Lindberg, M. Pålsson, F. Nicklasson, M. Sjöström and T. Lundstedt (2004). Multivariate methods in the development of a new tablet formulation: optimization and validation. *Drug. Dev. Ind. Pharm.*, **30**, 1037-1049.

GAMS Development Corporation (2010). GAMS Distribution 23.6.5. GAMS Development Corp., Washington D.C. (U.S.A.).

García-Muñoz, S., T. Kourti, J.F. MacGregor, A.G. Mateos and G. Murphy (2003). Troubleshooting of an industrial batch process using multivariate methods. *Ind. Eng. Chem. Res.*, **42**, 3592-3601.

García-Muñoz, S., T. Kourti and J.F. MacGregor (2004). Model Predictive Monitoring for Batch Processes. *Ind. Eng. Chem. Res.*, **43**, 5929-5941.

García-Muñoz, S. (2004). *Batch process improvement using latent variable methods*. McMaster University PhD Thesis.

García-Muñoz, S., J.F. MacGregor and T. Kourti (2005). Product transfer between sites using Joint-Y PLS. *Chemom. Intell. Lab. Syst.*, **79**, 101-114.

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

García-Muñoz, S., J.F. MacGregor, D. Neogi, B.E. Letshaw and S. Mehta (2008). Optimization of batch operating policies. Part II. Incorporating process constraints and industrial applications. *Ind. Eng. Chem. Res.*, **47**, 4202-4208.

García Muñoz, S. and D. Settell (2009). Application of multivariate latent variable modeling to pilot-scale spray drying monitoring and fault detection: Monitoring with fundamental knowledge. *Computers Chem. Eng.*, **33**, 2106-2110.

García Muñoz, S., L. Zhang and M. Cortese (2009). Root cause analysis during process development using Joint-Y PLS. *Chemom. Intell. Lab. Syst.*, **95**, 101-105.

García Muñoz, S. (2009). Establishing multivariate specifications for incoming materials using data from multiple scales. *Chemom. Intell. Lab. Syst.*, **98**, 51-57.

García-Muñoz, S. and C.A. Oksanen (2010). Process modeling and control in drug development and manufacturing. *Computers Chem. Eng.*, **34**, 1007-1008.

García-Muñoz, S. and A. Camody (2010). Multivariate wavelet texture analysis for pharmaceutical solid product characterization. *Int. J. Pharm.*, **398**, 97-106.

García-Muñoz, S. and D. Gierer (2010). Coating assessment for colored immediate release tablets using multivariate image analysis. *Int. J. Pharm.*, **395**, 104-113.

García-Muñoz S., S. Dolph and H.W. Ward II (2010). Handling uncertainty in the establishment of a design space for the manufacture of a pharmaceutical product. *Computers Chem. Eng.*, **34**, 1098-1107.

García-Muñoz, S. and M.A. Polizzi (2012). WSPLS – A new approach towards mixture modeling and accelerated product development. *Chemom. Intell. Lab. Syst.*, **114**, 116-121.

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

Gernaey, K.V., A.E. Cervera-Padrell and J.M. Woodley (2012). A perspective on PSE in pharmaceutical process development and innovation. *Computers Chem. Eng.*, **42**, 15-29.

Halstensen, M., P. de Bakker and K.H. Esbensen (2006). Acoustic chemometric monitoring of an industrial granulation production process – a PAT feasibility study. *Chemom. Intell. Lab. Syst.*, **84**, 88-97.

Hamad, M.L., K. Bowman, N. Smith, X. Sheng and K.R. Morris (2010). Multi-scale pharmaceutical process understanding: from particle to powder to dosage form. *Chem. Eng. Sci.*, **65**, 5625-5638.

Hatzantonis, H., A. Goulas, and C. Kiparissides (1998). A comprehensive model for the prediction of particle size distribution in catalyzed olefin polymerization fluidized-bed reactors. *Chem. Eng. Sci.*, **53**, 3251-3267.

Hermanto, M.W., R.D. Braatz and M.-S. Chiu (2011). Integrated batch-to-batch and nonlinear model predictive control for polymorphic transformation in pharmaceutical crystallization. *AIChE J.*, **57**, 1008-1019.

Hinz, D.C. (2006). Process analytical technologies in the pharmaceutical industry: the FDA's PAT initiative. *Anal. Bioanal. Chem.*, **384**, 1036-1042.

