UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale in Ingegneria Chimica e dei Processi Industriali

IMPACT OF MEASUREMENT ERROR IN BAYESIAN DESIGN SPACE DETERMINATION FOR PHARMACEUTICAL PROCESSES

Relatore: Prof. Massimiliano Barolo

Correlatore: Gabriele Bano

Laureando: Marco Cattaldo

Anno Accademico: 2017 - 2018

A chi mi è stato vicino

Abstract

The pharmaceutical industry is facing a period of drastic change in the way products are conceived and produced, due to the introduction of the *Quality by Design* (QbD) initiative put forth by the pharmaceutical regulatory agencies, such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA). A key concept introduced in the QbD framework is that of design space (DS) of a new pharmaceutical product, defined as "the multidimensional combination and interaction of raw material properties and process parameters that have been demonstrated to provide *assurance* of quality" for the final product. Once the DS has been approved by the regulatory agency, the process can be run within the DS without any further regulatory approval, thus significantly increasing process flexibility and allowing pharmaceutical companies to continuously optimize the operation of their processes.

Bayesian modelling techniques (such as Bayesian multivariate linear regression and joint Bayesian/latent variable modelling) have recently proved to play a key role for the quantification of "assurance" of quality advocated by the regulatory agencies. With these techniques, the different sources of uncertainty (parametric; structural; measurement uncertainty in the calibration dataset) that affect the DS prediction can be handled in a unified framework, and the "assurance" of quality can be quantified as the *probability* that the product under development will meet its specifications. However, the methodologies that are currently available in the literature primarily focus on the incorporation of parametric and structural uncertainty in the Bayesian framework, while systematic approaches for the incorporation of measurement uncertainty are still missing.

The aim of this Thesis is to propose a systematic approach for the incorporation of measurement uncertainty in the Bayesian identification of the DS of a new pharmaceutical product. Specifically, the proposed approach extends a joint Bayesian/latent variable methodology for DS identification recently proposed by Bano *et al*. (2018). A step-by-step methodology is proposed to handle measurement uncertainty in the calibration dataset, and three case studies (two of which involving experimental data of pharmaceutical granulation) are used to test its performance. The results show that the incorporation of measurement uncertainty can, under certain circumstances, significantly affect the prediction of the DS.

Riassunto

Con l'introduzione del concetto di *Quality-by-Design* i vari enti regolatori internazionali hanno di fatto dato inizio ad una rivoluzione nel modo nel quale le aziende farmaceutiche si devono approcciare al processo di produzione e di sviluppo di un prodotto medicale. Questa proposta di cambiamento sistematico è avvenuta in un momento topico; le industrie infatti vedevano la profittabilità dei loro investimenti diminuire mentre i costi dovuti agli errori di produzione continuavano ad aumentare (Abboud and Hensley, 2003). In questa nuova modalità di sviluppo ed operazione dei processi i punti focali dovranno essere la conoscenza approfondita del processo e le decisioni basate su dati scientifici, non sarà più accettabile un approccio basato sull'esperienza dove l'effettiva qualità di un prodotto farmaceutico è testata a campione tra i prodotti finiti. Tra gli altri si aggiunge infatti anche la necessità che la qualità finale di un prodotto sia inserita tra i parametri di sviluppo ed assicurata fin dalla concezione del prodotto e del suo processo produttivo. Oltre agli ovvi benefici al consumatore finale anche le stesse aziende farmaceutiche possono trarre vantaggio da questa iniziativa, infatti un controllo integrato, unito ad una approfondita conoscenza del processo ed alla possiblità di non fallire alcun batch portano benefici economici interessanti. Ai benefici economici si sommano anche degli sgravi regolatori che permettono al processo di essere condotto in modo ottimale senza essere costretti a seguire la ricetta che è stata approvata. Quest'utimo beneficio è legato al concetto di spazio di progetto (*design space*), spazio multivariato delle variabili di progetto ed operative e della loro relazione, che permette di assicurare l'ottenimento di una qualità richiesta. Un movimento dei parametri operativi all'interno delle configurazioni delimitate dallo spazio di progetto non consiste, nel *Quality-by-Design*, in un cambiamento del processo produttivo e di conseguenza non richiede autorizzazioni aggiuntive; un vantaggio per l'azienda come detto, ma anche per l'ente regolatore. La determinazione di uno spazio di progetto deve avvenire con l'ausilio di modelli atti a rappresentare il processo in corso senza dimenticare l'interconnessione tra i parametri operativi e la loro influenza gli uni sugli altri. Uno dei punti chiave nel concetto di spazio di progetto è il fatto che lo spazio di progetto dovrebbe indicare i punti che danno una certa assicurazione di qualità come appartenenti al suddetto, e quindi come condizioni operative e proprietà delle materie prime che possono essere usate nella conduzione del processo. Il fulcro, a parità di validità, deve essere l'effettiva capacità di un metodo per la definizione dello spazio di progetto di dare o meno un effettiva assicurazione della qualità, non solo un indicazione qualitativa. È in questo frangente che i metodi statistici di impostazione Bayesiana si sono rivelati più efficaci. L'obiettivo di questa Tesi è incorporare in una metodologia Bayiesiana per l'identificazione di uno spazio di progetto un ulteriore fonte di incertezza, finora non esplorata in questo frangente: l'incertezza dovuta all'errore di misura.

Al fine del conseguimento di questo obiettivo viene illustrata una metodologia che integra quella porposta in un recente articolo di Bano *et al*., (2018).

La Tesi è strutturata in questo modo. Nel Capitolo 1 è presente una trattazione generale del nuovo approccio proposto dagli enti regolatori con definizione dei termini chiave e delle principali sfide ancora aperte. Nel Capitolo 2 è sono discussi i metodi matematici che sono stati usati per l'ottenimento dei risultati conseguiti o durante le analisi preliminari. Il Capitolo 3 mostra come i metodi matematici sono stati applicati per la definizione di uno spazio di progetto, e come uno spazio di progetto Bayesiano sia in grado di dare l'assicurazione di qualità alla base della ricerca in oggetto, oltre a mostrare la metodologia che il presente lavoro espande. Nel Capitolo 4 il problema dell'errore di misura viene trattato più approfonditamente. Cosa comporta, che fonti ha ed infine come viene affrontato nella metodologia proposta. Nello stesso capitolo è inoltre presentata la metodologia per l'incorporazione dell'errore di misura nella determinazione di uno spazio di progetto Bayesiano per un nuovo prodotto farmaceutico, assieme ad un esempio nel corso del quale la metodologia viene sviscierata e spiegata. Il Capitolo 5 presenta alcuni dei risultati ottenuti mettendo ogni volta in luce come i risultati sono ottenuti e sottolineando il corretto metodo di azione nell'applicazione della metodologia. Una breve conclusione chiude il documento con delle considerazioni finali sulla metodologia e sulle direzione di indagine possibili per ulteriori approfondimenti sull'argomento.

Summary

Introduction

In 2004, the ICH released a set of guidelines in its Q8 (R2) document, adopted in 2006 by the Food and Drugs Administration (FDA) and the European Medicine Agency (EMA), stressing the importance of a Quality by Design (QbD) approach in the pharmaceutical industry.
In the guideline, QbD is described as "a systematic approach to development that begins with predefined objectives and emphasises product and process understanding and process control, based on sound science and quality risk management" (ICH, 2009).
The adoption of the QbD framework in the pharmaceutical context requires a paradigm shift from the traditional Quality by Testing (QbT) approach, where product quality is tested and assessed at the end of the production process, to an enhanced one, based on scientific knowledge, where quality is "built" into the final product since its original conception.
The practical implementation of QbD in the pharmaceutical industry requires the adoption of novel modelling approaches to pharmaceutical product development. Specifically, mathematical modelling is required in order to link the desired critical quality attributes (CQAs) of the final product with the critical process parameters (CPPs) and raw material properties of the process exploited for the production of the desired product. CQA are defined as "physical, chemical, biological, or microbiological properties or characteristics that should be within an appropriate limit, range, or distribution to ensure the desired product quality" (ICH, 2009). On the other hand, CPPs are defined as "process parameters whose variability has an impact on a critical quality attribute and therefore should be monitored or controlled to ensure the process produces the desired quality" (ICH, 2009). Once a mathematical model (data-driven or semi-empirical or first-principles) is obtained in order to link the CPPs and raw material properties to the product CQAs, the so-called design space (DS) of the pharmaceutical product can be obtained. The DS is defined 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" (ICH, 2009). The identification of the DS has several key benefits, many of which stem from the fact that as long as the CPPs and raw materials proprieties are changed inside the boundaries set by the approved DS, regulatory post-approval change procedures are not needed. From the perspective of a pharmaceutical company, this represents a competitive advantage since the process can be run (and therefore its operation can be optimized) within the DS without initiating any regulatory post-approval procedure.
The concept of "assurance" of quality contained in the regulatory definition of DS requires the quantification by the manufacturer of the reliability of the proposed design Space (Peterson,

2008). Several studies (Stockdale and Cheng, 2009; Peterson and Lief, 2010; Debrus *et al*., 2011; Peterson *et al*., 2017) demonstrated the ability of Bayesian methodologies to give a rigorous scientific metric (i.e. the probability that the product will meet its specifications) in order to scientifically quantify the concept of "assurance" of quality. The advantage of these methodologies is that different sources of uncertainty (parametric, structural, measurement uncertainty in the calibration dataset) can be incorporated in a unified framework and can therefore be treated in a straightforward fashion. However, while all the studies available in the literature on Bayesian DS determination focus on parameteric and structural uncertainties, reliable methodologies to include measurement uncertainty in the Bayesian framework are still missing.

The aim of this Dissertation is to present a possible way to include this source of uncertainty in Bayesian DS identification exercise. Specifically, we show how measurement uncertainty can be included in the Bayesian DS determination methodology recently proposed by Bano *et al.*, (2018). The proposed approach has been tested on three case studies. The first is a highly correlated mathematical case study. The second and the third are derived from industrial or experimental practice and regards granulation processes. The second one uses historical data to establish the design space of a roll compaction dry granulation, the third one for a high shear wet granulation.

The Dissertation is divided into five chapters. In the first chapter, the Quality by Design initiative is explained in depth, and the fundamental concepts and terms are explained. In the second chapter, background on some of the principal mathematical tools used in this work is presented. In the third chapter, state of the art in the probabilistic techniques for the determination of a design space is reported along with the Bayesian/Latent variable approach (Bano *et al*., 2018) methodology. In the fourth chapter, a contextualization on the measurement error is presented, and the methodology to include the added uncertainty in the design space determination exercise is reported. In the fifth chapter, three case studies are shown supporting and explaining the methodology. Some final remarks and possible direction for future studies will conclude the thesis.

Chapter 1

Motivation and state of the art

In this Chapter, an overview of the Quality by Design (*QbD*) initiative that has recently been put forth by the pharmaceutical regulatory agencies will be given, with particular focus to its implications for the analyses proposed in the Dissertation.

1.1 Quality by Design

Although it is perceived as cutting edge for its societal impact, the pharmaceutical industry has been known to rely on outdated techniques in product development and manufacture (Tomba , 2013). Product quality and performance were achieved by restricting flexibility in the manufacturing process and by random end-product testing (Yu, 2008). In 2004, the International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use (ICH) released a set of guidelines in its Q8 document, adopted in 2006 by the Food and Drugs Administration (FDA) and the European Medicine Agency (EMA), stressing the need of a different approach in pharmaceutical product design and manufacturing. The issue was not only the inability to quantify the assurance of achievement of the required level of quality, but the cost of inefficiencies linked to the methods in use. In a survey on the state of the pharmaceutical industry. In 2003, the Wall Street Journal in an article from Abboud and Hensley, quantified the percentage of product that failed to pass the desired quality threshold due to manufacturing shortcomings between the 5% and 10% of the produced medicines. To give an additional measure of the manufacturing deficiencies impact on the industry, the number of recalled medicines and the manufacturing expenses linked to those percentages of recall were reported in the article; both increased yearly. In 2002, the FDA counted 354 prescription-drugs recall, up from 248 in 2001, and the linked manufacturing expenses accounted to the 36% of the industry total, more than double what at the time was devoted to research and development (Abboud and Hensley, 2003). The proposal of this change happened at a pivotal time. Many patents on high profitable drugs were approaching, or had already passed, expiration date and development of new molecules to substitute them was lagging. In 2005, the FDA approved 18 new molecules, compared to the 31 of 2000 and to the 53 of 1996 (Hughes, 2009). In these circumstances, the regulatory agencies prompted a new

approach to product development and manufacturing introducing the concept of Quality by Design (QbD). Taking inspiration from other industrial fields the quality by design initiative focuses on the promotion of robust, systematic, science based tools rather than fixed traditional procedures based on experience and custom.

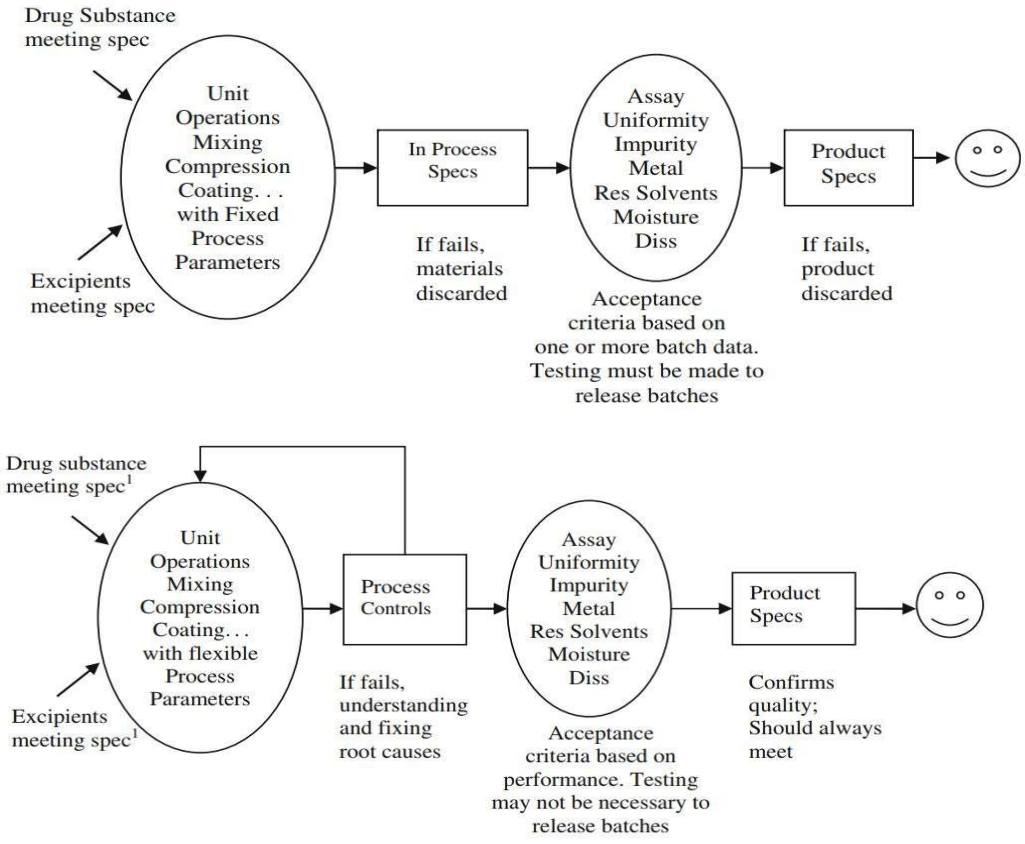

Figure 1.1 *Comparison between two simplified quality assurance diagrams for generic drugs; the upper one refers to the Quality by Testing methodology, the lower one to the Quality by Design one. (Adapted from Yu, 2008)*

According to the Quality by Design paradigm, in contrast to previous Quality by Testing approach, quality must be built into the end-product from its design rather than being evaluated at the end of the manufacturing process. This is cemented in the definition of QbD given in the ICH guidelines as "a systematic approach to pharmaceutical development that begins with predefined objectives and emphasizes product and process understanding based on sound science and quality risk management" (ICH, 2009).

In order to inspire a practical implementation of the QbD initiative, a number of different elements and criteria must be identified and defined. First, it is necessary to define what "quality" is. ICH defines quality as the suitability of a drug substance or drug product for its

intended use. This term includes such attributes as the identity, strength, and purity (ICH, 1999). With this definition in mind, it is possible to introduce the concept of criticality. The concept of criticality can be used to describe any feature or material attribute, property or characteristic of a drug substance, component, raw material, drug product or device, or any process attribute, parameter, condition or factor in the manufacture of a drug product (Garcia et al, 2008). The assessment of a parameter as critical or non-critical is a crucial step of a risk-based approach and leads to the identification of a Quality Target Product Profile (QTPP) for the product under development Critical Quality Attributes (CQAs) and Critical Process Parameters (CPPs).

1.2 Quality by design: terminology

In the following, the main terminology adopted in the Quality by Design framework is briefly summarized.

1.2.1 Quality Target Product Profile (QTTP)

QTPP is defined as a prospective summary of the quality characteristics of a drug product that ideally will be achieved to ensure the desired quality, taking into account safety and efficacy (ICH, 2009). In other words, QTPP are attributes of the desired drug, based on intended use, method of administration, pharmacokinetic mechanisms and so on. Examples of QTPP are, for instance, dosage form and strength, physical qualities such as odour, shape and uniformity, stability or even container closure system, to cite a few (Sun, 2010).
Once defined the QTPP, it is possible to conduct a risk assessment based on data and process knowledge to determine the criticality of what influences the process based on cause and effect relationships, relative to probability, severity, detectability, and sensitivity. Probability is the likelihood of deviation from QTPP, while severity is the entity of the aforementioned deviation. Detectability refers to the ability to discover or determine the existence, presence, or fact of this deviation and sensitivity is the attenuation of interactions between multivariate dimensions (Garcia et al, 2008). With this risk-based assessment, it is possible to define Critical Quality Attributes (CQAs) and Critical Process Parameters (CPPs).

1.2.2 Critical Quality Attributes (CQAs) and Critical Process Parameters (CPPs)

A CQA is defined as a physical, chemical, biological or microbiological property or characteristic that should be within an appropriate limit, range, or distribution to ensure the desired product quality (ICH, 2009). Examples of CQAs are for instance the Content Uniformity of the least concentrated Active Pharmaceutical ingredient (API) in a fixed-dose

combined instant release medicinal capsule or API Dissolution, linked to tablet density, in an extended release coated medicinal capsule (Maguire and Peng, 2015).
A CPP is defined as a process parameter whose variability has an impact on a critical quality attribute and therefore should be monitored or controlled to ensure the process produces the desired quality (ICH, 2009). Examples of CPP are for instance the number of blender revolution in a blending process or the screen opening size in a screening milling process (Maguire and Peng, 2015). 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) as more data and a deeper understanding of process become available to update the criticality assessment.

1.2.3 Design space (DS)

The concept of design space is one of the key definitions on which the QbD framework is based, and its description is expected to be one of the results of the pharmaceutical development investigation (Tomba , 2013). The ICH definition of design space is: “the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality” (ICH, 2009).
With the previous QbT, paradigm manufacturers were not permitted to make changes to operating parameters or other process change without filing Chemistry, Manufacturing and Controls (CMC) supplements as part of the post-approval change process. As a result the regulatory agencies were overwhelmed by post-approval changes requests, for example in 2005 and 2006 FDA Office of Generic Drugs received over 3000 post-approval changes requests annually (Yu, 2008).
The QbT approach placed little or no emphasis on a properly designed process, one that is both efficient and effective can ensure product quality. Furthermore, the regulatory process with its cumbersome requirements to assure strict adherence to the experience-based design, proved in the end to be pyrrhic since it inhibited continuous improvement and real time assurance of quality (Yu, 2008). In contrast with the QbT paradigm, QbD and the introduction of the concept of design space drastically changed the classical way agencies used to supervise pharmaceutical development. Movements inside the design space are not considered as a change and are not undergo a regulatory post-approval procedure.
The design space is considered the ultimate result of the manufacturing process understanding in the development of a new product, due to the necessity of identifying all critical process parameters and critical quality attributes of the feed and of the product. Since it is needed to predict a working zone that quantifies in some way the assurance of quality, the ability to predict, based on historical data and similar production campaign, is honed throughout the product lifecycle and as more data becomes available, culminating as the knowledge of the

process become complete. According to FDA, a process is considered well understood when (FDA, 2004):

- 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.

The ICH pharmaceutical development Q8 document provides a general guideline on how to construct and present a Design Space, without setting a precise methodology, but leaving the initiative to the company to select the most appropriate methods and tools.

While not stated explicitly in the first draft of the document, the second revision of the pharmaceutical development Q8 (R2) document specifies that the multivariate nature of the Design Space, as per its definition, must be taken into account. It is important not only to select the CCPs and CQAs with care but also to measure their interaction with one another and with other parameters or quality attributes that could surpass the edge of criticality and become relevant if not critical to the process. 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. In accordance to the general spirit of the QbD initiative, the regulatory agencies left the companies the possibility of choice on how to establish and present Design Spaces as long as all choices made in it are correctly justified and explained. Design Spaces could be presented for the single units or a single design spaces could be established for multiple operations up to the whole production line. For example, in the case of a drug product that undergoes degradation in solution before lyophilisation, the design space to control the extent of degradation (e.g., concentration, time, temperature) could be expressed for each unit operation or as a sum over all unit operations (ICH, 2009). The same spirit applies to multiple-scale spanning design Spaces. It is necessary for the company to provide data on scale-relevance of the various CQAs and CPPs, along with the specification on which scale the data at the base of the design Space were collected. In this case dimensionless numbers and/or models for scaling can be included as part of the design space description (ICH, 2009).

1.2.4 Process Analytical Technology (PAT)

The Process Analytical Technology (PAT) framework is defined by FDA as “a system for designing, analysing 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. The goal of PAT is to give tools and principles to enhance understanding and control of the manufacturing process (FDA, 2004) ; these tools are then used for clarification via scientific study of the problems encountered in product and process development. In the PAT framework, these tools can be categorized according to the following (FDA, 2004):

- Multivariate tools for design, data acquisition and analysis,
- Process analysers,
- Process control tools,
- Continuous improvement and knowledge management tools.

Multivariate tools include all the multivariate mathematical techniques including, for example, latent variable modelling and Bayesian statistical analysis, applied to the scientific understanding of the relevant multifactorial relationship between inputs and outputs of a specific process and their generalization to a broader class of processes.

Process analysers are all the tools used to collect process data. These measurements can be obtained by removing, isolating and analysing the sample in proximity to the process stream, by diverting the sample from the manufacturing process and returning it to the process stream after the measurement or by keeping the sample inside the process stream, where the measurement can be made invasive or not. Process analysers usually produce a massive amount of data, and dedicated multivariate analysis techniques must be applied to render them usable.

Process control tools are intended to monitor the state of a process and actively manipulate it to maintain the desired trend. Strategies should accommodate the attributes of input materials, the ability and reliability of process analysers to measure critical attributes, and the achievement of process endpoints to ensure consistent quality of the output materials and the final product (FDA, 2004).

The results of the application of the PAT framework is the availability of a large amount of data, from which comprehensive models can be derived, either from first principles - if the process knowledge is mature enough - through empirical modelling or a combination of the two. Since their origin is found in the application of PAT frameworks, models are part of the framework too, and as such are PAT tools themselves.

1.2.5 Control strategy

To maintain a process inside the boundaries defined by its design space, thus ensuring that a product of required quality will be produced consistently, it is necessary to employ an efficient control strategy. It should be noted that the term control does not usually refer to the traditional engineering understanding of process control. A control strategy is defined by the ICH in its pharmaceutical development Q8 (R2) document 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. The elements of the control strategy should describe and justify how in-process controls of the CPPs and the controls of input materials, intermediates and drug products CQAs contribute to the final product quality (ICH, 2009).

Understanding sources of variability and their impact on downstream processes or processing, in-process materials, and drug product quality can provide an opportunity to shift controls upstream and minimise the need for end-product testing. The goal is to design adaptive process steps and appropriate control strategy to ensure that the variability can be addressed in an adaptable way to deliver consistent product quality. This process would an alternative manufacturing paradigm where the input variability could be less tightly constrained.

Enhanced understanding of product performance and appropriate variability control can justify the use of alternative approaches to determine that the material is meeting its CQAs. The use of such alternatives could support real time release testing.

The ICH in its pharmaceutical development Q8 (R2) document reports a few examples. For instance, disintegration that could serve as a surrogate for dissolution for fast-disintegrating solid forms with highly soluble drug substances. Another example is the unit dose uniformity performed in-process (e.g., using weight variation coupled with near infrared (NIR) assay) can enable real time release testing and provide an increased level of quality assurance compared to the traditional end product testing. From this point of view by assuring continuous quality, real time release testing could supplant the time intensive end product testing.

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).

When establishing and submitting a control strategy, companies are not limited on the type of approach they have to take. For example a control step could use an end product testing approach and another could use real time release testing. What is required though is to explain and justify the selection of one strategy over another.

As with the design space, also the control strategy should be honed throughout the lifecycle of the product, as new data are collected and analysed and more knowledge is obtained. With this

goal in mind continuous process verification tools should be employed. Continuous process verification is an approach to process validation that includes the continuous monitoring and evaluation of manufacturing process performance (ICH, 2009). Examples of this process verification made by the ICH are for instance trend analysis of a manufacturing process as additional experience is gained during routine manufacture. Continuous process verification can utilize in-line, on-line or at-line monitoring or controls to evaluate process performance, and can enhance the evaluation of the manufacturing process if it provides substantially more information on process variability and control. 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 in turn fosters an increased confidence in the applied control strategy and the process design space, since they are continuously controlled, improved and verified.

1.3 Benefits of QbD Implementation

The QbD initiative provides an enhanced approach to pharmaceutical development and manufacturing, based on scientific and engineering principles for assessing and mitigating risks of production shortcomings, with the underlying goal of enhancing the quality of the drug products. The goal of pharmaceutical development with QbD is to achieve a near complete understanding of the scientific aspects of the process. The level of actual knowledge achieved is based on well-designed drug formulation and manufacturing efficacy and efficiency. In pharmaceutical manufacturing, the goal of QbD is to provide systems able to assure quality in real time and to identify and address disturbances entering the process. This endeavour is certainly simplified by using appropriate experimental design (DOE) methods and multivariate statistical analysis techniques to deepen the process knowledge.

In Table 1.1, a comparison between the QbT and the QbD approaches. The increased degree of importance of data and process knowledge is evident in the shift of the focus from an empirical, experience-based one, to a science-based one, where the ability to infer relations, in absence of the true mechanistic models, in crucial to the compliance.

This QbD approach has a monetary value not directly derived from the regulatory relief that comes from the design space and control strategy, but also from the opportunity for an early optimization of the manufactured drug product.

Table 1.1. *Comparison between QbT-based and QbD-based approaches to the pharmaceutical development and manufacturing, adapted from (ICH, 2009)*

Aspect	QbT-based 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
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	− Process analytical 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

In Fig. 1.2 the trend of total revenues of a total product are reported, from the discovery to the expiration of the patent. The solid line represents a product developed with the QbT paradigm, while the dashed line represents a product developed with the QbD best practices. In the pre-launch phase a lot of monetary investment have to be made in order to finance the necessary testing phase that usually lasts around ten years after which the product is launched and the sales start to increase revenue.

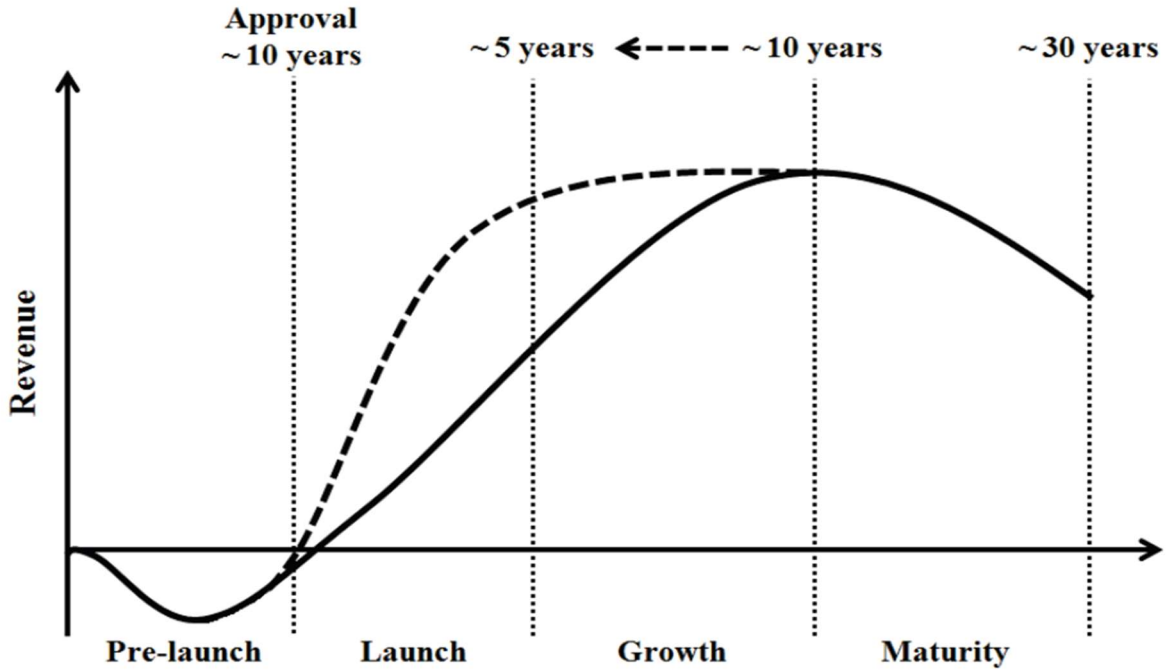

Figure 1.2: *Comparison of revenue trends for a drug product during its lifetime if a QbT (solid line) or a QbD (dashed line) approach were used for development and manufacturing (IBM Business Consulting Services, 2005).*

Sales continues to rise afterwards and after around ten other years a peak of revenue is achieved. After that sales could decline or continue to be relevant. More often than not products are set to launch well before the manufacturing process is optimized. IBM estimated that improving a new product and process development with science and risk based analysis and continuous improvement of the process, using strictly speaking a QbD-approach, could help reducing the period to launch to peak by as much as five years. This gives the product an enormous amount of added value. 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 (IBM Business Consulting Services, 2005).

Another useful guideline to implement and obtain a QbD approach is the one described in ICH Q10 “Pharmaceutical Quality System” (PQS). The PQS is a model for an effective quality management system for the pharmaceutical industry that is based on International Standards Organisation (ISO) quality concepts, includes applicable Good Manufacturing Practice (GMP) regulations and complements ICH Q8 “Pharmaceutical Development” (ICH, 2008). ICH Q10 demonstrates industry and regulatory authorities’ support of an effective pharmaceutical quality system to enhance the quality and availability of medicines around the world in the interest of public health. Implementation of ICH Q10 should facilitate innovation and continual improvement and strengthen the link between pharmaceutical development and manufacturing activities.

Table 1.2*: potential opportunity to enhance science and risk based regulatory approaches, from ICH Q10 (ICH, 2008)*

Scenario	Potential Opportunity
1. Comply with Good Manufacturing Practices	Compliance – status quo
2. Demonstrate effective pharmaceutical quality system, including effective use of quality risk-management principles (e.g., ICH Q9 and ICH Q10).	Opportunity to: increase use of risk based approaches for regulatory inspections.
3. Demonstrate product and process understanding, including effective use of quality risk-management principles (e.g., ICH Q8 and ICH Q9).	Opportunity to: facilitate science based pharmaceutical quality assessment; enable innovative approaches to process validation; establish real-time release mechanisms.
4. Demonstrate effective pharmaceutical quality system and product and process understanding, including the use of quality risk management principles (e.g., ICH Q8, ICH Q9 and ICH Q10).	Opportunity to: increase use of risk based approaches for regulatory inspections; facilitate science based pharmaceutical quality assessment; optimise science and risk based post-approval change processes to maximise benefits from innovation and continual improvement; enable innovative approaches to process validation; establish real-time release mechanisms.

In Table 1.2, a prospect of the potential opportunity to enhance science and risk based regulatory approaches obtainable by implementing the ICH Q10 Pharmaceutical Quality System. The implementation of Q10 means the achievement of three main objectives, achieve product realisation, establish and maintain a state of control, and facilitate continual improvement. Product realization is defined as "Achievement of a product with the quality attributes appropriate to meet the needs of patients, health care professionals, and regulatory authorities (including compliance with marketing authorisation) and internal customers' requirements." (ICH, 2008). This is done through the implementation of a QbD logic and the use of a well-maintained and derived design space. Establishing and maintaining a state of control means using quality risk management and an optimal control strategy providing assurance of continued suitability and capability of processes. Facilitate continual improvement means keeping in place continuous process verification, among other systems, to identify and implement appropriate product and process improvements thereby increasing the ability to fulfil quality needs consistently.

It is also necessary to stress the importance of a model-based approach to the establishment of the design space. Model-based design spaces maintain all the benefits of submitting a design space, such as more freedom of selection of the process parameter while providing some

powerful tools to conduct different analysis. A data driven and model driven design space can be used to limit the portion of the process knowledge space in which experiments to determine the optimal operating conditions are needed, greatly simplifying the ensuing experimental design. Another use in which the design space of the process can be helpful is online fault detection. The design space can be used to develop an accurate detection chart as it embodies process understanding and can be easily used to ensure that it is working as anticipated to deliver product quality attributes as predicted by the design space. This monitoring could include trend analysis of the manufacturing process as additional experience is gained during routine manufacture. Even the perceived deficiency of the model based approach, that is the need of periodic maintenance, can be seen as an opportunity to deepen the knowledge of the process that is, as mentioned above, crucial in the compliance to the QbD initiative.

Chapter 2

Mathematical Background

In this chapter, the basis of the main mathematical techniques employed in this work are presented, starting from the description of two dimensional reduction techniques, *Principal Components Analysis* (PCA) and *Partial Least Squares* (PLS) and then continuing with an overview on Bayesian statistics.

2.1 Principal Component Analysis

Let let $\mathbf{X}[I \times J]$ be a historical dataset composed of I rows and J columns. Rows correspond to samples, while columns correspond to variables. PCA is a statistical tool that allows describing the historical dataset by means of few variables, called *latent* variables, that describe the maximum multidimensional variance of the historical dataset. The notable advantage is that, in the presence of strong collinarity in the original dataset, the number of latent variables is much smaller than the number of original input variables.
PCA employs an eigenvector decomposition of the covariance matrix of the data.
The covariance of $\mathbf{X}$ is equal to:

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{I-\mathbf{1}}. \tag{2.1}$$

Eigenvector of the covariance matrix are calculated according to:

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_n = \boldsymbol{\lambda}_n\mathbf{p}_n \tag{2.2}$$

For each eigenvector $\mathbf{p}$ the corresponding vector $\mathbf{t}$ is found as:

$$\mathbf{X}\mathbf{p}_n = \mathbf{t}_n\,. \tag{2.3}$$

And can be used to write the original $\mathbf{X}$ matrix as the sum of the outer products of the $\mathbf{tp}$ pairs:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \mathbf{t}_3\mathbf{p}_3^{\mathrm{T}} + \cdots + \mathbf{t}_N\mathbf{p}_N^{\mathrm{T}} + \mathbf{E}\,, \tag{2.4}$$

where N is a number no greater than the smaller dimension of $\mathbf{X}$, i.e: $N \leq \min(I,J)$; $\mathbf{t}_n$ are the model scores; $\mathbf{p}_n$ are the model loadings; $\mathbf{E}$ is the matrix of the residuals. The scores are orthogonal, and convey information on how samples relate to each other, while the loadings are orthonormal if $\mathbf{X}$ is autoscaled, otherwise they are orthogonal, and convey information on how variables are correlated with each other.

The model scores $\mathbf{t}_n$ are collected in the score matrix $\mathbf{T}$, while the model loadings are collected in the loading matrix $\mathbf{P}$.

The decomposition of the historical dataset can be reformulated as the following optimisation problem:

$$\begin{aligned} &\max\ (\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}) \\ &\text{s.t.}\ \ \mathbf{P}^{\mathrm{T}}\mathbf{P} = 1 \end{aligned} \tag{2.5}$$

By applying the method of Lagrange multipliers, Equation (2.5) can be rewritten as:

$$\mathrm{L}(\mathbf{P}) = (\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}\ - \lambda(\mathbf{P}^{\mathrm{T}}\mathbf{P} - 1))\,. \tag{2.6}$$

According to Lagrange multipliers method Equation (2.6) has to be derived by $\mathbf{P}$ and λ and the resulting equations have to be set to zero**:**

$$\frac{\mathrm{dL}(\mathbf{P})}{\mathrm{d}\mathbf{P}}\left(\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}\ - \lambda(\mathbf{P}^{\mathrm{T}}\mathbf{P} - 1)\right) = \mathbf{P}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}) + 2\lambda\mathbf{P}^{\mathrm{T}} = 0, \tag{2.7}$$

$$\frac{\mathrm{dL}(\mathbf{P})}{\mathrm{d}\lambda}\left(\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}\ - \lambda(\mathbf{P}^{\mathrm{T}}\mathbf{P} - 1)\right) = (\mathbf{P}^{\mathrm{T}}\mathbf{P} - 1) = 0. \tag{2.8}$$

Knowing that cov($\mathbf{X}$) is symmetric and taking into account Equation (2.7) and Equation (2.8) it is possible to write:

$$\mathbf{P}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}) + 2\lambda\mathbf{P}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P} - \lambda\mathbf{P} = 0, \tag{2.9}$$

which is similar to the eigenvalue problem of Equation (2.2).

With the results of Equation (2.6) in mind it is possible to write Equation (2.4) as

$$\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P} = \mathbf{P}^{\mathrm{T}}\lambda\mathbf{P} = \lambda\mathbf{P}^{\mathrm{T}}\mathbf{P} = \lambda. \tag{2.10}$$

Considering the maximisation problem of equation (2.5), λ is the greatest eigenvalue for the eigenvalue problem of equation (2.2) (Trefethen and Bau, 1997).

It can be observed that the terms $\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}, \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} \ldots \mathbf{t}_N\mathbf{p}_N^{\mathrm{T}}$ of Equation (2.4) can be interpreted as the first, second and up to the N-th eigenvector ordered by the magnitude of the associated eigenvalue. According to the rule of thumb that an eigenvalue explains a number of variables roughly equal to it's value it is possible to say they are consequently also ordered by the amount of variance explained.