Holland, P.W. and P.R. Rosenbaum (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, **14**, 1523-1543.

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.

Huang, J., G. Kaul, C. Cai, R. Chatlapalli, P. Hernandez-Abad, K. Ghosh, A. Nagi (2009). Quality by design case study: an integrated approach to drug product and process development. *Int. J. Pharm.*, **382**, 23-32.

Hwang, D., G. Stephanopoulos and C. Chan (2004). Inverse modeling using multi-block PLS to determine the environmental conditions that provide optimal cellular function. *Bioinf.*, **20**, 487-499.

IBM Business Consulting Services (2005). Transforming industrialization. A new paradigm for pharmaceutical development. Available at http://www-935.ibm.com/services/uk/igs/pdf/ge510-3997-transforming-industrialization.pdf (Last accessed 18/12/2012).

ICH (1999). ICH harmonised tripartite guide. Specifications: test procedures and acceptance criteria for new drug substances and new drug products: chemical substances. Q6A.

ICH (2005). ICH harmonised tripartite guide. Quality risk management Q9.

ICH (2008). ICH harmonised tripartite guide. Pharmaceutical quality system Q10.

ICH (2009). ICH harmonised tripartite guide. Pharmaceutical development Q8(R2).

ICH (2010). Quality implementation working group on Q8, Q9 and Q10. Questions & Answers (R4).

ICH (2011). ICH quality implementation working group. Points to consider (R2). ICH-endorsed guide for ICH Q8/Q9/Q10 implementation.

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

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

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

Jaeckle, C.M. and J.F. MacGregor (2000b). Product transfer between plants using historical process data. *AIChE J.*, **46**, 1989-1997.

Johnson, R.A. and D. W. Wichern (2007). *Applied multivariate statistical analysis (6th ed.)*. Pearson Education, Inc., Upper Saddle River, NJ (U.S.A.).

Kadlec, P., B. Gabrys and S. Strandt (2009). Data-driven soft sensors in the process industry. *Computers Chem. Eng.*, **33**, 795-814.

Kapsi, S.G., L.D. Castro, F.X. Muller and T.J. Wrzosek (2012). Development of a design space for a unit operation: illustration using compression-mix blending process for the manufacture of a tablet dosage form. *J. Pharm. Innov.*, **7**, 19-29.

Katdare, A. and M.V. Chaubal (2006). *Excipient development for pharmaceutical, biotechnology, and drug delivery systems*. Informa Healthcare, Inc., New York (U.S.A.).

Ketterhagen, W.R., M.T. am Ende and B.C. Hancock (2009). Process modeling in the pharmaceutical industry using the discrete element method. *J. Pharm. Sci.*, **98**, 442-470.

Kim, M., H. Chung, Y. Woo and M.S. Kempes (2007). A new non-invasive, quantitative Raman technique for the determination of an active ingredient in pharmaceuical liquids by direct measurement through a plastic bottle. *Anal. Chim. Acta*, **587**, 200-207.

Kirdar, A.O., K.D. Green and A.S. Rathore (2008). Application of multivariate data analysis for the identification and successful resolution of a root cause for a bioprocessing application. *Biotech. Progr.*, **24**, 720-726.

Kourti, T., P. Nomikos and J.F. MacGregor (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process Control*, **5**, 277-284.

Kourti, T. (2003). Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemom.*, **17**, 93-109.

Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *Int. J. Adapt. Control Signal Process.*, **19**, 213-246.

Kourti, T. (2006). Process analytical technology beyond real-time analyzers: the role of multivariate analysis. *Crit. Rev. Anal. Chem.*, **36**, 257-278.

Krasnyk, M., M. Mangold, S. Ganesan and L. Tobiska (2012). Numerical reduction of a crystallizer model with internal and external coordintaes by proper orthogonal decomposition. *Chem. Eng. Sci.*, **70**, 77-86.

Kremer, D.M. and B.C. Hancock (2006). Process simulation in the pharmaceutical industry: a review of some basic physical models. *J. Pharm. Sci.*, **95**, 517-529.

Ku, W., R.H. Storer and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemom. Intell. Lab. Syst.*, **30**, 179-196.

Kucheryavski, S., K.H. Esbensen and A. Bogomolov (2010). Monitoring of pellet coating process with image analysis – a feasibility study. *J. Chemom.*, **24**, 472-480.