Following this reasoning Equation (2.3) can be written as a sum of *A*, accepted principal components and (*n-A)* discarded principal components without losing representativeness.

These *n*-A principal components can be stored in the matrix $\mathbf{E}$, which will now store the unexplained variability in addition to residuals

$$\mathbf{X} = \sum_1^A \mathbf{t}_n\mathbf{p}_n^{\mathrm{T}} + \sum_{A+1}^N \mathbf{t}_n\mathbf{p}_n^{\mathrm{T}} + \mathbf{E} = \sum_1^A \mathbf{t}_n\mathbf{p}_n^{\mathrm{T}} + \hat{\mathbf{E}}. \tag{2.11}$$

Methods are avaible to assess the performance of the PCA model. These methods can be divided in observation, variable and model diagnostics. (Eriksson *et al*., 2001)

Observation diagnostics are used to recognise outliers. Outliers are observations which appear to be inconsistent with the rest of the data, relative to an assumed model (Everitt and Skrondal,

2010), hence the importance of their removal when data analysis is carried out. A diagnostic used to identify these inconsistent observations is Hotelling's T^2 statistic (MacGregor and Kourti, 1995), a mulivariate generalisation of student's t-test that checks the adherence to multivariate normality.

An observation T_i^2 statistic is given by :

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{ia}}{s_{ta}^2}, \tag{2.12}$$

where $s_{ta}^2$is the score variance with respect to principal component *a* and A is the number of principal components used for the PCA.

A sample is deemed to be an outlier if T_n^2 is greater than a critical value T_{crit}^2 expressed by (Wikström, *et al*., 1998)

$$T_{crit}^2 = \frac{A(I-1)}{I(I-A)} F(A, I-A, \alpha) \tag{2.13}$$

where A is the number of selected Principal Components, *I* is the number of observations in the model training set and F(A, *I*-A, α) is the function that computes the value that should exceed α% of the samples from an F distribution with A degrees of freedom in the numerator and (*i*-A) degrees of freedom in the denominator. The threshold value α is selected based on the confidence that is needed for the critical value, usually 5% is used giving a confidence of 95%.

Variables diagnostics can be used, for example, to assess the explained variation of a variable with respect to the selected principal components. This indicator ranges from zero (no explanation) to one (complete explanation).

One of the ways to compute this indicator is by analysing the $\hat{\mathbf{E}}$ matrix column-wise sum of squared elements:

$$\hat{\mathbf{E}}_{aj} = \sum_{i=1}^{I} \mathbf{e}_{ija}^2 = \sum_{i=1}^{I} (\mathbf{X}_{ij} - \mathbf{t}_{ia}\mathbf{p}_{ja})^2 \tag{2.14}$$

where *j* stands for one of the *J* variables of the **X** dataset, *a* stands for the *a*-th principal component up to the selected A of equation (2.11) and *I* is the total number of repeated measurements in the **X** dataset.

The matrix from Equation (2.14) can be related to the sum of squares of the original variable to obtain a normalised portion of unexplained variance for that variable (Varmuza and Filzmoser, 2009). Subtracting that from one gives ${}_aR_j^2$, which is the above-mentioned indicator of variation for *a* principal components,

$${}_aR_j^2 = 1 - \frac{\sum_{i=1}^{I} \mathrm{e}_{ija}^2}{\sum_{i=1}^{I} \mathbf{x}_{ij}}. \tag{2.15}$$

Given a fixed number of principal components, it is desirable to have the maximum possible. This is not always the case as it can happen that a variable may have a very low explained variance while all the others variables may have satisfactory values. If that happens it could be

appropriate to add another principal component to the analysis, if in accordance to the model diagnostic.

Model diagnostics can be used for a statistical based estimation of the optimum number of principal components. In this analysis is essential to evaluate the "goodness of fit" (R^2) statistic along with the "goodness of prediction" (Q^2) statistic. To be done in a statistically meaningful way data have to be resampled, for example by bootstrapping or by cross validation (Varmuza and Filzmoser, 2009).

Cross validation is one of the most used techniques and consists in splitting the original data into group into subgroups and then proceeding with a leave-one-out criterion.

The subgroup left out will be the validation group to the training one, comprised of all the other groups.

PCA is carried out on the training group, loadings and scores are calculated and residuals are computed. From this, a goodness of fit statistic can be extracted calculating an indicator similar to that of Equation (2.15) but considering all *J* variables instead of just one at a time. As for the variable diagnostic, values of this goodness of fit statistic range from zero to one and are ever increasing as new principal components are added to the analysis.

It is now clear that, as mentioned above, a goodness of prediction statistic is also necessary; this statistic is computed using the validation dataset.

The validation group scores are calculated according to (2.3) using the training group loadings:

$$\mathbf{X}_{val}\mathbf{P}_{train} = \mathbf{T}_{val}, \tag{2.16}$$

where the subscripts "val" and "train" refer to the validation and training group respectively.

As for the training set residuals are computed as

$$\hat{\mathbf{E}}_{val} = \mathbf{X}_{val} - \mathbf{T}_{val}\mathbf{P}_{train} \tag{2.17}$$

in addition, the goodness of prediction statistic is calculated as:

$$Q_a^2 = 1 - \frac{||\,_a\hat{\mathbf{E}}_{val}||^2}{||\mathbf{X}||^2} \tag{2.18}$$

where the *a* stands for the number of principal components used for this projection as in equation (2.14) and (2.15).

Criteria for selecting the number of optimal principal components using these statistics are manifold. For example, Eriksson *et al*. (2001) proposed a slightly different criterion for Q2 using the residual sum of squares of the previous dimension,

$$Q_a^2 = 1 - \frac{||\,_a\hat{\mathbf{E}}_{val}||^2}{||_{a-1}\hat{\mathbf{E}}||^2}, \tag{2.19}$$

instead of $||\mathbf{X}||^2$ in equation (2.18), stopping when the added prediction power is lower than a critical limit.

2.2 Partial Least Squares Regression (PLS)

Partial Least Squares Regression (PLS) is a multivariate regression technique that introduces a reduced set of variables, called latent variables, by maximising the covariance between a input and output dataset. PLS can be used for dimensional reduction, regression analysis or classification problems among other uses (Rosipal and Krämer, 2006).

Let $\mathbf{Y}[I \times K]$ be a mean-centered historical dataset composed of the output generated by the input in $\mathbf{X}$. The matrix $\mathbf{Y}$ is composed of I rows and K columns. Using the PLS decomposition it is possible to write $\mathbf{X}$ and $\mathbf{Y}$ as:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \end{aligned} \tag{2.20}$$

Where $\mathbf{T}$ and $\mathbf{U}$ are $[I \times A]$ matrices of the A extracted score vectors. The matrices $\mathbf{P}$ $[J \times A]$and $\mathbf{Q}$ $[K \times A]$ are composed of the loadings and $\mathbf{E}$ $[I \times J]$and $\mathbf{F}$ $[I \times K]$ are the residuals.

The covariance between $\mathbf{X}$ and $\mathbf{Y}$ is:

$$\mathrm{cov}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{Y}}{I-1}. \tag{2.21}$$

The decomposition of Equation (2.20) is obtained through two weight vectors, $\mathbf{w}$ and $\mathbf{c}$, one for $\mathbf{X}$ and one for $\mathbf{Y}$. Weights $\mathbf{w}$ and $\mathbf{c}$ are $[I \times 1]$ vectors and every latent variable will have different weight vectors adding up to $[I \times A]$ matrices $\mathbf{W}^*$ and $\mathbf{C}^*$ after the A latent variables have been selected.

These weights are subject to the following optimisation problem,

$$\begin{aligned} &\max\ (\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}) \\ &\text{s.t.}\ \ \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1 \end{aligned} \tag{2.22}$$

and

$$\begin{aligned} &\max\ (\mathbf{c}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{c}) \\ &\text{s.t.}\ \ \mathbf{c}^{\mathrm{T}}\mathbf{c} = 1. \end{aligned} \tag{2.23}$$

Equations (2.22) and (2.23) are similar to the optimisation problem of Equation (2.5), and the methods of Equations (2.6)-(2.10) can be applied. As derived in Equation (2.10) $\mathbf{w}$ and $\mathbf{c}$ are the eigenvectors of the greatest eigenvalue for the matrices above.

Once computed the weight vectors can be used to obtain the score vectors $\mathbf{t}$ and $\mathbf{u}$ as the linear coefficient of:

$$\mathbf{t}_{\boldsymbol{a}} = \mathbf{X}\mathbf{w}_{\boldsymbol{a}}, \tag{2.24}$$

$$\mathbf{u}_{\boldsymbol{a}} = \mathbf{Y}\mathbf{c}_{\boldsymbol{a}}, \tag{2.25}$$

where a stands for the a-th component of the PLS and, as such, as the a-th step of the algorithm.

To obtain the full latent structure it is necessary to proceed iteratively on the deflated data matrix. The process of Deflation means subtracting the projection of **X** and **Y** to the **X** and **Y** matrix to be used in the next iteration.

The deflation step is one of the critical features of the PLS methodology since it allows to remove the projection of the data matrix in the direction of the selected component, and to maximise the information gained from the method.

Consider this inequality, proved among others, by Rao (Rao, 1979):

$$s_i^2(\mathbf{A}-\mathbf{B}) \geq s_{i+k}^2(\mathbf{A}) = a_{i+k}^2. \tag{2.26}$$

Where $s_i^2(\mathbf{A})$ is the i-th singular value of matrix **A** and B is a matrix of rank *k.*

When this is applied to the PLS it becomes (Hoskuldsson, 1988):

$$s_1^2\left(\mathbf{X}_{\mathrm{i+1}}^{\mathrm{T}}\mathbf{Y}\right) = s^2\left(\mathbf{X}_i^{\mathrm{T}}\mathbf{Y} - \mathbf{p}_i\mathbf{t}_i^{\mathrm{T}}\mathbf{Y}\right) \geq s_2^2\left(\mathbf{X}_i^{\mathrm{T}}\mathbf{Y}\right) \tag{2.27}$$

This means that the largest singular value at step *i*+1, after deflation, is greater than the second largest singular value at step *i*. Equation (2.27) proves the previous assumption of that deflation maximises the information gained.

As mentioned above deflation is carried out by:

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \frac{\mathbf{t}_{a-1}\mathbf{t}_{a-1}^{\mathrm{T}}}{\mathbf{t}_{a-1}^{\mathrm{T}}\mathbf{t}_{a-1}}\mathbf{X}_{a-1}, \tag{2.28}$$

$$\mathbf{Y}_a = \mathbf{Y}_{a-1} - \frac{\mathbf{u}_{a-1}\mathbf{u}_{a-1}^{\mathrm{T}}}{\mathbf{u}_{a-1}^{\mathrm{T}}\mathbf{u}_{a-1}}\mathbf{Y}_{a-1}, \tag{2.29}$$

Where *a* is the *a*-th step while step 0 is the original dataset for **X** and **Y** and vector of zeros of the appropriate dimension for **t** and **c**.

Loadings for the reconstruction of the matrices are calculated as:

$$\mathbf{p}_a = \frac{\mathbf{t}_a^{\mathrm{T}}\boldsymbol{X}_a}{\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a}, \tag{2.30}$$

and

$$\mathbf{q}_a = \frac{\mathbf{u}_a^{\mathrm{T}}\mathbf{Y}_a}{\mathbf{u}_a^{\mathrm{T}}\mathbf{u}_a}. \tag{2.31}$$

Both scores and loading of each step are concatenated as columns to create the **P**, **Q**, **T** and **U** matrices of Equations (2.20).

As with PCA, there are some diagnostics useful for troubleshooting the model and the analysis. The actual diagnostics are in many way similar to those of the PCA, with only minor variations necessary to adress the dimensionality issues between the two methods.

Hotelling's T^2 for PLS methods is equivalent to the one applied to PCA, since it tests the adherence to multivariate normality. Other diagnostics used to troubleshoot Observations are the scores plot, which in PLS develop a further level of complication since cross variable scores have to be inspected too.

Variable diagnostics are easily adapted from PCA. The analysis of column-wise sum of residual (2.14) and R^2 statistics (2.15) can be also computed for PLS. Since PLS has one more residuals matrix, **F**, these statistics should be computed for this matrix too evaluating the degree of explaination that the method gives to the *K* output variables.
In the context of PLS model diagnostics can, and should, be used. Following the trend set by the other diagnostics type only minimal modifications have to be applied to use the PCA model diagnostic. While cross validation stands, it is necessary to use a validation dataset to investigate the goodness of the model for both matrices. The modification needed for the goodness of prediction statistic is that Equation (2.18) and (2.19) need to be applied to both matrices, **Y** and **X**. In the case of PLS regression instead, the goodness of fit statistics of Equation (2.18) and (2.19) have to be applied to the results of the regression of matrix **X** on **Y** and thus on the prediction of the output matrix **Y** instead of the imput matrix **X.**

2.3 Bayesian statistical methods

Bayes rule is a theorem in statistical science that originates from the work of Pierre-Simon Laplace on a posthumous work by Thomas Bayes (Bayes, 1763).
This theorem is formulated as:

$$\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}. \tag{2.32}$$

It states that the probability of an observation **y** given the that the parameters are **θ** is equal to the probability of the parameter vector being **θ** having observed **y**, multiplied by the probability of the parameters being **θ**, divided by the marginal distribution of **y**. The marginal distribution, also called the prior predictive distribution, can also be written as:

$$\mathrm{P}(\mathbf{y}) = \int \mathrm{P}(\mathbf{y}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{2.33}$$

for continuos quantities, or

$$\mathrm{P}(\mathbf{y}) = \sum_{\theta}^{\theta} \mathrm{P}(\mathbf{y}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta}) \tag{2.34}$$

where the summation sign stands for every possible value of **θ**, for discrete values.
Bayes theorem is a propriety of conditional probability, but it also represents a tool for updating and revising the probability of an occurrence based on new evidence (Everitt and Skrondal, 2010). This new evidence is fed in Equation (1.30) by the prior distribution of the parameters **θ** and different degree of knowledge of the priors give rise to different distribution for P(**y**| **θ**). The idea at the core of Bayesian statistics is that the uncertainty of the investigator about an inferred quantity is expressed by using a probability distribution (Jaynes, 2003), and the region under the *x* percent of the probability function is called a *x* credibility interval (Gelman, et al., 2013).

One of the most used tools of statistics is linear regression and the simplest and most widely used linear regression tecnique is the normal linear model (Gelman, *et al*., 2013).
The normal linear model represent **Y** as normally distributed with mean composed as a linear combination of the values of **X**, with **B** as the combination coefficients:

$$\mathrm{E}(\mathbf{y_i}|\mathbf{X},\mathbf{B}) = \begin{matrix}\mathbf{B}_{11}\\ \vdots\\ \mathbf{B}_{1N}\end{matrix}\mathbf{x}_1 + \begin{matrix}\mathbf{B}_{21}\\ \vdots\\ \mathbf{B}_{2N}\end{matrix}\mathbf{x}_2 + \begin{matrix}\mathbf{B}_{31}\\ \vdots\\ \mathbf{B}_{3N}\end{matrix}\mathbf{x}_3 + \cdots + \begin{matrix}\mathbf{B}_{J1}\\ \vdots\\ \mathbf{B}_{JN}\end{matrix}\mathbf{x}_J. \tag{2.35}$$

Equation (2.33) represents a multivariate case in which *N* is different from one and as such the distribution of **Y** is multivariate normal instead of normal.
This can be written as:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E} \tag{2.36}$$

where **E** is the matrix of the residuals.
From a Bayesian perspective, this model should also include a distribution for **X** given a parameter vector $\boldsymbol{\xi}$, since **X** is also part of the dataset.
It follows that the regression model should incorporate a joint probability:

$$\mathrm{P}(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta},\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\theta},\boldsymbol{\xi}). \tag{2.37}$$

A defining characteristic of regression from a Bayesian point of view is the consideration of prior independence of the parameter vector $\boldsymbol{\theta}$ that determines $\mathrm{P}(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})$ and the parameter vector $\boldsymbol{\xi}$ that determines $\mathrm{P}(\boldsymbol{\xi}|\mathbf{X})$. From this condition follows the posterior distribution:

$$\mathrm{P}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{X},\mathbf{Y}) = \mathrm{P}(\boldsymbol{\xi}|\mathbf{X})\mathrm{P}(\boldsymbol{\theta}|\mathbf{X},\mathbf{Y}). \tag{2.38}$$

It is possible to study only the second part of the equation, given that usually data comprising **X** are observed and directly chosen and as such their probability $\mathrm{P}(\mathbf{X})$ is known and there is no parameter vector $\boldsymbol{\xi}$.
Applying Bayes Rule to equation (2.36) we have:

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{X},\mathbf{Y}) \propto \mathrm{P}(\mathbf{Y}|\boldsymbol{\theta},\mathbf{X})\mathrm{P}(\boldsymbol{\theta}). \tag{2.39}$$

Let $\boldsymbol{\theta}$ be $[\mathbf{B}, \sigma^2]$ with σ^2 equal to the variance of unit weight of **Y** and its distribution as normal. This identifies a subset of the normal linear model called the ordinary linear model, characterised by homoscedasticity and equal variance throughout **Y**.
In short, for this model we can write:

$$(\mathbf{Y}|\mathbf{B},\boldsymbol{\sigma}^{\mathbf{2}},\mathbf{X}) \sim \mathrm{N}(\mathbf{BX},\boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I}^{\mathrm{N}}). \tag{2.40}$$

At this point, if **X** is full rank, the analysis of this distribution still yields the same maximum likelihood estimate as the classical frequentist approach, taken by maximising the logarithm of the exponential form of the normal distribution. This maximum likelihood estimate are:

$$\widehat{\mathbf{B}}_{MLE} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\,, \tag{2.41}$$

and

$$\widehat{\boldsymbol{\sigma}}^2_{MLE} = \frac{\mathbf{1}}{N}\mathbf{e}^{\mathrm{T}}\mathbf{e}\,. \tag{2.42}$$

Where*N* is the number of observations and **e** is the model error with the **B** parameters from equation (2.40) (Lynch, 2007).

In the full Bayesian approach at this point a definition of a prior for equation (2.37) is necessary. A standard uninformative prior for regression is the uniform prior, which assigns an equal probability to every possible value of the beta parameters. This is not always the case.

From these information it is possible to compute a posterior predictive distribution. The goal of Bayesian regression is not simply the estimate of the values of the parameters and their deviation, but to infer on unknown quantities, thus the target distribution is the distribution of the response variable *y* given a new data point taken from the same data pool as **X** but not included in the derivation of the parameters,

$$\mathrm{P}(\mathbf{y}|\mathbf{x}, \mathbf{B}, \boldsymbol{\sigma}^{\mathbf{2}}). \tag{2.43}$$

This probability is obtainable through calculations only in limited cases due to the difficulty of the required integrations and it has to be computed through a simulation. The most used simulation is the Markov Chain Monte Carlo method. The Markov Chain Monte Carlo (MCMC) is a method for constructing and sampling from an arbitrary posterior distribution $\boldsymbol{\theta}$ adjusting the draws to better represent the posterior distribution $\mathrm{P}(\boldsymbol{\theta}|\mathbf{Y})$ (Gelman, *et al.*, 2013). The MCMC method is based on the concept of Monte Carlo process, a process composed of random draws and that of Markov chains. A Markov chain is a stocastic process for which the distribution of the parameters $\boldsymbol{\theta}^{\mathbf{t}}$ at iteration *t*, given the distribution of all the other parameters $\boldsymbol{\theta}$, depends only on the value at $\boldsymbol{\theta}^{\mathbf{t}-1}$.