Kumar, V. (2003). Computational material science: the emergence of predictive capabilities of material behaviour. *Sadhana*, **28**, 815-831.

Kushner IV, J. and F. Moore (2010). Scale-up model describing the impact of lubrication on tablet strength. *Int. J. Pharm.*, **399**, 19-30.

Lakshminarayanan, S., H. Fuji, B. Grosman, E. Dassau, and D.R. Lewin (2000). New product design via analysis of historical databases. *Computers Chem. Eng.*, **24**, 671-676.

Landín, M., R.C. Rowe and P. York (2009). Advantages of neurofuzzy logic against conventional experimental design and statistical analysis in studying and developing direct compression formulations. *Europ. J. Pharm. Sci.*, **38**, 325-331.

Lepore, J. and J. Spavins (2008). PQLI design space. *J. Pharm. Innov.*, **3**, 79-87.

Leuenberger, H. (2001). New trends in the production of pharmaceutical granules: batch versus continuous processing. *Europ. J. Pharm. Biopharm.*, **52**, 289-296.

Li, B., J. Morris and E.B. Martin (2002). Model selection for partial least squares regression. *Chemom. Intell. Lab. Syst.*, **64**, 79-89.

Li, W., H.H. Yue, S. Valle-Cervantes and S.J. Qin (2000). Recursive PCA for adaptive process monitoring. *J. Process Control*, **10**, 471-486.

Lince, F., D.L. Marchisio and A.A. Barresi (2008). Strategies to control the particle size distribution of poly-ε-caprolactone nanoparticles for pharmaceutical applications. *J. Coll. Interf. Sci.*, **322**, 505-515.

Lince, F., D.L. Marchisio and A.A. Barresi (2009). Smart mixers and reactors for the production of pharmaceutical nanoparticles: proof of concept. *Chem. Eng. Res. Des.*, **87**, 543-549.

Lince, F., D.L. Marchisio and A.A. Barresi (2011a). A comparative study for nanoparticle production with passive mixers via solvent-displacement: use of CFD models for optimization and design. *Chem. Eng. Process.*, **50**, 356-368.

Lince, F., S. Bolognesi, B. Stella, D.L. Marchisio, F. Dosio (2011b) Preparation of polymer nanoparticles loaded with doxorubicin for controlled drug delivery. *Chem. Eng. Res. Des.*, **89**, 2410-2419.

Liu, Z., M.-J. Bruwer, J.F. MacGregor, S.S.S. Rathore, D.E. Reed and M.J. Champagne (2011a). Modeling and optimization of a tablet manufacturing line. *J. Pharm. Innov.*, **6**, 170-180.

Liu, Z., M.-J. Bruwer, J.F. MacGregor, S.S.S. Rathore, D.E. Reed and M.J. Champagne (2011b). Scale-Up of a pharmaceutical roller compaction process using a Joint-Y partial least squares model. *Ind. Eng. Chem. Res.*, **50**, 10696-10706.

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

Lourenço, V., T. Herdling, G. Reich, .C. Menezes and D. Lochmann (2011). Combining microwave resonance technology to multivariate data analysis as a novel PAT tool to improve process understanding in fluid bed granulation. *Europ. J. Pharm. Biopharm.*, **78**, 513-521.

Lourenço, V., D. Lochmann, G. Reich, J.C. Menezes, T. Herdling and J. Schewitz (2012). A quality by design study applied to an industrial pharmaceutical fluid bed granulation. *Europ. J. Pharm. Biopharm.*, **81**, 438-447.

Lu, J. and F. Gao (2008a). Process modeling based on process similarity. *Ind. Eng. Chem. Res.*, **47**, 1967-1974.

Lu, J. and F. Gao (2008b). Model migration with inclusive similarity for development of a new process model. *Ind. Eng. Chem. Res.*, **47**, 9508-9516.

Lu, J., K. Yao and F. Gao (2009). Process similarity and developing new process models through migration. *AIChE J.*, **55**, 2318-2328.

Lundstedt-Enkel, K., J. Gabrielsson, H. Olsman, E. Seifert, J. Pettersen, P.M. Lek, A. Boman and T. Lundstedt (2006). Different multivariate approaches to material discovery, process development, PAT and environmental process monitoring. *Chemom. Intell. Lab. Syst.*, **84**, 201-207.

Mantanus, J., E. Ziémons, P. Lebrun, E. Rozet, R. Klinkenberg, B. Streel, B. Evrard and P. Hubert (2010). Active content determination of non-coated pharmaceutical pellets by near-infrared spectroscopy: method development, validation and reliability evaluation. *Talanta*, **80**, 1750-1757.

Martens, H. (2001). Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. *Chemom. Intell. Lab. Syst.*, **58**, 85-95.

Matero, S., S. Poutiainen, J. Leskinen, K. Järvinen, J. Ketolainen, A. Poso and S.P. Reinikainen (2010). Estimation of granule size distribution for batch fluidized bed granulation process using acoustic emission and N-way PLS. *J. Chemom.*, **24**, 464-471.

Mathworks (2012). Matlab 8.0 (R2012b). The MathWorks Inc., Natick, MA (U.S.A.).

MacGregor, J.F., C. Jaeckle, C. Kiparissides and M. Koutoudi (1994). Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.*, **40**, 826-838.

MacGregor, J.F. and T. Kourti (1995). Statistical process control of multivariate processes. *Control Eng. Practice*, **3**, 403-414.

MacGregor, J.F. and M-.J. Bruwer (2008). A framework for the development of design and control spaces. *J. Pharm. Innov.*, **3**, 15-22.

Maltesen, M.J., S. Bjerregaard, L. Hovgaard, S. Havelund and M. van de Weert (2008). Quality by Design – Spray drying of insulin intended for inhalation. *Europ. J. Pharm. Biopharm.*, **70**, 828-838.

Mardia, K.V., J.T. Kent and J.M. Bibby (1979). *Multivariate analysis*. Academic Press Limited, London (U.K.).

Meyer, C.D. (2000). *Matrix analysis and applied linear algebra*. SIAM, Philadelphia, PA (U.S.A.).

Moghimi, S.M., A.C. Hunter and J.C. Murray (2001). Long-circulating and target-specific nanoparticles: Theory to practice. *Pharmacol. Rev.*, **53**, 283-318.

Montgomery, D.C. (2005a). *Design and analysis of experiments. 6th edition*. John Wiley & Sons, Inc., New York (U.S.A.).

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

Montgomery, D.C. and G.C. Runger (2010). *Applied statistics and probability for engineers*. John Wiley & Sons, Inc., New York (U.S.A.).

Mortier, S.T.F.C., T. De Beer, K.V. Gernaey, J.P. Remon, C. Vervaet and I. Nopens (2011). Mechanistic modelling of fluidized bed drying processes of wet porous granules: a review. *Europ. J. Pharm. Biopharm.*, **79**, 205-225.

Moteki, Y. and Y. Arai (1986). Operation planning and quality design of a polymer process. *IFAC DYCORD*, Pergamon Press, Bournemouth (U.K.), 159-165.

Mullard, A. (2011). 2010 FDA drug approvals. *Nat. Rev. Drug Discov.*, **10**, 82-85.

Muteki, K., J.F. MacGregor and T. Ueda (2006). Rapid development of new polymer blends: the optimal selection of materials and blend ratios. *Ind. Eng. Chem. Res.*, **45**, 4653-4660.

Muteki, K. and J.F. MacGregor (2007a). Sequential design of mixture experiments for the development of new products. *J. Chemom.*, **21**, 496-505.

Muteki, K. and J.F. MacGregor (2007b). Multi-block PLS modeling for L-shape data structures with applications to mixture modeling. *Chemom. Intell. Lab. Syst.*, **85**, 186-194.

Muteki, K. and J.F. MacGregor (2008). Optimal purchasing of raw materials: a data-driven approach. *AIChE J.*, **54**, 1554-1559.

Muteki, K., K. Yamamoto, G.L. Reid and M. Krishnan (2011). De-risking scale-up of a high-shear wet granulation process using latent variable modeling and near-infrared spectroscopy. *J. Pharm. Innov.*, **6**, 142-156.

Muteki, K., V. Swaminathan, S.S. Sekulic and G.L. Reid (2012). Feed-forward process control strategy for pharmaceutical tablet manufacture using latent variable modeling and optimization technologies. *Proc. of the 8th IFAC Symposium on Advanced Control of Chemical Processes*, July 10-13 2012, Furama Riverfront, Singapore.