Applied to the MCMC method this means that the transition distribution $\mathrm{T_t}$,

$$\mathrm{T}_t(\boldsymbol{\theta}_t|\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{t-1}) = \mathrm{T}_t(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}), \tag{2.44}$$

that is the distribution that governs the direction of the Markov chain evolution, must have a specific stationary distribution, which is the distribution the chain tends to with infinite draws. For any specific distribution $\mathrm{P}(\boldsymbol{\theta}|\mathbf{Y})$ or un-normalized density $\mathrm{Q}(\boldsymbol{\theta}|\mathbf{Y})$, it possible to build some Markov chains with the desired stationary distribution. Within the context of successive refinement of the approximation, the sampling algorithm has a crucial spot. For these reasons the research on the topic of optimal sampling algorithm is still open.

One of the most used algorithm is the Gibbs Samples, also called alternating conditional sampling. In this algorithm the parameter vector $\boldsymbol{\theta}$ is divided into *d* sub-vectors such as:

$$\boldsymbol{\theta}_{gibbs} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_d). \tag{2.45}$$

At each step *t* then the Gibbs sampler cycles through the *d* parameters sub-vectors and the new sub-vector values are sampled as conditional on all the others sub-vectors. Each iteration *t* has

then *d* steps, at every step a different ordering of the *d* sub-vectors is selected and the new value of the proposed posterior probability is computed.
The new values are conditional on the current values of the others:

$$\mathrm{P}(\boldsymbol{\theta}_j^t|\boldsymbol{\theta}_1^t,\dots,\boldsymbol{\theta}_{j-1}^t,\boldsymbol{\theta}_{j+1}^{t-1},\dots,\boldsymbol{\theta}_d^{t-1},\mathbf{Y}), \tag{2.46}$$

thus the sub-vector *j* at iteration *t* is calculated as conditional to the value at time *t* of the previous sub-vectors and to the value of the following ones at time t −1. This algorithm is the simplest of the MCMC sampling and is thus mainly used when it is possible to direclty sample the conditional posterior distribution, such as in conditionally conjugate model.
Other examples of sampling algorithms are the Metropolis and its generaliation the Metropolis-Hastings. The Metropolis algorithm is a modified random walk model with an acceptance criterion to speed up convergence to the desired stationary distribution. The random walk is a stochastic process in which a particle *r* starts at a position $r = r^0$ and, at each step *t*, has the possibility *p* of moving up and (1−*p*) of moving down; *p* is called the jumping distribution.
The Metropolis algorithm starts with a draw for which $\mathrm{P}(\boldsymbol{\theta}^0|\mathbf{Y})$ is greater than zero, taken form a given starting distribution. Then for *t* steps a proposal $\boldsymbol{\theta}^*$is generated from a symmetric jumping distribution:

$$\mathbf{J}_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}) = \mathbf{J}_t(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b) = \mathbf{J}_t(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a), \tag{2.47}$$

for every *a*, *b* and *t*.
The algorithm then calculates the ratio of densities,

$$\text{ratio} = \frac{\mathbf{P}(\boldsymbol{\theta}^*|\mathbf{Y})}{\mathbf{P}(\boldsymbol{\theta}^{t-1}|\mathbf{Y})} \tag{2.48}$$

and subsequently generates an uniform random number to simulate the random walk, using it as acceptance criterion for the proposed $\boldsymbol{\theta}^*$

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(\text{ratio},1) \\ \boldsymbol{\theta}^{t-1} & \text{otherwise} \end{cases} \tag{2.49}$$

The Metropolis Hastings algorithm has some fundamental differences, first the jumping distribution of Equation (2.47) does not need to be symmetric, and as such the second and third therm are not present, second the ratio of Equation (2.48) is calculated as a ratio of ratios:

$$\text{ratio}_{MH} = \frac{\frac{\mathbf{P}(\boldsymbol{\theta}^*|\mathbf{Y})}{\mathbf{J}_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}}{\frac{\mathbf{P}(\boldsymbol{\theta}^{t-1}|\mathbf{Y})}{\mathbf{J}_t(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}} \tag{2.50}$$

As this is always defined even for asymmetric jump distributions.
It is possible to prove that the Gibbs sampler is a special case of the Metropolis algorithm with a particular jumping distribution and ratio equal to 1.
The Metropolis and Metropolis Hastings algorithm are more flexible than the Gibbs sampler and can be used even in cases where the conditional posterior distribution is not in the same distribution family as the prior distribution.

Many other sampling algorithm have been proposed, each used in different cases and with their own pros and cons, for example Slice Sampling (Gilks and Wild, 1992), Reversible Jump Metropolis (Green, 1995) or preconditioned Crank-Nicolson algorithm (Cotter, *et al.,* 2013).

Chapter 3

Design space determination: a Bayesian approach

In this chapter, a brief introduction to the link between mathematical modelling and the Quality by design paradigm discussed in the previous chapter is given. Moreover, a recently developed technique that can be used to identify the design space of a new pharmaceutical product in a Bayesian framework is briefly reviewed.

3.1 Quality by Design and modelling

The new paradigm of Quality by Design for pharmaceutical products and process design can be seen, from the chemical engineering point of view, as the application of Process System Engineering (PSE) to manufacture and development of pharmaceutical products (García-Muñoz and Oksanen, 2010). In line with the approach suggested by the regulatory agencies, it is crucial to analyse CPPs and raw materials properties and to link them, through a scientific analysis, to the CQAs of the products. Models can be derived either from first principles - if the process knowledge is mature enough - through data driven methods or by a combination of the two. Since their origin is found in the application of PAT frameworks (see chapter §1), models are part of the framework too, and as such are PAT tools themselves.

The use of models is crucial in the effort to shift the product design paradigm from an experience-based one towards a science-based one, integrating it in the design of the production process. This fact is fully understood by the regulatory agencies, which encouraged model-based support for QbD implementation. The International Conference on Harmonisation in its 2011 guide for ICH Q8/Q9/Q10 implementation divided models into different categories, based on their contribution to assuring the quality of the end product.

Regarding the assurance of quality, a model can be divided into low, medium or high impact. Low-impact models are typically those used to support product and/or process development (e.g., formulation optimization). Medium-impact models are 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 instead are those whose prediction is a

significant indicator of the quality of the product (e.g., a chemometric model for product assay) (ICH, 2011).
Process System Engineering provides many tools for model development and application, which have already been developed and have undergone decades of improvements and evolution in other areas of industrial manufacturing. Provided that some challenges, unique to the field, have been studied and incorporated in the PSE, pharmaceutical sector stand to gain remarkably from PSE adoption.

3.2 Design Space identification

Modelling the multivariate relations between the raw materials properties, the CPPs and the CQAs of the products could seem straightforward, but some issues arise. Scientists were at first tempted to employ some well-known metrics to calculate the design space, for example using a combination of proven acceptable range (PAR). PAR are defined as "a characterised range of a process parameter for which operation within this range, while keeping other parameters constant, will result in producing a material meeting relevant quality criteria". As the definition suggests PAR is a univariate approach in its core and as such is too simplistic in treating the multivariate nature of an industrial process. In fact, during a subsequent review of the Q8 document, ICH clarified that a combination of proven acceptable ranges does not constitute a design space (ICH, 2009). Scientists then changed the approach to one well known and easily implemented; the overlapping mean responses (OMR) approach. The OMR approach is a classical response surface methodology, which primarily focuses on inference about mean response surfaces. This method also had the benefit of having many point-and-click oriented statistical packages such as such as Design Expert™ or JMP equipped with integrated functionalities that made the construction of an OMR plot relatively easy; furthermore, at the time some studies proposed the OMR as the correct approach (Peterson and Lief, 2010).
The OMR method consists of merely overlapping two or more mean responses chart and to look for "sweet spots" (Anderson and Whitcomb, 1998) where the desired responses are obtained and propose that area as the design space. The OMR can be expressed mathematically as (Peterson *et al.*, 2017):

$$\{\mathbf{x} \in \boldsymbol{\chi} : \mathrm{E}(\mathbf{y}|\mathbf{x}) \in \mathbf{S}\} \tag{3.1}$$

where $\boldsymbol{x}$ is an arbitrary configuration of the process inside the "sweet spot", $\boldsymbol{\chi}$ is the "sweet spot", that is the combination of desired responses, expressed by the expectation E of the responses $\mathbf{y}$ given the data vector $\boldsymbol{x}$, that are inside the desired response interval $\mathbf{S}$.
Another approach that enhances OMR is to use desirability functions. Desirability functions are functions that map the mean response of a single CQA to a scalar value ranging from zero to one, with zero an undesirable outcome and one a wholly desirable, ideal, outcome.

$$\begin{aligned} &\boldsymbol{Y}_i(\mathbf{x}) = d_i(\mathbf{Y}_i); \\ &d_i \in [0,1]. \end{aligned} \tag{3.2}$$

Desirability functions have different forms whether the objective is to maximise, minimise or obtain target desirability. The most used forms are those proposed by Derringer and Suich (Derringer and Suich, 1980). Let U_i, L_i and T_i be the upper, lower and target desired value for the i-th CQA, $d_i(\mathbf{Y}_i)$ is then equal to:

$$d_i(\mathbf{Y}_i) = \begin{cases} 0 \text{ if } \mathbf{Y}_i(\mathbf{x}) > \mathrm{U}_i; \\ \frac{\mathbf{Y}_i(\mathbf{x}) - \mathrm{U}_i}{\mathrm{T}_i - \mathrm{U}_i} \text{ if } \mathrm{T}_i \le \mathbf{Y}_i(\mathbf{x}) \le \mathrm{U}_i; \\ 1 \text{ if } \mathrm{Y}_i(\mathbf{x}) < \mathrm{T}_i\,, \end{cases} \tag{3.3}$$

if the target is to minimize the value of $\mathbf{Y}_i(\mathbf{x})$,

$$d_i(\mathbf{Y}_i) = \begin{cases} 0 \text{ if } \mathbf{Y}_i(\mathbf{x}) < \mathrm{L}_i; \\ \frac{\mathbf{Y}_i(\mathbf{x}) - \mathrm{L}_i}{\mathrm{T}_i - \mathrm{L}_i} \text{ if } \mathrm{T}_i \ge \mathbf{Y}_i(\mathbf{x}) \ge \mathrm{L}_i; \\ 1 \text{ if } \mathrm{Y}_i(\mathbf{x}) > \mathrm{T}_i, \end{cases} \tag{3.4}$$

If the target is to maximize the value of $\mathbf{Y}_i(\mathbf{x})$ and

$$d_i(\mathbf{Y}_i) = \begin{cases} 0 \text{ if } \mathbf{Y}_i(\mathbf{x}) > \mathrm{U}_i; \\ \frac{\mathbf{Y}_i(\mathbf{x}) - \mathrm{L}_i}{\mathrm{T}_i - \mathrm{L}_i} \text{ if } \mathrm{T}_i \ge \mathbf{Y}_i(\mathbf{x}) \ge \mathrm{L}_i; \\ \frac{\mathbf{Y}_i(\mathbf{x}) - \mathrm{U}_i}{\mathrm{T}_i - \mathrm{U}_i} \text{ if } \mathrm{T}_i \le \mathbf{Y}_i(\mathbf{x}) \le \mathrm{U}_i; \\ 0 \text{ if } \mathrm{Y}_i(\mathbf{x}) < \mathrm{L}_i, \end{cases} \tag{3.5}$$

If the response is of "target is best" kind. These responses are then combined, usually as a geometric mean, to obtain the total desirability of the CPP and material attributes configuration.

$$\mathrm{D} = \sqrt[n]{d_1(\mathbf{Y}_1) \times d_2(\mathbf{Y}_2) \times \cdots \times d_n(\mathbf{Y}_n)} \tag{3.6}$$

where n is the total number of CQAs. The total desirability D is then maximised (Derringer and Suich, 1980).

The endorsement from the ICH and the ISPE PQLI came even though both the OMR and the desirability function approaches present two debilitating flaws. They both fail to account for the model parameters uncertainty and to describe the influence of the correlation structure of the regression models residuals. Furthermore, for the establishment of a DS, it is necessary to give an assurance of quality as stated by the regulatory agencies. The flaws mentioned above render the assurance of quality almost impossible, as shown by Peterson (Peterson J. J., 2008). In fact, even the best point in the "sweet spot" can have small reliability with regards to meeting QTPP. The following two examples that prove that in a simple way are taken from the works of Peterson J. J. (2008) and Peterson and Lief (2010).

Consider an acceptance region $\mathbf{S}$ of Equation (3.1) composed of $(-\infty, 1] \times (-\infty, 1] \times (-\infty, 1] \times (-\infty, 1]$ and a length four response vector $\mathbf{y} \sim N(0, \boldsymbol{\Sigma})$. Now consider the variance covariance matrix $\boldsymbol{\Sigma}$, if it is composed of all ones in the diagonal and all 0.9 in the off diagonal

the probability $P(\mathbf{y}\in\mathbf{S}) = 0.75$, if the off-diagonal values are smaller, it is found that $P(\mathbf{y}\in\mathbf{S})$ can be substantially lower.

Furthermore, consider a mean response vector $\mathbf{y}$ composed of n mean response values of the CQAs to the CPP and material attributes configuration. Suppose to need the CQAs to be lower than a specific threshold u_i. OMR and desirability functions work with mean responses without considering that these responses are instead random variables, $\mathbf{y}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Setting a response $\mathbf{y}_i$ to be below the desired threshold u_i means, since $\mathbf{y}_i$ is normally distributed, that the $P(y_i<u_i)\geq 0.5$ in virtue of the shape of a normal distribution. Extending this to the whole vector composed of n mean responses means that $P(y_1,y_2,\ldots,y_n<u_i)\geq 0.5^n$ using some basic probability propriety. This is true only if the responses have no correlation, i.e. they are independent. If they are positively correlated, this will result in a more significant probability, while the contrary is true if they are negatively correlated. From these examples, it is clear that a new approach is necessary, one that accounts for the correlation structure of the data, the uncertainty of model parameters and the variability of the process distribution.

Furthermore, it is necessary to provide the assurance of quality that the regulatory agencies require.

3.3 Assurance of Quality in the Bayesian Approach

Two options are identified for the manufacturers to give the required assurance of quality. First, they should demonstrate that the operating conditions are fully under control, without statistical modelling of his process. Second, evidence should be shown that the quality of the outcome or the product remains within acceptable limits, for changes in input variables within identified limits. The first option is generally difficult or too expensive to achieve. The second option is more realistic because it considers the inevitable variability in the achievement of quality (Peterson and Lief, 2010). Peterson (2008) first introduced the concept of Bayesian probabilistic design space as the most suitable technique to quantify the "assurance" of quality as defined by the regulatory agencies. The Bayesian approach offers other theoretical benefits, such as the robustness of the derived model, which can easily accommodate for "noisy variables" and the ability to account for missing values. Furthermore, the Bayesian approach does not present the flaws found in the OMR approach; it processes inherently correlations, uncertainty and variability. The next section describes the mathematical formulation of this approach.

3.3.1 Bayesian design space: mathematical formulation

According to Peterson (2008) the Bayesian DS can be defined as:

$$DS = \{\mathbf{x} : P(\mathbf{y}|\mathbf{x}, \text{data}) \geq \theta_{th}\}, \tag{3.7}$$

with $\boldsymbol{x}$ one of the possible configuration of the CPP and material attributes for the process, $\mathbf{y}$ a vector of CQAs and θ_{th} a user-defined probability threshold. A necessary comment must be made on the value of θ_{th}. The regulatory agencies have not stated a minimal threshold for a presented DS. However, as proved by Peterson (2008), a minimum value can be set to 80%.
To use Equation (3.7) it is necessary to have a model that links the process inputs, CPP and raw materials properties, to the CQAs of the products. Throughout this study, the model used is a multivariate linear regression as in Equation (2.36):

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}_S, \tag{3.8}$$

with the same variables and dimensions and the residuals supposed independent and following a normal distribution. In order to establish a DS, it is necessary to apply the Bayesian method discussed in chapter §two to obtain a posterior predictive distribution as the one of Equation (2.43) with the multivariate parameter $\boldsymbol{\Sigma}$ in place of the variance of unit weight.
The posterior predictive distribution conditional on data and regression parameters is

$$g(\mathbf{y}|\mathbf{x}, \text{data}) = \iint f(\mathbf{x}|\mathbf{y}, \mathbf{B}, \boldsymbol{\Sigma_B}) \text{P}(\mathbf{B}, \boldsymbol{\Sigma_B}|\text{data}) d\mathbf{B} d\boldsymbol{\Sigma_B}, \tag{3.9}$$

where $\text{P}(\text{B},\Sigma_B|\text{data})$ is the joint posterior distribution of the model parameters. From Equation (3.5) Equation (3.4) can be obtained with integration at the desired threshold.

$$\text{P}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{Y}) = \int_{\theta_{th}} g(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{Y}) d\mathbf{y} \tag{3.10}$$

It is easy to see, from Equations (3.7) and (3.9) that it is necessary to obtain the joint posterior distribution of the model parameters to solve Equation (3.10).
The joint posterior distribution can be computed using the Bayes' theorem of Equation (2.32), thus giving:

$$\text{P}(\mathbf{B}, \boldsymbol{\Sigma_B}|\text{data}) \propto \mathcal{L}(\mathbf{B}, \boldsymbol{\Sigma_B}|\mathbf{Y}) \text{P}(\mathbf{B}|\boldsymbol{\Sigma_B}), \tag{3.11}$$

where $\mathcal{L}(\mathbf{B}, \boldsymbol{\Sigma_B}|\mathbf{Y})$ is the likelihood function of the parameters conditional to the response contained in the data and $\text{P}(\mathbf{B}|\boldsymbol{\Sigma_B})$ is the joint distribution of the model parameters. The likelihood function of Equation (3.11) can be expressed as (Lenk, 2001):

$$\mathcal{L}(\mathbf{B}, \boldsymbol{\Sigma_B}|\mathbf{Y}) = |\boldsymbol{\Sigma_B}|^{-\frac{n}{2}} \text{e}^{\left(\frac{1}{2} \text{trace}\, [\boldsymbol{\Sigma_B}^{-1} (\mathbf{Y}-\mathbf{XB})^{\text{T}} (\mathbf{Y}-\mathbf{XB})]\right.}. \tag{3.12}$$

The model error term, which accounts for the model variability and the error stemming from the parameters variability, $\mathbf{E}_s$ in Equation (3.8), has to be calculated in order to correctly estimate $\mathbf{Y}$. The model error is assumed as normally distributed with error $\boldsymbol{\Sigma}_S$:

$$\mathbf{E}_S \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_S). \tag{3.13}$$

To account for an unknown variance parameter it is known (e.g. Gelman *et al.*,(2013)) that the appropriate non-informative prior distribution is the Inverse-Wishart distribution, the multivariate generalisation of the Inverse-Gamma distribution. The likelihood of the model error is then calculated as:

$$\mathcal{L}(\boldsymbol{\Sigma}_S|\upsilon, \mathbf{G}, \text{data}) = \frac{\boldsymbol{\Sigma}_S^{\frac{\upsilon-n-1}{2}}}{\mathbf{G}^{\frac{\upsilon}{2}}} \mathrm{e}^{\left[-\frac{1}{2}\text{trace}(\mathbf{G}^{-1}\boldsymbol{\Sigma}_S\right]}, \tag{3.14}$$

where $\upsilon > n$-1 is the degree of freedom of the distribution, n is the dimension of $\boldsymbol{\Sigma}_s$ and $\mathbf{G}$ is a positive definite, symmetric, $n \times n$ scale matrix. Once all the factors have been calculated, Equation (3.10) can be solved and a Bayesian DS can be established.

This Bayesian approach to design space has several advantages already mentioned above, but can still be improved.

There are problems in the representation of the obtained results; this derivation of the DS suffers from the "curse of dimensionality" as did the OMR approach. Furthermore, the contribution of measurement error to the established DS has to be considered from a Bayesian point of view.

3.3 Design space determination: a joint Bayesian/latent variable approach

In order to address the dimensionality issue of the Bayesian DS, the main strategies used are results table or arrays of bivariate matrices, keeping the interpretation of the results not straightforward and hard for non-scientist or technicians. In a recent paper from Bano *et al.* (2018) a new approach has been proposed, utilising the dimensionality reductions of the PLS method explained in chapter §2.2 in conjunction with the Bayesian methodology to establish a DS explained above. This methodology has the clear benefit of permitting the easy representation of the resulting Bayesian DS, permitting ease of interpretation from technicians and non-technicians alike. This data-driven method employs historical data from known processes defined similar (e.g. by the method proposed by Jaeckle and MacGregor(1998))

Keeping in mind the definition of knowledge space (KS) and historical dataset in chapter §two it is possible to rewrite Equation (3.7) as (Bano *et al*, 2018):

$$\{\mathbf{x} \in \text{KS} : \text{P}(\mathbf{y} \in \text{AR}|\mathbf{x}, \mathbf{X}, \mathbf{Y}) \geq \theta_{th}\}. \tag{3.15}$$

where the Acceptance Region (AR) is the region in which all product quality attributes meet their acceptance criteria. The KS is identified in the latent variables space starting from the historical dataset, it is then discretized and the belonging to the AR of each discrete point is assessed according to Equation (3.15). Discretization points belonging to the DS are marked in green, those rejected in red. The method is summarized in Fig 3.1.

The step by step methodology is as follows:

1) Given an historical dataset [**X**, **Y**], a PLS model is built in order to relate **X** to **Y**. The number of latent variables A (see chapter §2.2) is chosen in order to maximise the information retained while maintaining the ease of representation given by the latent variable modelling. Once calibrated a region defined by the 95% confidence limit on the Hotelling's T^2 statistics (Equation 2.13) is identified as the KS.
2) The KS is then discretised; a very large number N_a (e.g. 2000 in the current study) of

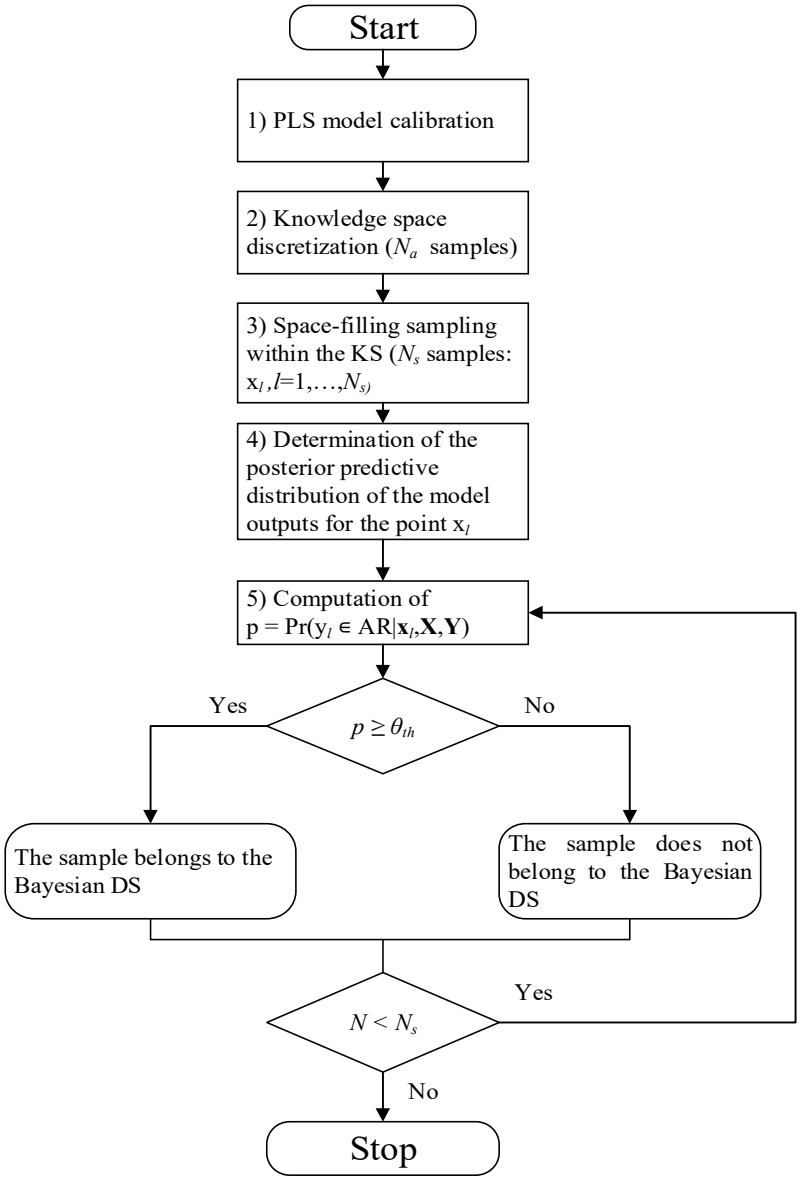

Figure 3.3: *Flowchart for the proposed methodology for the determination of a DS for a new pharmaceutical product taken from Bano et al., 2018*

samples are generated in the latent space inside a square of unitary sides. With a small number of geometrical transformations, the square is scaled to the confidence hyper-ellipsoid size.

3) A representative number of samples $N_s < N_a$ is chosen. The number of chosen samples is a trade-off between good, uniform coverage of the KS and computational load. In this study, a Kennard-Stone algorithm (Kaiser, 1960) has been used for space filling. The Kennard Stone algorithm is a sequential sampling algorithm that selects a subset of N_s samples, given N_a candidates by selecting, at every iteration, the farthest points in term of Euclidean distance. Given the importance of the trade-off between coverage and computational load, a method has been devised and reported in the original publication (Bano *et al.*, 2018) for the selection of the optimal number of points.
4) For each of the N_s selected samples, the joint posterior predictive distribution (PPD) has been computed via Bayesian calculation as explained in section §3.2. The rationale used for the calculation was a sequential metropolis, in which new candidates for the various parameters were generated with an MCMC procedure with a Metropolis-Hastings sampling algorithm (Equations 2.46 and 2.49). Figure 3.2 presents the sampling sequence.

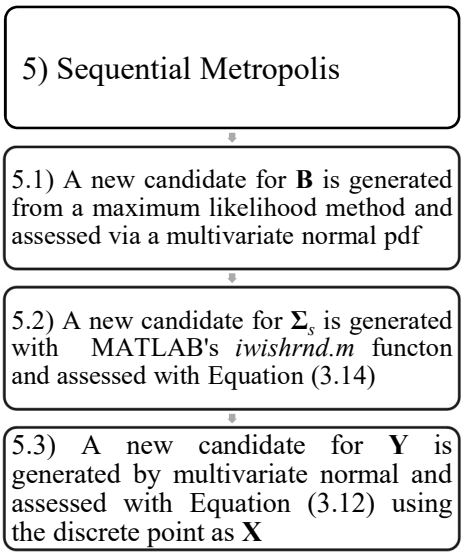

Figure 3.4. *Sampling logic in the Sequential Metropolis algorithm used to compute the PPD for the reported method of establishing a Bayesian DS.*

5) Given the desired quality target $\mathbf{y}_{des}$ (or quality interval) if the condition expressed in Equation 3.3 is satisfied the point is considered inside the DS, if not it is rejected as it does not belong to the probabilistic DS.

An example of a DS established with this methodology is presented in Figure 3.3.

3.4 Latent Space representation of the DS: a case study

As mentioned above this method has the benefit of giving an easy way of illustrating the proposed DS to non-technicians. In almost all cases the accuracy of the latent space 2-

dimensional representation has been satisfactory, but not in all cases. Hereby a case in which the lack of latent space representativeness fails to represent accurately the calculated DS.

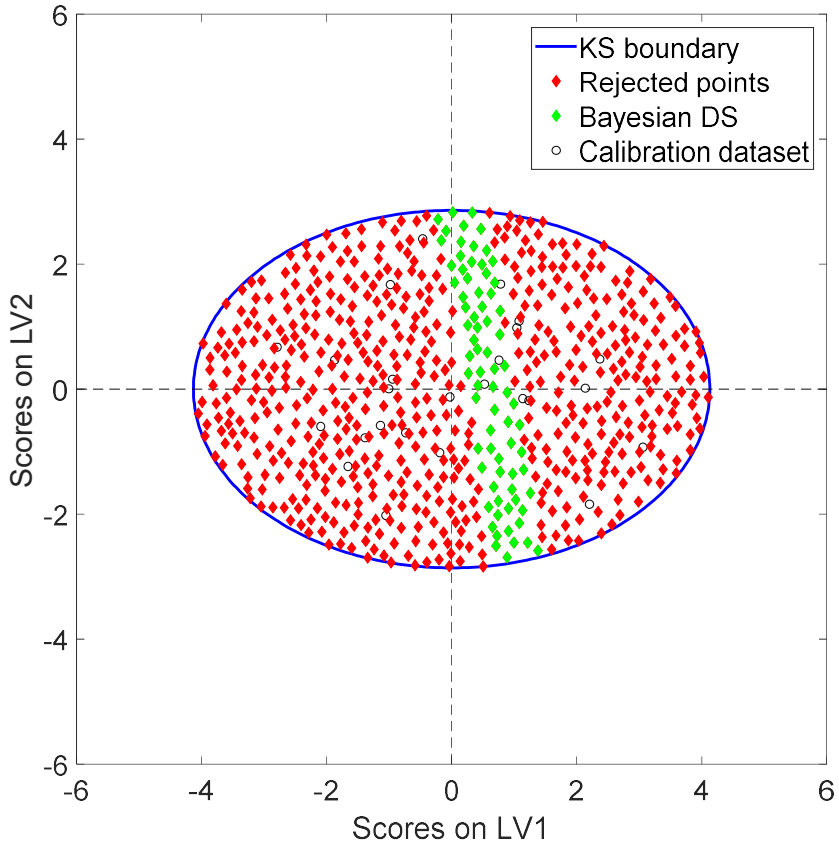

Figure 5.3*: graphical representation of a DS established with the proposed methodology; the green diamonds are points that satisfy Equation 3.3 and thus contained in the DS. The red diamonds do not satisfy Equation 3.3 and are therefore rejected.*

3.4.1 Case study: Simulated Roll Compaction

This case study concerned the dry granulation of a pharmaceutical blend by roller compaction. The data used to model this process were generated by Facco *et al.* (2015), based on the model of Johanson (1965). Roll compaction is a method to provide pre-densification and improve the flowability of powders commonly employed in the pharmaceutical and chemical industries. It coalesces small size particles in larger agglomerate ready for the subsequent milling and tabletting processes (Souihi, 2014).

In a roll compaction process an initial powder mix, in the pharmaceutical industry usually a compressible filler and some API, passes through the gap between two counter-rotating rolls. The powder can be fed, depending on the setup of the machinery, by gravity or by screw feeding. The powder is gripped in the decreasing roll gap by the friction on the roll surfaces and conveyed to the region close to the minimum roll gap, where compression happens, and a compact ribbon is formed. The formed ribbon is then pushed forward and released from the roll. It is generally considered that there are three zones of material behaviour in the roll compaction process, the slip, nip and release zones as shown in Figure 3.4

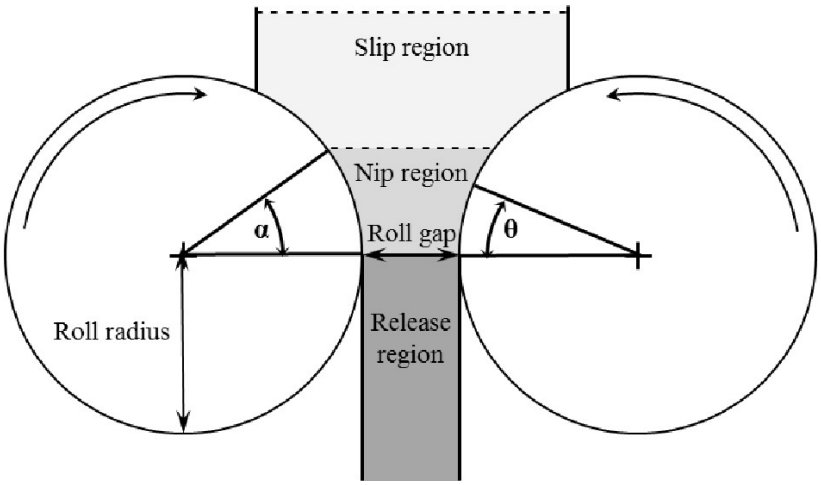

Figure 3.6: *Schematic representation of a Roll Compaction process. The powder is fed to the process in the slip region, is gripped by friction forces in the nip region and compacted. In concordance with the variable names of Table 3.1: Roll gap is s_{roll}, Roll radius is half of D_{roll}, the friction angle α is γ_{FR} and the effective friction angle θ is γ_{EFR}. (Souihi, 2014)*

The powder is fed in the slip zone, characterised by particles slipping on the surface of the roll, i.e. wall velocity of the particles is different from that of the rolls. Particle rearrangements occur, and relatively little pressure is efficiently transferred to the powder in the slip zone. The nip zone, short for non-slip, starts at a roll angle α (γ_{FR} from Table 3.1) when the wall velocity of the powder becomes equal to that of the rolls. After initial rearrangements, the bulk of the pressure is exerted at roll angle θ. Powder densification mainly takes place in this zone. The powder is dragged to the smallest gap and compressed by the substantial increase in the pressure. The release zone starts when the roll cap begins to increase again in which elastic recovery (linked to spring back factor of Table 3.1) can happen. One important factor in roll compaction is that binding of particles results only from the compaction forces requiring a certain degree of compressibility of the powder blend.

This technique offers advantages compared to wet granulation since it does not employ liquid binders and does not require drying stages, this could positively affect ease of processing in the presence of drug substances that are sensitive to moisture, solvent or heat.

3.4.2 Case study results

The historical dataset is composed by ninety input variables (compressibility factor, roller diameter, roller width, roller speed, pressure force, friction angle between solid granulate, and roller compactor, effective friction angle and springback factor) and one response variable (intravoid fraction of the solids out of the compactor). A summary of the input/output variables and the characterisation of the input dataset is reported in Table 3.1.

Table 3.1*: Latent Space misrepresentation case study: list of the input and response Variables (Data from Facco et al., (2015) based on the model of Johanson (1965) and characterisation of the input dataset (Columns 5 and 6))*

ID	Variable name	Units	Symbol	Mean	Std.
Inputs					
1	Compressibility factor	[-]	k	9.85	2.528
2	Roller width	[m]	s_{roll}	0.134	0.015
3	Roller diameter	[m]	D_{roll}	0.398	0.0734
4	Roller speed	[rpm]	ν_{roll}	10.239	6.434
5	Pressure force	[kN]	F_{roll}	13867	6951.2
6	Friction Angle	[rad]	γ_{FR}	27.507	8.778
7	Effective friction angle	[rad]	γ_{EFR}	48.167	31.763
8	Spring back factor	[-]	F_{sb}	0.109	0.0287
Response					
R1	Intra void fractions of solids	[m^3/m^3]	ϕ_s	0.521	0.117

The problem faced in this early case study was the application of the proposed methodology for the mitigation of the Impact of measurement error in Bayesian DS determination on one of the case studies from the paper of Bano et al. (2018). That is the development of a granulate with intra void fraction of solids of 0.641 [m3/m3] with a probability of reaching specifications of at least 90%. As per the followed methodology, a PLS model has been calibrated with the whole dataset. Two latent variables were chosen for ease of representation, but these two latent variables captured only 47.36% of the variation of **X**. The methodology has been applied nevertheless, selecting a multivariate normal prior distribution for the parameters and an Inverse-Wishart prior to the variance-covariance terms. These prior distributions were made as uninformative as possible since no prior knowledge was available to justify a different approach.

The results are presented in Figure 3.4. The blue circles represent a subset of the real DS of the process, derived from first principles modelling by trial and error. As it can be clearly seen the proposed DS does not incorporate most of the real DS points. This confirms one of the issues mentioned in the paper by Bano et al. (2018) on the limitations on the use of the proposed methodology. The amount of the cumulative **X**-variability explained has a relevant influence on the projection of the knowledge space and lowers the precision of the representation of the original input space. It is possible, when the amount of cumulative **X**-variability explained is low, to not represent part of the KS that belong to the DS. The study from Bano et al. (2018) suggests as a rule of thumb to set a minimum cumulative **X**-variability explained to 90%, modified the single investigator risk aversion. To achieve this target it could be necessary to increase the number of selected latent variables, lowering the ease of representation of the proposed DS.