Nagy, Z.K., M. Fujiwara and R.D. Braatz (2008). Modeling and control of combined cooling and antisolvent crystallization processes. *J. Process Control*, **18**, 856-864.

Nelson, P.R.C., P.A. Taylor and J.F. MacGregor (1996). Missing data methods in PCA and PLS: score calculations with incomplete observations. *Chemom. Intell. Lab. Syst.*, **35**, 45-65.

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

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

Nomikos, P. and J.F. MacGregor (1995b). Multivariate SPC charts for monitoring batch processes. *Technometrics*¸ **37**, 41-59.

Norioka, T., S. Kikuchi, Y. Onuki, K. Takayama and K. Imai (2011). Optimization of the manufacturing process for oral formulations using multivariate statistical methods. *J. Pharm. Innov.*, **6**, 157-169.

Peinado, A., J. Hammond and A. Scott (2011). Development, validation and transfer of a near infrared method to determine in-line the end point of a fluidised drying process for commercial production batches of an approved oral solid dose pharmaceutical product. *J. Pharm. Biomed. Anal.*, **54**, 13-20.

Peterson, J.J. (2008). A Bayesian approach to the ICH Q8 definition of design space. *J. Biopharm. Stat.*, **18**, 959-975.

Peterson, J.J. and M. Yahyah (2009). A Bayesian design space approach to robustness and system suitability for pharmaceutical assays and other processes. *Stat. Biopharm. Res.*, **1(**4), 441-449.

Plumb, K. (2005). Continuous processing in the pharmaceutical industry. Changing the mind set. *Chem. Eng. Res. Des.*, **83**, 730-738.

Polizzi, M.A. and S. García-Muñoz (2011). A framework for *in-silico* formulation design using multivariate latent variable regression methods. *Int. J. Pharm.*, **418**, 235-242.

Pöllanen, K., A. Hakkinen, S.P. Reinikainen, J. Rantanen, M. Karjalainen, M. Louhi-Kultanen and L. Nystrom (2005). IR spectroscopy together with multivariate data analysis as a process analytical tool for in-line monitoring of crystallization process and solid-state analysis of crystalline product. *J. Pharm. Biomed. Anal.*, **38**, 275-284.

Pomerantsev, A.L. and O.Y. Rodionova (2012). Process analytical technology: a critical view of the chemometricians. *J. Chemom.*, **26**, 299-310.

Poon, J.M., R. Ramachandran, C.F.W. Sanders, T. Glaser, C.D. Immanuel and F.J. Doyle III (2009). Experimental validation studies on a multi-dimensional and multi-scale population balance model of batch granulation. *Chem. Eng. Sci.*, **64**, 775-786.

Pordal, H.S., C.J. Matice and T.J. Fry (2002). The role of computational fluid dynamics in the pharmaceutical industry. *Pharm. Tech.*, **26**(2), 72-77.

Portillo, P.M., M. Ierapetritou, S. Tomassone, C. Mc Dade, D. Clancy, P.P. Avontuur and F.J. Muzzio (2008). Quality by design methodology for development and scale-up of batch mixing processes. *J. Pharm. Innov.*, **3**, 258-270.

Qin, S.J. (1998). Recursive PLS algorithms for adaptive data modeling, *Computers Chem. Eng.*, **22**, 503-514.

Quesada, I. and I.E. Grossman (1992). An LP/NLP based branch and bound algorithm for convex MINLP optimization problems. *Computers Chem. Eng.*, **16**, 937-947.

Rajalahti, T. and O.M. Kvalheim (2011). Multivariate data analysis in pharmaceutics: a tutorial review. *Int. J. Pharm.*, **417**, 280-290.

Ramachandran, R., C.D. Immanuel, F. Stepanek, J.D. Litster and F.J. Doyle III (2009). A mechanistic model for brekeage in population balances of granulation: theoretical kernel development and experimental validation. *Chem. Eng. Res. Des.*, **87**, 598-614.

Ramachandran, R., J. Ariunan, A. Chaudhury and M. Ierapetritou (2011). Model-based control-loop performance of a continuous direct compaction process. *J. Pharm. Innov.*, **6**, 249-263.

Rambali, B., L. Baert and D.L. Massart (2001). Using experimental design to optimize the process parameters in fluidized bed granulation on a semi-full scale. *Int. J. Pharm.*, **220**, 149-160.