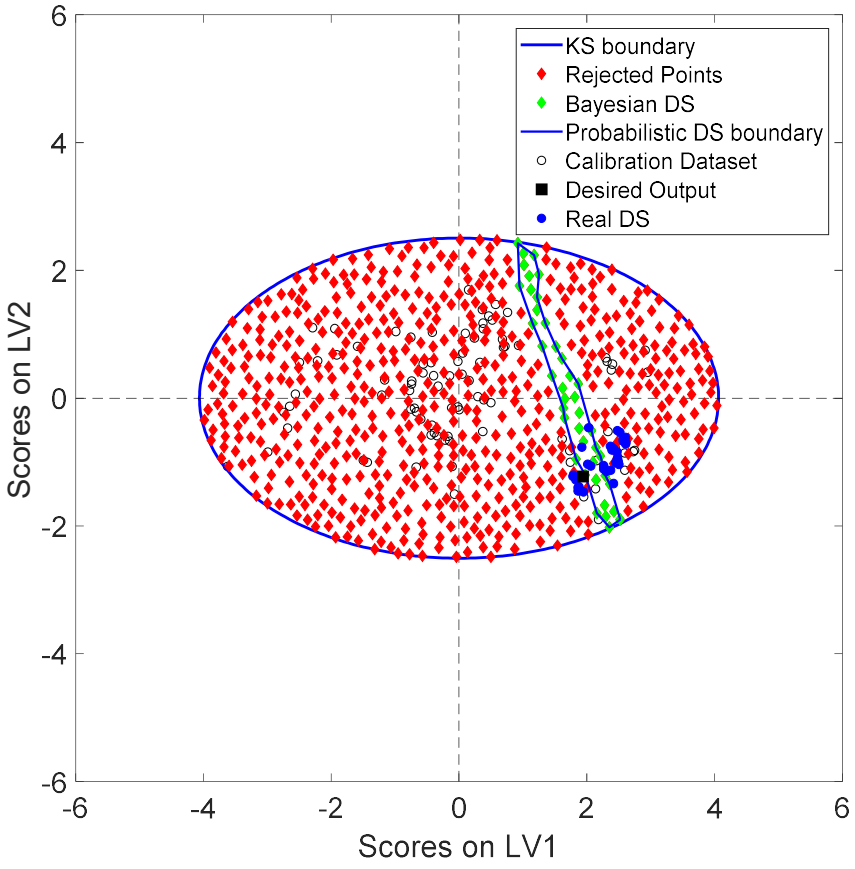

Figure 3.5: *The DS misrepresentation case study: the green diamonds represent the proposed design space. While the red ones are rejected points. Due to the small cumulative **X**-variability explained, the proposed DS is represented with a great uncertainty. As a counterproof the reported Blue circles are a subset of the actual DS, derived by trial and error with a first principles model.*

This misrepresentation gave rise to further study on the topic, and two essential results can be reported. First, the definition of a KS in *n*-dimensions proves difficult since most computational methods calculate sums of convex spaces; the risk is to represent the KS as a sum of acceptable ranges instead of having a continuous function defining it. Secondly, the number of points necessary to discretise the space increase manifold.
This could be explained in a simple thought experiment. It is possible to think of a filling strategy for the KS such as the centre of unitary radius n-dimensional spheres are used as discretisation points up to the point in which no more unitary radius n-dimensional spheres can be added. Using the concept of kissing number, defined as the number of non-overlapping unit spheres that can be arranged such that they each touch another given unit sphere, (Pfender and Ziegler, 2004) it is possible to understand that to fill space, discretising it in a meaningful way, many more points have to be selected. The higher the dimension, the higher the needed points to discretise the space. In the above examples the kissing number of 8-dimensional space is 240 compared to the six of a 2-dimensional space (Pfender and Ziegler, 2004), so if the same "degree of filling" is desired, many more points have to be selected, possibly manyfold. These results suggest that applying a latent space approach to map the original input space and the KS has benefits that go beyond the ease of representation and of inversion of the model.

3.5 Objective of the Dissertation

As discussed in the previous sections, handling the different types of uncertainties that affect a DS identification exercise is a key step for the implementation of the concept of "assurance" of quality put forth by the regulatory agencies.
The aim of this Dissertation is to extend methodologies discussed in the previous sections by accounting for an additional source of uncertainty, namely the uncertainty added by the presence of measurement errors on the calibration dataset.
An extention to the bayesian framework described above has been developed and the methematical details will be thoroughly discussed in the next chapter.

Chapter 4

Handling measurement uncertainty in Bayesian design space determination

In this chapter a review of the sources and impact of the measurement error will be presented in brief. Proceeding with the chapter a methodology for its incorporation in the Bayesian approach to the determination of the design space of a new pharmaceutical product will be proposed, followed by an explanatory example of its application.

4.1 Measurement uncertainty: introduction

In the manufacturing sector measurements have an essential role as they are used as a diagnosis of a process, from its start to its end. With the QbD initiative data from measurement retains their importance because several decisions of compliance or non-compliance are based on measured results and process modelling. Even more than before the industry demands the acquisition of reliable in-process analytical data (ICH, 2009). As a consequence of these requirements, manufacturers should demonstrate the quality of their results and their fitness for purpose by giving a measure of the confidence that can be placed on the results. Furthermore based on the needs for an "assurance" of quality it is important to be able to account for all sources of uncertainty and their impact on the exercise of establishing a design space.

4.1.1 Classification of the sources of measurement errors

Measurement errors can be divided into those arising during the measurement process and those that arise due to a later corruption of the measurement signal, e.g. while traveling from the sensor to a transducer. From those arising during the measurement process the most relevant are systematic and random errors. Systematic errors are errors in the output readings that are consistently on one side of the correct observation. System disturbance is one of the most frequent sources of systematic measurement error (Morris and Langari, 2016). The act of measuring a system perturbs the system in question. This disturbance could be minimal, as the one caused by using a cold glass thermometer to measure the temperature of boiling water, or as major as the pressure drop introduced in a pipe by a measure of flow with an orifice plate. In the first case, the cold glass thermometer would lower the water temperature, even if at a

minimal extent, while in the second case an intense pressure drop would be generated in the fluid. In view of the above examples, it is clear that the magnitude of the disturbance varies between different systems from and has to be accounted when designing the measurement system. To design a measurement system that minimizes disturbance a knowledge of the mechanism that generates the disturbance and an evaluation of the effective impact on the measurement are needed. In the examples above in the first case, the disturbance is so minor that the measurement system – the glass thermometer – cannot detect it. In the second case, the pressure drop is known and extensively modelled, so its effect can be efficiently balanced. Another primary source of systematic error is the presence of environmental inputs. An environmental input is defined as an apparently real input to a measurement system that is actually caused by a change in the environmental conditions surrounding the measurement system (Morris and Langari, 2016). Static and dynamic charateristics of an instrument are only valid for particular environmental conditions, these specified conditions must be reproduced during calibration in order to avoid wrong results. Away from the specified conditions the performance of the instrument can vary and cause an error in the measured quantity. The amount of this variation from the true measure is quantified by two constants, zero drift and sensitivity drift. In the presence of zero and sensitivity drift it is often difficult to determine how much of the response is due to environmental input and how much to an actual change to the measured variable. The magnitude of environmental inputs should be measured before the real input can be determined. Generally it is very difficult to avoid the influence of environmental inputs, because it is rarely pratical, or even possible, to control che environmental conditions sorrounding the measurement system (Morris and Langari, 2016) and accurate design choices are needed to minimize these influences. Other sources of systematic errors can include poorly calibrated or uncalibrated instruments, poorly maintained measurement systems of drift in the instrument characteristics. As mentioned above, even with good calibration and maintenance standards, coupled with properly trained and attentive technicians, some error remains that are inherent in the manufacture of an instrument. These errors are quantified by the accuracy value published in the instrument data sheet along with zero and sensitivity drifts.

In addition to systematic errors, random errors also occur in normal measurement practice. They are perturbations of the measurement either side of the true value caused by random and unpredictable effects, such that positive and negative errors occur in approximately equal numbers for a series of measurements made of the same quantity. Such perturbations are mainly small, but large perturbations occur from time to time (Morris and Langari, 2016). Random errors can be caused by faulty observations of an instrumental reading by an inattentive technician (in lab scale experiments), from electrical noises (in plant scale measurements) or from random environmental changes. Random errors can be mitigated by repeated measurements and by statistical analysis. Because of the unpredictable nature of these random

errors, the best approximations are obtained in probabilistic terms. A measurement could be assigned a confidence of 95% and a span of ±1, but in 5% of the cases the actual value would be outside the set boundaries. An accuracy of 100% can never be attained in measured quantities that are subject to random errors (Morris and Langari, 2016).
In addition to systematic and measurement errors, that arise during the process of measurement by an instrument, another type of error can affect a measurement without being, per se, a measurement error. This is the case of induced electrical noise, or induced voltage noise. These noise terms arise when an electrical output signal generated by sensors or transducers are corrupted by induced currents. (Morris & Langari, 2016).
The principal induced voltage noise action modes are differential and common. Differential noise mode arise inside a circuit when the noise source acts in serie with the voltage output of a sensor or a transductor, these can cause significant errors in the measured output. The magnitude of corruption from a differential noise is called signal-to-noise ratio defined as:

$$SNR_{dB} = 20\log_{10}\left(\frac{V_s}{V_n}\right). \tag{4.1}$$

where V_s is the mean voltage of the signal and V_n is the mean voltage of the noise.
In the case of AC differential mode noise voltages, the root-mean-squared value of the voltage is used in place of V_n (Morris and Langari, 2016). Common mode noise voltages are less impacting, because they usually affect both side of the signal circuit modifiying both outputs by the same level, thus having no effect on the level of output signal, but they must still be considered as sometimes common mode noise voltage can generate a differential mode noise voltages. Induced voltages can arise both inside the measurement circuit and during the transmission of the signal. The most common source of noise during the transmission of the signal is the proximity to other electric appliances, power sources or radio signals. The most common sources of noise voltages inside a measurement circuit include thermoelectric potentials, shot noise and potentials due to electrochemical action (Morris and Langari,2016).
Although much can be made to enhance measurement systems in order to better process error sources, either random, systematic or due to noise voltages, these uncertainties can never be removed altogether prompting all the research that have been done in the area.

4.1.2 Impact of measurement errors on process modeling

As discussed in the previous section, many sources of measurement error exist. With the ever increasing importance placed on modeling by the regulatory agencies and the call for accurateness and reliability of those models (see chapter §1), it is necessary to account for the effect of measurement error at the calibration stage of any model adopted to assist a DS identification exercise.

Measurement error in the variables of a model can potentially have three notable effects (Carroll *et al*, 2006):

- it may cause a bias in the parameter estimation for statistical models;
- it may lead to a loss of power, sometimes profound, for detecting interesting relationship among variables;
- it may mask the main features of the data, making graphical model analysis difficult.

While most of the work in literature is focused on the first two effect, relatively little work has been done to account for the third effect. To understand the effect of measurement error, a brief example adapted from Carroll *et al,* (2006) is presented.

Let x be an error prone variable uniformly distributed on an interval $[-\pi, \pi]$. Suppose that y is a response with a mean of sin $(2x)$ with a small standard deviation of 0,1. In Figure 4.1, the blue circles represent 628 draws from the variable x measured without error. Now suppose that instead of detecting the true variable x it is only possible, due to the measurement system in use, to observe w, a variable with mean equal to x and standard deviation of 0.8.

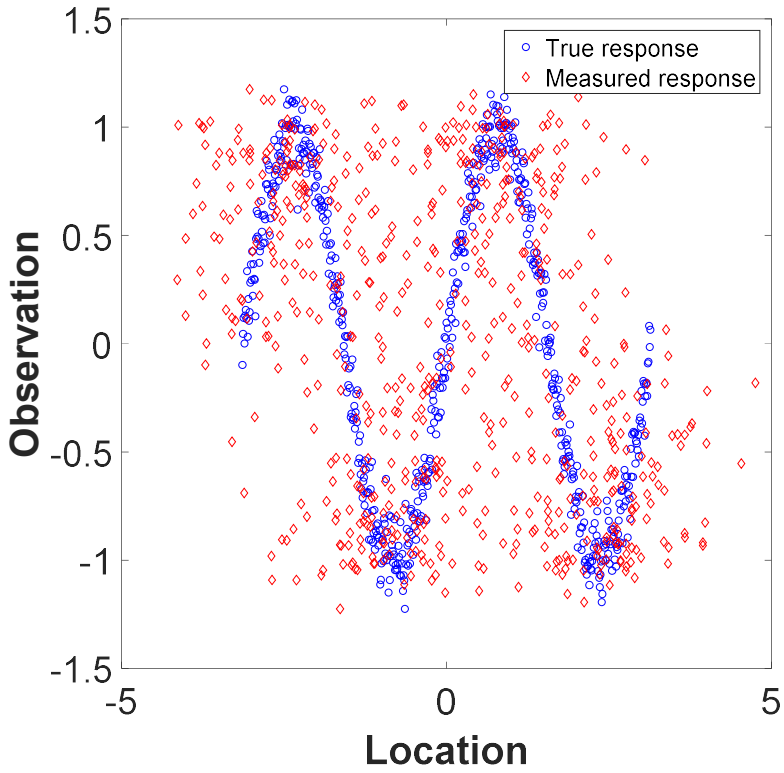

Figure 4.7: *Comparison between a true quantity and a measured one, observations have mean equal to sin (2**x**) and small standard deviation, blue circle are shown at the right location **x**, while red diamonds are shown at a location **w** measured with error; **w**~N (**x** ,0.8). (Adapted from Carroll et al, 2006)*

It is clear that the sinusoidal response is not readily visible in the red diamonds, the error has hidden the key feature of the data and someone trying to construct a basic regression from w would probably not use a sinusoidal interpolating function. It is also possible to note the loss

of predictive power, the variability of (x, y) around the sinusoidal function is lower than any curve that could be fitted from (w, y).

In presence of variables measured with error it is not possible to calculate the parameters of the regression of y on x, since x is not observed. The usual goal of measurement error modelling then is to build a bridge that will permit to compute the parameters of y on x starting from w. This goal is not as straightforward as it seems and if the values of w were to be used instead of x**,** obvious errors would arise (Carroll et al, 2006).

The effect on measurement error on parameter estimation has been extensively studied in the literature (Abramowitz *et al.*, 2005) (Allegrini *et al*, 2018) (Aoki *et al*, 2001) (Mallick *et al*., 2002). In particular, in the context of regression analysis, it has been proved that the measurement error, even if a simple additive error, causes the parameters of a regression of y on x to be biased towards zero:

$$\mathrm{B}^* = \mathrm{B}(\frac{\sigma_x}{\sigma_w}). \tag{4.2}$$

Equation (4.2) is valid only for univariate x. For the regression of a multivariate matrix of variables **X** on a vector of responses **y** the regression parameters **B** will be biased accordingly to the following equation (Carroll *et al*, 2006):

$$\mathbf{B}^* = (\boldsymbol{\Sigma_{XX}} + \boldsymbol{\Sigma_{UU}})^{-1}[(\boldsymbol{\Sigma_{XX}})\mathbf{B} + \boldsymbol{\Sigma_{UE}}], \tag{4.3}$$

where Σ_{ab} stands for the covariance between a and b, **U** is the matrix containing the measurement error of **X** and **E** is the matrix of the residuals as in Equation (2.36).

The simplest type of error model is the classical error model, expressed as:

$$w_{ij} = \mathbf{x}_i + \Delta \mathrm{x}_{ij}. \tag{4.4}$$

In the classical error model, the true variable is measured with additive error, usually with constant standard deviation (i.e.: unchanging in subsequent repeated measurement).

In this model, the observed quantity w is an unbiased measure of the true value x, and as such it must have zero mean:

$$\mathrm{E}(w_{ij}|x_i) = 0\,. \tag{4.5}$$

In the classical measurement error model, the error can be heteroscedastic or homoscedastic, which means that the error variance could or could not depend on the measured quantity. In the case of homoscedasticity of the error, the $\boldsymbol{\Delta X}$ is independent and identically distributed with a normal distribution, that is:

$$\boldsymbol{\Delta X} \sim \mathrm{N}(0, \boldsymbol{\Sigma}_{\Delta x}). \tag{4.6}$$

In the case of heteroscedasticity, the variance of the error depends on the value of another variable. In this case, the variance of the error term could be, e.g.:

$$\boldsymbol{\Delta X} \sim \mathrm{N}(0, \alpha_0 + \alpha_1 f(\mathbf{X}) + \alpha_2 f(\mathbf{Z})), \tag{4.7}$$

where **Z** is a matrix of covariates of **X**.

Another error model that could be of importance in industrial applications is the Berkson error model. With this model, the measured variable is in some way approximated, i.e.:

$$x_{ij} = \mathbf{w}_i + u_{ij}, \tag{4.8}$$

with $\mathbf{w}_i$ an unbiased measured of $\mathbf{x}_i$;

$$\mathrm{E}(u_{ij}|\mathbf{w}_i) = 0, \tag{4.9}$$

information of the actual variation of the true variable $\mathbf{x}_i$ is lost in the measurement. This type of error is common when a fixed value is assumed for the **x** variable (e.g.: a fixed minimum distance between rolls in a roll compactor, or a fixed value of weight for a powder fed in different sieve tray analysis). When measurement are taken continuously it can be assumed that the errors are independent and identically distributed, clearly this assumption has to be verified. A simple procedure that can be used to do so is the visual inspection of the quantile-quantile plot of the differences between the measurements. A Quantile-Quantile plot is a plot of points whose coordinate are the ordered points of the dataset versus the expected value of the distribution at the same quantile. If the plot is linear, the sample likely comes from the distribution used to compute the other quantiles (Everitt and Skrondal, 2010). While applied to the whole dataset this test would give no indication of the distribution of the errors applying it

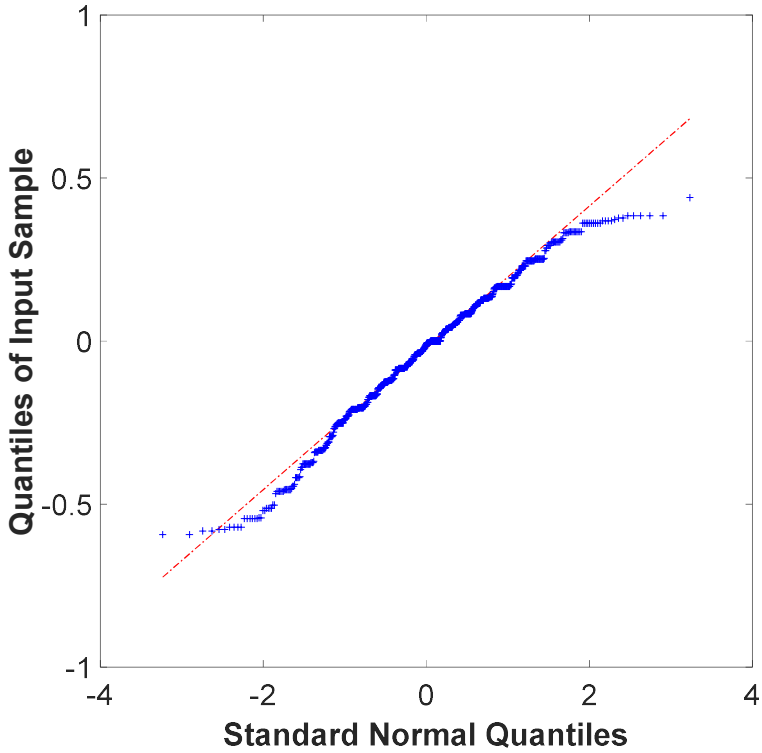

Figure 4.2: *Quantile-Quantile plot of one of the variable used for case study #2 in chapter §5, as it can be seen the intra variable difference appears to be normally distributed*

to the difference between replicates of the same measurement can give an idea of the

characteristics of the error structure. While it provides a quick visual way to assess normality of a distribution the use of a q-q plot has its downsides as it is reported that it can lead to false assumption of adherence to a normal distribution for samples that are distributed according to distributions that have similar shapes. Non-normality is evident only in the presence of heavier-than-Gaussian tails (Carroll *et al*, 2006), such as those of a lognormal or a pareto distribution.
The models presented in Equations (4.4) and (4.8) are the most simple models for the modelling of errors, but most of the assumption at the base of their formulation hold in industrial cases if the measurement systems are accurately designed.
Other error models include the "general classical measurement model" or the "regression calibration model" that are generalization of the classical and Berkson models already presented, and models tuned for specific applications, where the structure and the nature of the error term is problem-specific and well-understood and can be modelled extensively, e.g.: the model to correlate the estimated glomerular filtration rate, found via correlation equations, to the actual glomerular filtration rate, used to predict the progression of coronary kidney disease.
It is worth noting that outside the statistical and medical literature most applied models are focussed on linear, univariate application.

4.2 Modeling measurement error for Bayesian DS determination

In the following, a procedure to account for measurement uncertainty in the Bayesian DS determination methodology proposed in the previous Chapter is presented.
First, the problem statement will be presented. Secondly, the mathematical formulation of the proposed approach will be disclaimed. Lastly, the step-by-step methodology will be presented and discussed with an illustrative case study.

4.2.1 Problem statement

The methodology will be discussed for the multivariate linear regression model of Eq. (3.8):

$$\mathbf{Y} = \mathbf{WB} + \mathbf{E}_s \tag{4.10}$$

where **W** now substitutes **X** as it is assumed that **X** is measured with error.
It has been reported in chapter §4.1.1 that measurement error could be of random and systematic nature or due to induced noise voltages. Herein only random errors are considered, since in most industrial cases continuous measurements taken with well-maintained and well-designed measurement systems, remove all systematic measurement errors and induced electrical noise up to the instrument accuracy. For this reason the measured quantity, **W**, is assumed to be an unbiased measure for the values of **X**.
The model used to incorporate the measurement error is the classical measurement error model of Equation (4.4).

By plugging in Equation (4.4), into Equation (4.10), Equation (4.10) can be rewritten as:

$$\mathbf{Y} = (\mathbf{X} + \mathbf{\Delta X})\mathbf{B} + \mathbf{E}_s. \quad (4.11)$$

Equation (4.11) can be rearranged to obtain an expression similar to the one of Equation (3.8), obtaining: of,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}_m + \mathbf{E}_s, \quad (4.12)$$

where $\mathbf{E}_m$ is $\mathbf{\Delta X} \times \mathbf{B}$, which is the error term related to the measurement noise.
Let $\mathbf{E}_{tot}$ be the sum of the residuals due to the measurement noise and the structural model error, i.e.:

$$\mathbf{E}_{tot} = \mathbf{E}_m + \mathbf{E}_s. \quad (4.13)$$

The model parameters are assumed to be normally distributed with mean $\bar{\mathbf{B}}$ and standard deviation equal to $\mathbf{\Sigma}_{tot}$, i.e.:

$$\mathbf{B} \sim \mathrm{N}(\bar{\mathbf{B}}, \mathbf{\Sigma}_{tot}). \quad (4.14)$$

It is important to note that this approach is different from the one used in classical Bayesian error modelling. In classical Bayesian error modelling, which is the hierarchical modelling, another model would be necessary as exposure model for the true variable **X** (Carroll *et al*, 2006), in addition to the necessity to compute an increased number of parameters. Other than **B**, and the components of $\mathbf{\Sigma}_{tot}$, most models ask for the Bayesian computation of at least structural error, measurement error and true variable standard deviation and means, in addition to other hyperparameters that could be necessary to express the final model (Carroll *et al*, 2006).
In order to perform a combined Bayesian calibration of **B** and $\mathbf{E}_{tot}$ it is necessary to:

- find a pdf for $\mathbf{E}_{tot}$, given the pdfs of $\mathbf{E}_m$ and $\mathbf{E}_s$. In other terms, the problems consists of determining the pdf for a summation of two multivariate Gaussian distributions.
- find a pdf for $\mathbf{E}_m$, given the pdfs of **ΔX** and **B.** In other terms, the problems consists of determining the pdf for a product of two multivariate Gaussian distributions.

A brief discussion on how to solve these two issued is described in the following.

4.2.2 Summation of two multivariate Gaussian distributions

The probability distribution of a sum of distributions is the convolution between the individual distributions (Bromley, 2014). The convolution theorem is used to this purpose:

$$\mathrm{P}_{f \otimes g}(\mathbf{x}) = \mathrm{F}^{-1}[\mathrm{F}(f(\mathbf{x}))\mathrm{F}(g(\mathbf{x}))] = f(\boldsymbol{x}) \otimes g(\boldsymbol{x}) \quad (4.15)$$

where F is the Fourier transform, F^{-1} denotes the inverse Fourier transform and $\otimes$ is the convolution symbol, while $f(\mathbf{x})$ and $g(\mathbf{x})$ are two probability density functions. The Fourier transform of $f(\mathbf{x})$ is defined as:

$$\mathrm{F}\big(f(\mathbf{x})\big) = \frac{e^{-2\pi i k \boldsymbol{\mu}_f}}{\sqrt{2\pi\boldsymbol{\Sigma}_f}} \int_{-\infty}^{\infty} e^{-\frac{\mathbf{x}'^2}{2\boldsymbol{\Sigma}_f}} e^{-2\pi i\ \mathbf{x}'}\, \mathrm{d}\mathbf{x}' \tag{4.16}$$

with $\mathbf{x}'=\mathbf{x}-\boldsymbol{\mu}_f$ mean-centred value of $\mathbf{x}$.
Euler's formula can then be applied to split the integral in $\mathbf{x}'$:

$$\mathrm{F}\big(f(\mathbf{x})\big) = \frac{e^{-2\pi i k \boldsymbol{\mu}_f}}{\sqrt{2\pi\boldsymbol{\Sigma}_f}} \int_{-\infty}^{\infty} e^{-\frac{\mathbf{x}'^2}{2\boldsymbol{\Sigma}_f}} [\cos(2\pi k\mathbf{x}' - i\sin(2\pi k\mathbf{x}')]\mathrm{d}\mathbf{x}' \ , \tag{4.17}$$

The second part of the formula contains the integral from -∞ to ∞ of the sine function. This function is odd, so its integral in a space of equal positive and negative spanning will be zero. In virtue of this, it can be crossed out leaving only the cosine term.
The remaining integral can be solved obtaining:

$$\mathrm{F}\big(f(\mathbf{x})\big) = e^{-2\pi i k \boldsymbol{\mu}_f} e^{-2\pi^2 k^2 \boldsymbol{\Sigma}_f} \ . \tag{4.18}$$

The same applies for $g(\mathbf{x})$ giving a similar result. Multiplying the two Fourier transform it is possible to write:

$$\mathrm{F}\big(f(\mathbf{x})\big)\mathrm{F}\big(g(\mathbf{x})\big) = e^{-2\pi i k \boldsymbol{\mu}_f} e^{-2\pi^2 k^2 \boldsymbol{\Sigma}_f}\, e^{-2\pi i k \boldsymbol{\mu}_g} e^{-2\pi^2 k^2 \boldsymbol{\Sigma}_g} \ , \tag{4.19}$$

$$\mathrm{F}\big(f(\mathbf{x})\big)\mathrm{F}\big(g(\mathbf{x})\big) = e^{-2\pi i k (\boldsymbol{\mu}_f + \boldsymbol{\mu}_g)} e^{-2\pi^2 k^2 (\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_g)} , \tag{4.20}$$

Since these Fourier transforms are invertible, Equation (4.15) can be solved to determine that the resulting probability density function will have a mean equal to the sum of the two means and a standard deviation equal to the sum of the individual standard deviations:

$$\mathrm{F}^{-1}[\mathrm{F}(f(\mathbf{x}))\mathrm{F}(g(\mathbf{x}))] \ = \frac{1}{2\pi^{\frac{n}{2}}\sqrt{|(\boldsymbol{\Sigma}_f + \boldsymbol{\Sigma}_g)|}} e^{-\frac{1}{2}(\mathbf{x}-(\boldsymbol{\mu}_f+\boldsymbol{\mu}_g))^{\mathbf{T}}(\boldsymbol{\Sigma}_f+\boldsymbol{\Sigma}_g)^{-1}(\mathbf{x}-(\boldsymbol{\mu}_f+\boldsymbol{\mu}_g))} \tag{4.21}$$

where n is the dimensionality of the PDFs.

$$\boldsymbol{\mu}_{sum} = \sum_{i=1}^{I} \boldsymbol{\mu}_i \tag{4.22}$$

$$\boldsymbol{\Sigma}_{sum} = \sum_{i=1}^{I} \boldsymbol{\Sigma}_i \tag{4.23}$$

From Equations 4.22 and 4.23 it is the possible to see that the sum of probability density functions (PDFs) via convolution is a PDF with mean equal to the sum of the mean and covariance matrix equal to the sum of the covariance matrices.

4.2.3 Multiplication of two multivariate Gaussian distributions

It is possible to define a canonical parametrization for a normal PDF that can be linked to the usual moments parametrization by:

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \tag{4.24}$$

$$\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \tag{4.25}$$

Where $\boldsymbol{\eta}$ is the information vector, $\boldsymbol{\Lambda}$ is Fisher information matrix, $\boldsymbol{\mu}$ is the *n*-vector of the means and $\boldsymbol{\Sigma}$ is the *n*-by-n covariance matrix.

Using the transformation of Equations 4.24 and 4.25 a PDF can be written in canonical notation, i.e.:

$$\mathrm{P}(\mathbf{x}) = \mathrm{e}^{[\boldsymbol{\alpha}+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}]}, \tag{4.26}$$

where the quantity α is a scale factor:

$$\boldsymbol{\alpha} = -(n\log(2\pi) - \log|\boldsymbol{\Lambda}| + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta}) \tag{4.27}$$

and *n* is the dimensionality of **x**.

Using the canonical notation is possible to compute the multiplication of the functions as a summation of the single PDFs, i.e.:

$$\prod_{i=1}^{I}\mathrm{P}_i(\mathbf{x}) = \exp[\,\boldsymbol{\alpha}_{i=1\ldots I} + (\sum_{i=1}^{I}\boldsymbol{\eta}_i)^{\mathrm{T}}\mathbf{x} - \tfrac{1}{2}\,\mathbf{x}^{\mathrm{T}}(\sum_{i=1}^{I}\boldsymbol{\Lambda}_i)\mathbf{x}], \tag{4.28}$$

with

$$\boldsymbol{\alpha}_{i=1\ldots I} = -\mathrm{In}(\,\log(2\pi) - \sum_{i=1}^{I}\log|\boldsymbol{\Lambda}|_i + \sum_{i=1}^{I}\boldsymbol{\eta}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\boldsymbol{\eta}_i). \tag{4.29}$$

It is possible to transform Equation (4.29) to a form in principle equal to the one of Equation (4.26) multiplied by a scale factor by adding and subtracting $\boldsymbol{\alpha}_I$, giving:

$$\prod_{i=1}^{I}\mathrm{P}_i(\mathbf{x}) = \exp[\boldsymbol{\alpha}_{i=1\ldots I} - \boldsymbol{\alpha}_I]\exp[\,\boldsymbol{\alpha}_I + (\boldsymbol{\eta}_I)^{\mathrm{T}}\mathbf{x} - \tfrac{1}{2}\,\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda}_I)\mathbf{x}] \tag{4.30}$$

where the total information vector is:

$$\boldsymbol{\eta}_I = \sum_{i=1}^{I}\boldsymbol{\eta}_i, \tag{4.31}$$

and the total Fisher information matrix is:

$$\boldsymbol{\Lambda}_I = \sum_{i=1}^{I}\boldsymbol{\Lambda}_i. \tag{4.32}$$

The total scale factor is given by:

$$\boldsymbol{\alpha}_I = -(n\log(2\pi) - \log|\boldsymbol{\Lambda}|_I + \boldsymbol{\eta}_I^{\mathrm{T}}\boldsymbol{\Lambda}_I^{-1}\boldsymbol{\eta}_I\,). \tag{4.33}$$

The resulting function is still a Gaussian PDF over **x**. Equations (4.31) and (4.32) are then transformed back to moments notation using Equations (4.24) and (4.25). The following formulae for the mean and the standard deviation of the product of PDFs are obtained, i.e.:

$$\boldsymbol{\mu}_{prod} = \boldsymbol{\Sigma}_{prod}^{-1}(\sum_{i=1}^{I}\boldsymbol{\Sigma}_i\,\boldsymbol{\mu}_i) \tag{4.34}$$

$$\boldsymbol{\Sigma}_{prod} = (\sum_{i=1}^{I}\boldsymbol{\Sigma}_i^{-1})^{-1} \tag{4.35}$$

where μ_{prod} is the mean of the product of two or more PDFs and Σ_{prod} is their covariance matrix

4.3 Proposed methodology

Based on the mathematical background described above, step d) of the methodology reported in chapter §3.3 for the Bayesian DS determination has been extended to account for the measurement errors in the calibration dataset. Applying the modelling strategy of the previous section is now possible to combine the results of Equations (4.22) and (4.23) with those of Equations (4.34) and (4.35) to obtain a method of applying Equation (4.13) as the model linking CPP and material attributes to CQAs.

The new proposed methodology is composed of 7 steps and is described as follows:

1. Assess the normality of the error distribution. In this study a q-q plot has been used as shown in Figure 4.2. To assess normality of the measurement error with a q-q plot take a sufficiently representative subset of each variable in the dataset, $\mathbf{x}_j^* \subseteq \mathbf{x}_j$ for all variables *j*, compute the differences between all variables in the subset, the proposed way is to use the *nchoosek.m* command in MATLAB to compute the possible combination and subtract according to these combinations, use the result of the subtraction with the MATLAB *qqplot.m* routine and assess the results. This can be done graphically or automated with a script assessing deviation from the normal distribution via distance from the normal quantile line.
2. PLS model calibration
3. Knowledge space discretisation
4. Space filling sampling within the KS to obtain a representative number of points
5. Use a two-step algorithm for generating a prior for $\mathbf{E}_m$. The first step of the algorithm is a singular value decomposition of the error of $\mathbf{\Delta X}$ to obtain the best rank-K approximation, where K is the number of variables in Y. The goodness of this approximation is assessed with an "eigenvalue greater than one" criteria. The second step is setting an objective function in order to find a $\mathbf{\Sigma}_m$ that satisfies this system

   $$\mathbf{\Sigma}_{tot} = \mathbf{\Sigma}_s + \mathbf{\Sigma}_m, \tag{4.36}$$

   $$\mathbf{\Sigma}_m = (\mathbf{\Sigma}_{\Delta X}^{-1} + \mathbf{\Sigma}_{tot}^{-1})^{-1}. \tag{4.37}$$

6. a) Generation of a new candidate for **B**

   b) Generation of a new candidate for $\mathbf{\Sigma_s}$

   c) At this step a new candidate is generated for $\mathbf{\Sigma}_m$. This new candidate is generated from and inverse-Wishart distribution where the scale matrix parameter of the distribution is modified to account for the presence of $\mathbf{\Sigma}_s$ and its probability is then assessed with Equation (3.14), a Metropolis Hastings is then used to decide to keep the sample or discard it as per procedure explained in chapter §2.3. $\mathbf{\Sigma}_{tot}$ is computed at every iteration using Equation (4.36)

   d) Generate a new candidate for **y**

7. Computation of $p = Pr(\mathbf{y}_l \in AR|\mathbf{x}_l,\mathbf{X},\mathbf{Y})$

This methodology is illustrated in Figure 4.3.

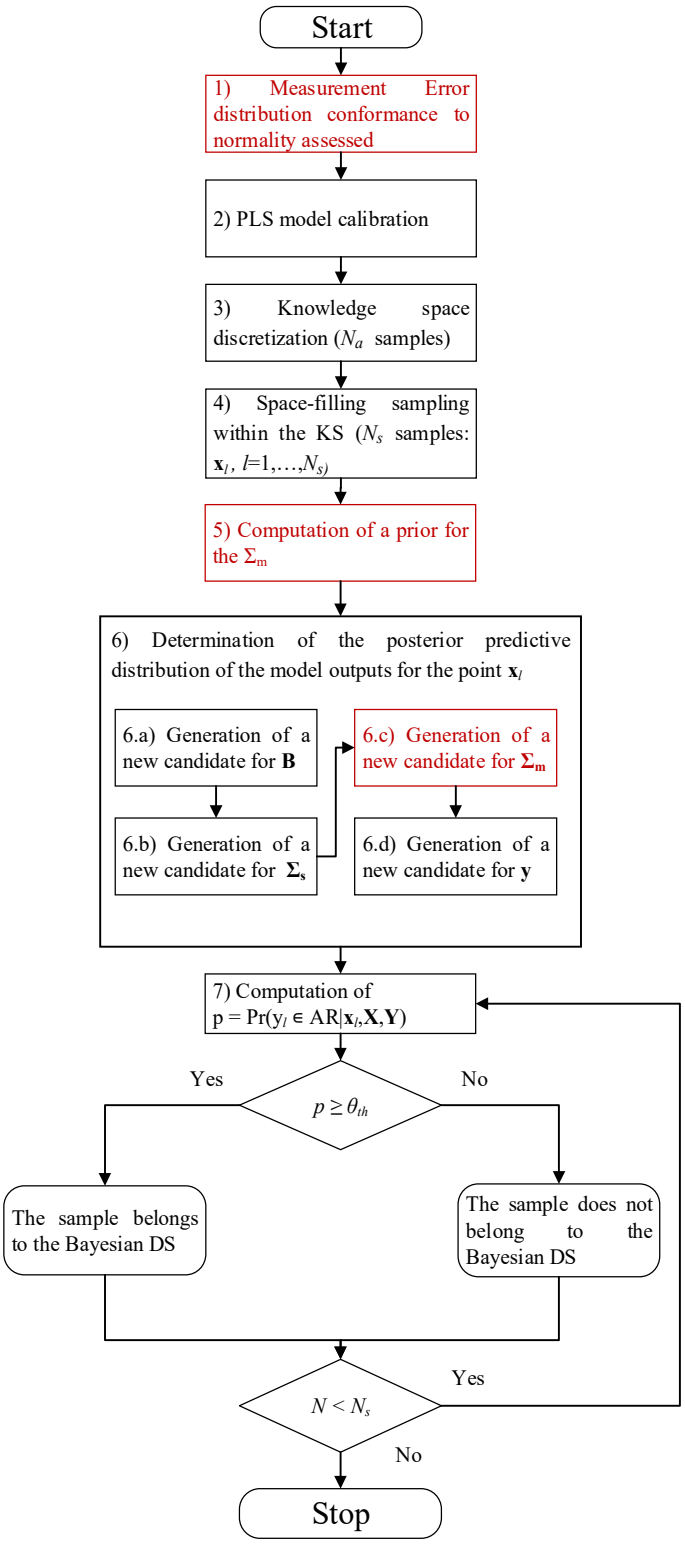

Figura 4.3: *Flowchart for the proposed methodology for the incorporation of uncertainty derived from measurement error in a Bayesian DS determination exercise, parts in red are modification to the original methodology proposed by Bano et al., (2018)*

An illustrative example of the procedure will now follow. The methodology will be thoroughly explained with exemplification of the crucial passages.

4.3.1 Applied methodology: an illustrative example

This illustrative exampled is focussed on the calculation of the posterior predictive distribution for **y** given an **x** not belonging to the original dataset and on the differences between the method proposed by Bano *et al*.,(2018) and the modifications proposed in this thesis.

One hundred random variables have been generated using MATLAB to sample from determined distributions.