Rännar, S., J.F. MacGregor and S. Wold (1998). Adaptive batch monitoring using hierarchical PCA. *Chemom. Intell. Lab. Syst.*, **41**, 73-81.

Remy, B., J.G. Khinast and B.J. Glasser (2009). Discrete element simulation of free flowing grains in four-bladed mixer. *AIChE J.*, **55**, 2035-2048.

Rowe, R.C., and R.J. Roberts (1998). Artificial intelligence in pharmaceutical product formulation: neural computing and emerging technologies. *Pharm. Sci. Tech. Today*, **1**, 200-205.

Saerens, L., L. Dierickx, T. Quinten, P. Adriaensens, R. Carleer, C. Vervaet, J.P. Remon and T. De Beer (2012). In-line NIR spectroscopy for the understanding of polymer-drug interaction during pharmaceutical hot-melt extrusion. *Europ. J. Pharm. Biopharm.*, **81**, 230-237.

Sahinidis, N.V. (1996). BARON – A general purpose global optimization software package. *J. Global Optim.*, **8**, 201-205.

Schaber, S.D., D.I. Gerogiorgis, R. Ramachandran, J.M.B. Evans, P.I. Barton and B.L. Trout (2011). Economic analysis of integrated continuous and batch pharmaceutical manufacturing: a case study. *Ind. Eng. Chem. Res.*, **50**, 10083-10092.

Sebzalli, Y.M. and X.Z. Wang (2001). Knowledge discovery from process operational data using PCA and fuzzy clustering. *Eng. App. Artif. Intel.*, **14**, 607-616.

Shah, R.B., M.A. Tawakkul and M.A. Khan (2007). Process analytical technology: chemometric analysis of Raman and near infra-red spectroscopic data for predicting physical properties of extended release matrix tablets. *J. Pharm. Sci.*, **96**, 1356-1365.

Shao, Q., R.C. Rowe and P. York (2007). Investigation of an artificial intelligence technology – model trees novel applications for an immediate release tablet formulation database. *Europ. J. Pharm. Sci.*, **31**, 137-144.

Sin, G., P. Ödman, N. Petersen, A. Eliasson Lantz and K.V. Gernaey (2008). Matrix notation for efficient development of first-principles models within PAT applications: integrated modeling of antibiotic production with *Steptomyces coelicor*. *Biotech. Progr.*, **25**, 1043-1053.

Singh, R., M. Ierapetritou and R. Ramachandran (2012). An engineering study on the enhanced control and operation of continuous manufacturing of pharmaceutical tablets via roller compaction. *Int. J. Pharm.*, **438**, 307-326.

Smith, B.V. and M.G. Ierapepritou (2010). Integrative chemical product design strategies: reflecting industry trends and challenges. *Computers Chem. Eng.*, **34**, 857-865.

Soh, J.L.P., F. Wang, N. Boersen, R. Pinal, G.E. Peck, M.T. Carvajal, J. Ceney, H. Valthorsson, J. Pazdan (2008). Utility of multivariate analysis in modeling the effects of raw material properties and operating parameters on granule and ribbon properties prepared in roller compaction. *Drug Dev. Ind. Pharm.*, **34**, 1022-1035.

Stockdale, G.W. and A. Cheng (2009). Finding design space and a reliable operating region using a multivariate bayesian approach with experimental design. *Quant. Management*, **6**, 391-408.

Streefland, M., P.F.G. Van Herpen, B. Van de Waterbeemd, L.A. Van der Pol, E.C. Beuvery, J. Tramper. D.E. Martens and M. Toft (2009). A practical approach for exploration and modeling of the design space of a bacterial vaccine cultivation process. *Biotech. Bioeng.*, **104**, 492-504.

Sun, Y., Y. Peng, Y. Chen and A.J. Shukla (2003). Appliation of artifical neural network in the design of controlled release drug delivery systems. *Adv. Drug Deliv. Rev.*, **55**, 1201-1215.

Takayama, K., M. Fujiwara, Y. Obata and M. Morishita (2003). Neural network based optimization of drug formulations. *Adv. Drug Deliv. Rev.*, **55**, 1217-1231.

ter Horst, J.H., H.J.M. Kramer and P.J. Jansens (2006). Towards a crystalline product quality prediction method by combining process modeling and molecular simulations. *Chem. Eng. Technol.*, **29**, 175-181.

Thirunahari, S., P. Shan Chow and R.B.H. Tan (2011). Quality by Design (QbD)-based crystallization process development for the polymorphic drug tolbutamide. *Cryst. Growth Des.*, **11**, 3027-3038.

Thomassen, Y.E., E.N.M. van Sprang, L.A. van der Pol, W.A.M. Bakker (2010). Multivariate data analysis on historical IPV production data for better process understanding and future improvements. *Biotech, Bioeng.*, **107**, 96-104.

Trygg, J. and S. Wold (2002). Orthogonal projections to latent strucutres (O-PLS). *J. Chemom.*, **16**, 119-128.

Valente, I., E. Celasco, D.L. Marchisio and A.A. Barresi (2012). Nanoprecipitation in confined impinging jets mixers: production, characterization and scale-up of pegylated nanospheres and nanocapsules for pharmaceutical use. *Chem. Eng. Sci.*, **77**, 217-227.

Valle, S., W. Li and S.J. Qin (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Ind. Eng. Chem. Res.*, **38**, 4389-4401.

Vanarase, A.U., M. Alcala, J.I.J. Rozo, F.J. Muzzio and R.J. Romanach (2010). Real-time monitoring of drug concentration in a continuous powder mixing process using NIR spectroscopy. *Chem. Eng. Sci.*, **65**, 5728-5733.

Varmuza, K. and P. Filzmoser (2009). *Introduction to multivariate statistical analysis in chemometrics*. CRC Press, Boca Raton, FL (U.S.A.).

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

Verma, S., Y. Lan, R. Gokhale and D.J. Burgess (2009). Quality by design approach to understand the process of nanosupsension preparation. *Int. J. Pharm.*, **377**, 185-198.

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

Walczak, B. and D.L. Massart (2001b). Dealing with missing data: Part II. *Chemom. Intell. Lab. Syst.*, **58**, 29-42.

Wan, J., O. Marjanovic and B. Lennox (2012). Disturbance rejection for the control of batch end-product quality using latent variable models. *J. Process Control*, **22**, 643-652.

Wangen, L.E. and B.R. Kowalski (1989). A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemom.*, **3**, 3-20.

Wassgren, C. and J.S. Curtis (2006). The aplication of computational modeling to pharmaceutical materials science. *MRS Bulletin*, **31**, 900-904.

Westerhuis, J.A. and P.M.J. Coenegracht (1997). Multivariate modelling of the pharmaceutical two-step process of wet granulation and tabletting with multiblock partial least squares. *J. Chemom.*, **11**, 379-392.

Westerhuis, J.A., P.M.J. Coenegracht and C.F. Lerk (1997). Multivariate modelling of the tablet manufacturing process with wet granulation for tablet optimization and in-process control. *Int. J. Pharm.*, **156**, 109-117.

Westerhuis, J.A., T. Kourti and J.F. MacGregor (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.*, **12**, 301-321.

Wiklund, S., D. Nilsson, L. Eriksson, M. Sjöström, S. Wold and K. Faber (2007). A randomization test for PLS component selection. *J. Chemom.*, **21**, 427-439.

Winkle, H.N. (2007). Implementing quality by design. *Proc. of PDA/FDA joint regulatory conference*, September 24-28 2007, Washington D.C., U.S.A.

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

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

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate analysis*, Academic Press Limited, New York (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 and H. Wold (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Lecture Notes in Math.*, **973**, 286-293.

Wold, S., M. Sjöström, R. Carlson, T. Lundstedt, S. Hellberg, B. Skagerberg, C. Wikström and J. Öhman (1986). Multivariate design. *Anal. Chim. Acta*, **191**, 17-32.

Wold, S., N. Kettaneh, H. Fridèn and A. Holmberg (1998). Modeling and diagnostics of batch processes and analogous kinetics experiments. *Chemom. Intell. Lab. Syst.*, **44**, 331-340.

Woo, X.Y., R.B.H. Tan and R.D. Braatz (2009). Modeling and computational fluid dynamics - population balance equation - micromixing simulation of impinging jet crystallizers. *Cryst. Growth Des.*, **9**, 156-164.

Yacoub, F. and J.F. MacGregor (2004). Product optimization and control in the latent variable space of nonlinear PLS models. *Chemom. Intell. Lab. Syst.*, **70**, 63-74.

Yacoub, F., J. Lautens, L. Lucisano and W. Banh (2011a). Application of quality by design principles to legacy drug products. *J. Pharm. Innov.*, **6**, 61-68.

Yacoub, F. and J.F. MacGregor (2011b). Robust processes through latent variable modeling and optimization. *AIChE J.*, **57**, 1278-1287.

Yu, L.X. (2008). Pharmaceutical quality-by-design: product and process development, understanding and control. *Pharm. Res.*, **25**, 781-791.

Zacour, B.M., J.K. Drennen III and C.A. Anderson (2012a). Development of a statistical tolerance-based fluid bed drying design space. *J. Pharm. Innov.*, DOI: 10.1007/s12247-012-9133-y.

Zacour, B.M., J.K. Drennen III and C.A. Andreson (2012b). Development of a fluid bed granulation design space using critical quality attribute weighted tolerance intervals. *J. Pharm. Sci.*, **101**, 2917-2929.

Zacour, B.M., J.K. Drennen III and C.A. Anderson (2012c). Hybrid controls combining first-principle calculations with empirical modeling for fully automated fluid bed processing. *J. Pharm. Innov.*, DOI: 10.1007/s12247-012-9137-7.

Ziémons, E., J. Mantanus, P. Lebrun, E. Rozet, B. Evrard and P. Hubert (2010). Acetaminophen determination in low-dose pharmaceutical syrup by NIR spectroscopy. *J. Pharm. Biomed. Anal.*, **53**, 510-516.

Zimmermann, V., H. Hennemann, T. Daußmann and U. Kragl (2007). Modelling the reaction course of *N*-acetylneuraminic acid synthesis form *N*-acetyl-d-glucosamine – New strategies for the optimisation of neuraminic acid synthesis. *Appl. Microbiol. Biotech.*, **76**, 597-605.

Zlokarnik, M. (2006). *Scale up in chemical engineering*. Wiley-VCH, Weinheim (Germany).

Zomer, S., M. Gupta and A. Scott (2010). Application of multivariate tools in pharmaceutical product development to bridge risk assessment to continuous verification in a quality by design environment. *J. Pharm. Innov.*, **5**, 109-118.

Acknowledgements

There are many people I need to thank for their intellectual, emotional and technical contribution to this work and for their support along this journey.

First of all I would like to express my most sincere gratitude to my supervisor, Prof. Massimiliano Barolo, and to Prof. Fabrizio Bezzo, for their guidance, support, encouragement, patience and help without which this Thesis would not have been possible.

Special thanks to Dr. Pierantonio Facco for his invaluable contribution, support and for the helpful discussions and comments which have enriched the value of this Thesis.

I will always be grateful to Dr. Salvador García-Muñoz for his incomparable help and guidance in this work and for his kindness and support during my visit in Pfizer. Sal, this Thesis would have not been the same without all your teachings and guidance! Thank you for your essential contribution to my professional growth and for your kind help which went far beyond the academical work. Thanks to all the kind guys I had the chance to work with in Pfizer.

Warm thanks to Dr. Simeone Zomer and Dr. John Robertson from GSK for their kindness and for the valuable discussions we had. Special thanks to Dr. Zomer for his invaluable support and precious friendship during my stay in GSK. Thanks to all the people I had the chance to meet there.

Thanks to Prof. Daniele Marchisio, Prof. Antonello Barresi for the helpful discussions and comments. Thanks to Tereza Zelenková for having been so kind to perform all the experiments I needed.

Thanks to Natascia Meneghetti and Martina Largoni for their kind help and contributions.

Finally, I would like to thank all my peers at CAPE-Lab, with whom I shared this journey: Federico, Matteo, Andrea, Ricardo, and all my nice office mates during these years.

Infine ringrazio la mia famiglia per essermi sempre stata vicino e avere assecondato tutte le mie scelte.

Grazie a Letizia per essere sempre presente, per la pazienza e l'Amore che mi dimostri tutti i giorni.

Un ringraziamento particolare ad Alessandro per i tre lunghi anni condivisi assieme. Un grazie a tutti gli amici di sempre per tutti i momenti che passiamo assieme.

Un pensiero particolare a chi non c'è più affinchè dall'Alto mi illumini sempre.