The distributions are as follows:

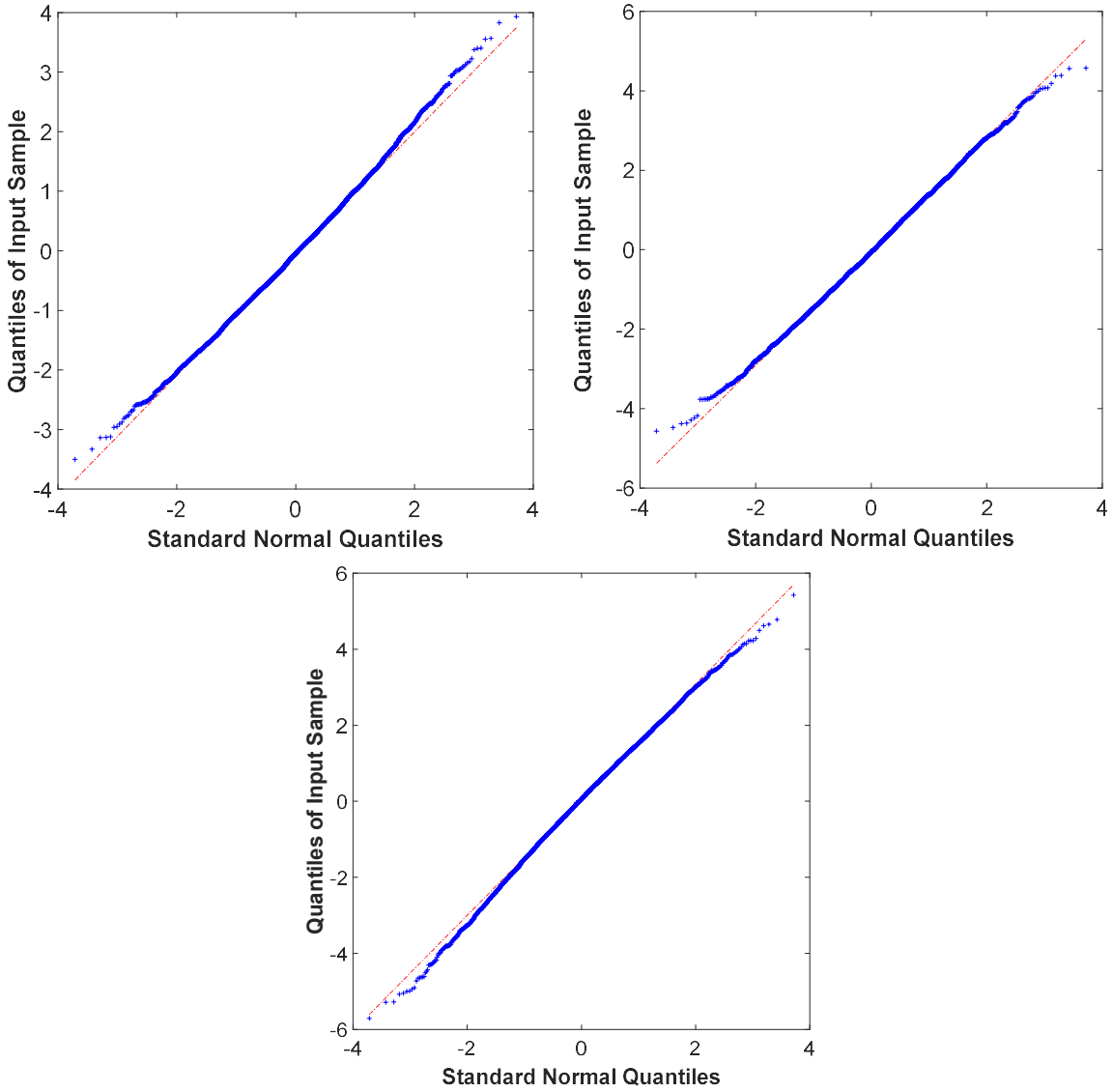

Figure 4.3: *Assessment of normality of the error of measurement of the variables* ***X*** *in Equation (4.34) using the q-q plot method (Carroll et al., 2006)*

$$\begin{cases} \mathbf{x_1} \sim \mathrm{N}(-0.007043, 1.15915) \\ \mathbf{x_2} \sim \mathrm{N}(-0.095056, 1.14323) \\ \mathbf{x_3} \sim \mathrm{N}(0.04298, 0.76840) \end{cases} \tag{4.34}$$

The response has been constructed via MATLAB as:

$$\begin{cases} \mathbf{y}_1 = \mathbf{x}_1 - 5\mathbf{x}_2 + 5\mathbf{x}_3 \\ \mathbf{y}_2 = 2\mathbf{x}_1 - 3\mathbf{x}_2 + 3\mathbf{x}_3 \end{cases} \tag{4.35}$$

The normality of the distribution of errors has been assessed from the inspection of the q-q plots shown in Figure 4.3.

A value for the measurement noise is usually known, given by the investigator that supplies the data. Since data for this case study have been simulated a value for the first guess of the measurement noise has been set as a percentage of the standard deviation of the mean calculated starting from a random point in the dataset adding a random point at the time up to the whole dataset. The tentative first guess is:

$$\mathbf{\Sigma_{\Delta X}} = 0.1 \times \begin{bmatrix} 0.1086 & 0 & 0 \\ 0 & 0.1715 & 0 \\ 0 & 0 & 0.1352 \end{bmatrix} \tag{4.36}$$

The values in Equation (4.36) are less than one percent of the simulated dataset variation, a diminutive measurement error.

A singular value decomposition (SVD) of the error is now carried out; incidentally, for Equation (4.36) and similar completely independent errors the SVD is just the diagonal values ordered by magnitude. The SVD decomposition is used to approximate this [3×3] matrix to a [2×2] matrix (the number of variable in the response dataset **Y)** to be used in the recursive calculation of step b) in chapter §4.3.

After the approximation, the value of the first discarded eigenvalue is used to assess the approximation. After the approximation an objective function has been set using Equation (4.33) and optimized via successive minimizations. The prior parameter for the $\mathbf{\Sigma}_s$ of the model is I^2 the two dimensional identity matrix. The result of the minimizations is the prior that will be used in the subsequent steps of the methodology, i.e.:

$$\Sigma_{M^0} \cong \begin{bmatrix} 0.1199 & 0 \\ 0 & 0.0980 \end{bmatrix} \tag{4.37}$$

Where the approximation sign has been used to imply that the value reported are truncated and not the ones used in the iteration. This problem is a subset of the whole methodology presented in chapter §3.3 if the point at which the PPD is calculated is thought as one of the N_s samples. Once that a prior is obtained it is possible to proceed with step c) of the methodology proposed in the previous section. At each iteration, once all other parameters have been updated a new candidate is generated for $\mathbf{\Sigma}_m$ with the MATLAB *iwishrnd.m* function and assessed with Equation (3.14) reported again here, i.e.:

$$\mathcal{L}(\boldsymbol{\Sigma}_s|\upsilon, \mathbf{G}, \text{data}) = \frac{\boldsymbol{\Sigma}_s^{\frac{\upsilon-n-1}{2}}}{\mathbf{G}^{\frac{\upsilon}{2}}} \mathrm{e}^{\left[-\frac{1}{2}\text{trace}(\mathbf{G}^{-1}\boldsymbol{\Sigma}_s\right]}. \tag{4.38}$$

While the same υ is used, in this step $\mathbf{G}_m$ is different from the one used in the original methodology, $\mathbf{G}_s$, calculated as:

$$\mathbf{G}_s = \begin{bmatrix} \upsilon & 0 \\ 0 & \upsilon \end{bmatrix} + [\mathbf{y} - \mathbf{B}\mathbf{x}]^{\mathrm{T}}[\mathbf{y} - \mathbf{B}\mathbf{x}]. \tag{4.39}$$

In the proposed methodology the accepted value of $\boldsymbol{\Sigma}_s$ at the current step is used to control the dimension of the proposed $\boldsymbol{\Sigma}_m$

$$\mathbf{G}_m = \begin{bmatrix} \upsilon & 0 \\ 0 & \upsilon \end{bmatrix} + [\mathbf{y} - \mathbf{B}\mathbf{x}]^{\mathrm{T}}[\mathbf{y} - \mathbf{B}\mathbf{x}] - \boldsymbol{\Sigma}_s \tag{4.40}$$

The resulting matrix, assessed at the ninety thousandth iteration is:

$$\boldsymbol{\Sigma}_{tot} \cong \left[\boldsymbol{\Sigma}_m \cong \begin{bmatrix} 0.0827 & -0.0643 \\ -0.0643 & 0.0840 \end{bmatrix} + \boldsymbol{\Sigma}_s \cong \begin{bmatrix} 0.0231 & 0.0165 \\ 0.0165 & 0.0247 \end{bmatrix}\right]. \tag{4.41}$$

While the matrix sigma calculated with the original methodology at the same step is:

$$\boldsymbol{\Sigma}_s \cong \begin{bmatrix} 0.0177 & -0.0118 \\ -0.0118 & 0.0189 \end{bmatrix}. \tag{4.42}$$

The resulting PPDs are presented in Figures 4.4 and 4.5. The original methodology resulted in a normal distribution with:

$$\begin{cases} \mathbf{y}_1 \sim \mathrm{N}(0.9924, 8.59529); \\ \mathbf{y}_2 \sim \mathrm{N}(2.080, 9.4363), \end{cases} \tag{4.43}$$

With expected values calculated by direct substitution of (1, 2).

The proposed methodology resulted in two normal distributions with similar mean parameters but higher standard deviation, i.e.:

$$\begin{cases} \mathbf{y}_1 \sim \mathrm{N}(1.097, 9.8326); \\ \mathbf{y}_2 \sim \mathrm{N}(2.118, 10.9158). \end{cases} \tag{4.44}$$

with the same expected values.

In this example the applied methodology to incorporate a method for accounting for the measurement error in the Bayesian DS determination exercise has been shown with in depth explanation of the necessary steps. It is possible, from the inspection of Figures (4.5) and (4.4), to see that the PPD computed with the proposed methodology has greater standard deviation. This visible added uncertainty is due to the incorporation of the measurement error.

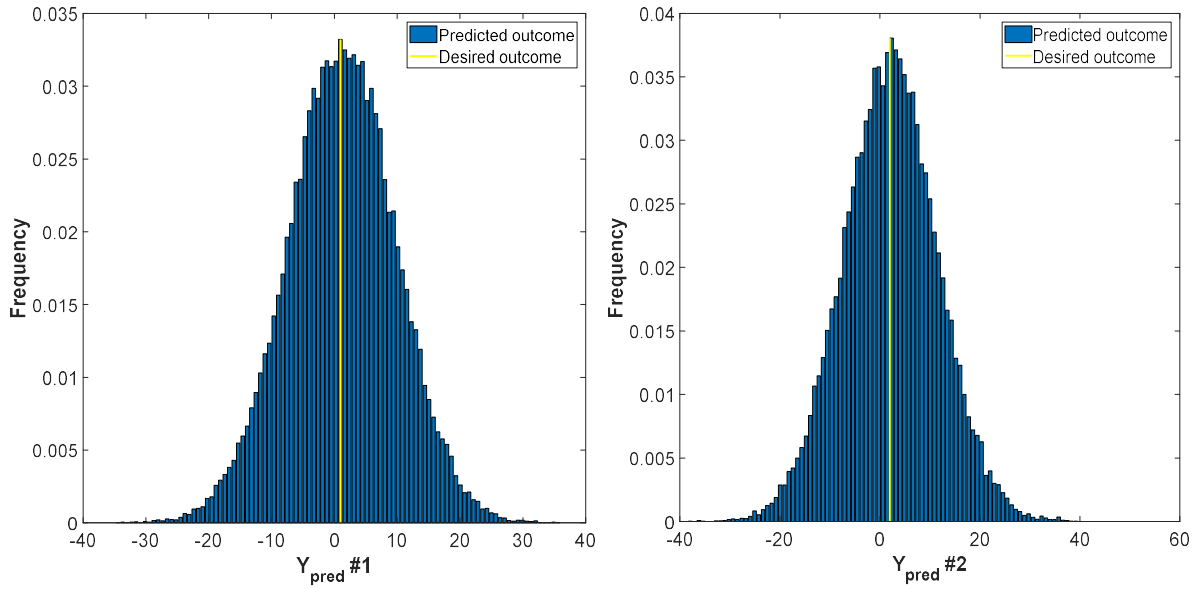

Figure 4.4: *PPD of the value of **Y** calculated with the original methodology the predicted values are very close to the expected outcome. The actual values are reported in Equation (4.43)*

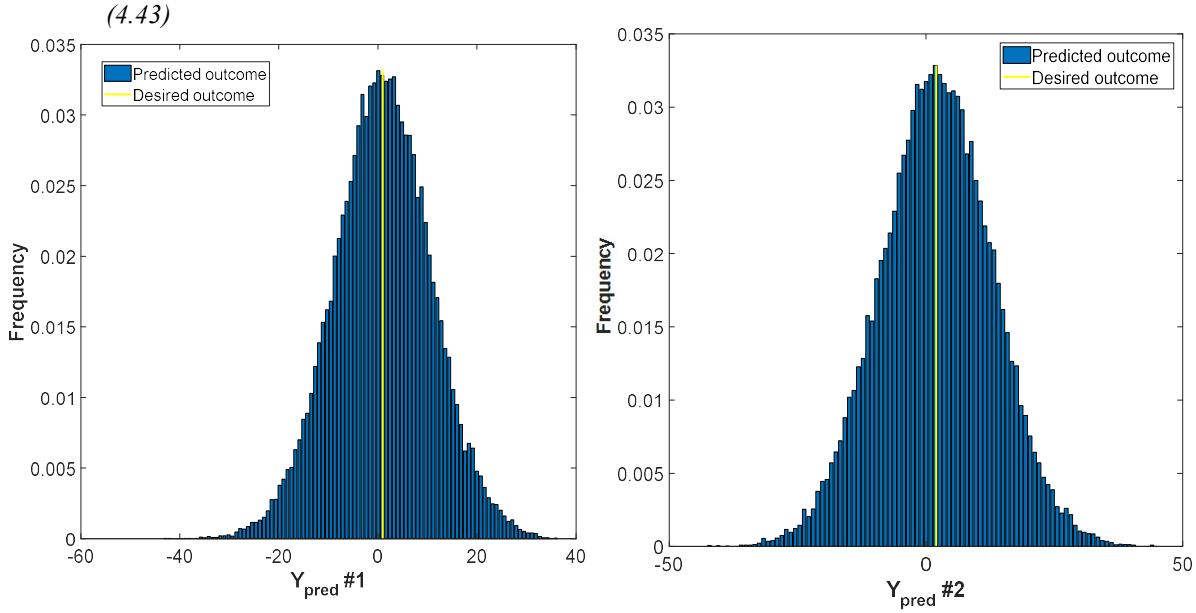

Figure 4.5*: Calculated PPD of the values of **Y** calculated with the modified methodology; the mean is still very close to the expected outcome. The standard deviation is visibly bigger. The actual values are reported in Equation (4.44)*

This result is replicated for all the other points of the discretized KS, and the overall effect is a shrinkage of the resulting DS. Results of the full methodology applied to mathematical and industrial cases will be reported in the next chapter.

Chapter 5

Case studies

In this Chapter, three case studies will be presented showing the completed DS determination exercise and comparing the obtained results to those coming from the application of the original methodology.

5.1 Case Studies

Three further examples of the methodology proposed in chapter §4.3 are presented herein. A mathematical case, chosen for its nonlinearity and correlation, one of the weaknesses of regression methods. An industrial case from a sieve tray analysis of two different API formulations. Lastly, industrial data are analysed in a DS building exercise targeting the compactability ratio of a wet granulation process.

5.1.1 Case #1: Mathematical Example

A highly correlated, nonlinear dataset is used as the first case study. This dataset is generated as reported in an article by Facco *et al.*, (2015). The dataset is composed of one hundred observations, each of five parameters that simulate a historical dataset. The input calibration matrix **M** [100 × 5] collects both the dependant and the independent inputs. The dependant input called $\mathbf{m}_1$ and $\mathbf{m}_2$ to maintain the convention of the original article are sampled from two normal distributions, i.e.:

$$\mathbf{m}_1 \sim \mathrm{N}(41.73, 16.07), \tag{5.1}$$

$$\mathbf{m}_2 \sim \mathrm{N}(11.13, 2.97). \tag{5.2}$$

While the dependent variables are defined as follows:

$$\mathbf{m}_3 = \mathbf{m}_1^2 \tag{5.3}$$

$$\mathbf{m}_4 = \mathbf{m}_2^2 \tag{5.4}$$

$$\mathbf{m}_5 = \mathbf{m}_1\mathbf{m}_2 \tag{5.5}$$

The calibration response dataset is generated by a mathematical formulation, i.e.:

$$\mathbf{y} = k_0 + k_1\mathbf{m}_1 + k_2\mathbf{m}_2 + k_3\mathbf{m}_3 + k_4\mathbf{m}_4 + k_5\mathbf{m}_5 \tag{5.6}$$

Where the parameter vector is equal to:

$$\boldsymbol{k} = [k_0 = -21.0, k_1 = 4.3, k_2 = 0.022, k_3 = -0.0064, k_4 = 1.1, k_5 = -0.12] \tag{5.7}$$

Values have been selected with MATLAB, via a pseudorandom number generator based on Mersenne twister with seed zero.
The characterisation of the dataset is reported in Table 5.1

Table 5.1: *Case study #1: Characterization of the input and output dataset*

Variable	Mean	Standard deviation
$\mathbf{u}_1$	43.708	18.68
$\mathbf{u}_2$	10.91	2.99
$\mathbf{u}_3$	2255.83	1801.50
$\mathbf{u}_4$	127.94	67.79
$\mathbf{u}_5$	481.20	265.68
$\mathbf{y}$	235.74	76.80

5.1.2 Case #2 Sieve analysis data

A historical dataset on a wet granulation process has been analysed with the proposed methodology to identify a probabilistic design space for wet granulation. The original dataset is composed of 38 repetitions of thirteen variables. A subset of the original dataset has been used. The subset is **M** [21 × 7] and is composed of 21 observations at different conditions. The scale of the experiment has been used to create this subset and is constant during the analysis. The variables to investigate have been chosen based on similarity with known systems.
A summary of the used variables is available in Table 5.2. The sieve tray analysis has been processed to obtain the sample d_{50} and d_{90}; the d_{50} has been used as a target for the present analysis.

Table 3.1*: Case study #2: list of the input and response variables.*

ID	Variable name	Units	Symbol
Inputs			
1	Tip speed	[m/s]	u_t
2	Conveying rate	[kg/hr]	F_c
3	PFN	[-]	PFN
4	Froude number	[-]	Fr
5	Torque data	[N/m]	τ
6	Power	[W]	W
7	SME	[J/g]	SME
Response			
R1	Median particle size	[mm]	d_{50}
R2	90-th percentile	[mm]	d_{90}

5.1.3 Case 3# Wet granulation experimental data

This case study considered the exercise of constructing a probabilistic design space for a new high-shear wet granulation process. The experiment data used are available from the work of Oka *et al*., (2015). The historical dataset is composed of 27 observations of three input characteristics and one response and is based on a full factorial DoE (3 × 3 × 3). For the design space construction exercise three input variables are considered, and their impact on the particle size distribution of the granulate has been studied. In Table 5.3 a summary of the outputs and inputs is presented.

Table 5.3*: Case study#3: list of the input and response variables (data from (Oka et al., 2015))*

ID	Variable name	Units	Symbol
Inputs			
1	Liquid to solid ratio	[-]	*L/S*
2	Impeller speed	[rpm]	ν_{imp}
3	Wet massing time	[min]	τ_{wet}
Response			
R1	Median particle size	[μm]	d_{50}

5.1.4 Wet granulation: process description

Wet Granulation is a particle size enlargement process employed in a variety of industries to convert fine powders into granules. It involves the use of a liquid binder for particle agglomeration and consolidation. The liquid binder is distributed on the powdered particles while a mixer blends the powder. The wetted particles will form nuclei that are then deformed and densified during the collisions with the vessel wall, the agitator's blades or other particles. The densification process will force the liquid out of the particles allowing coalescence and

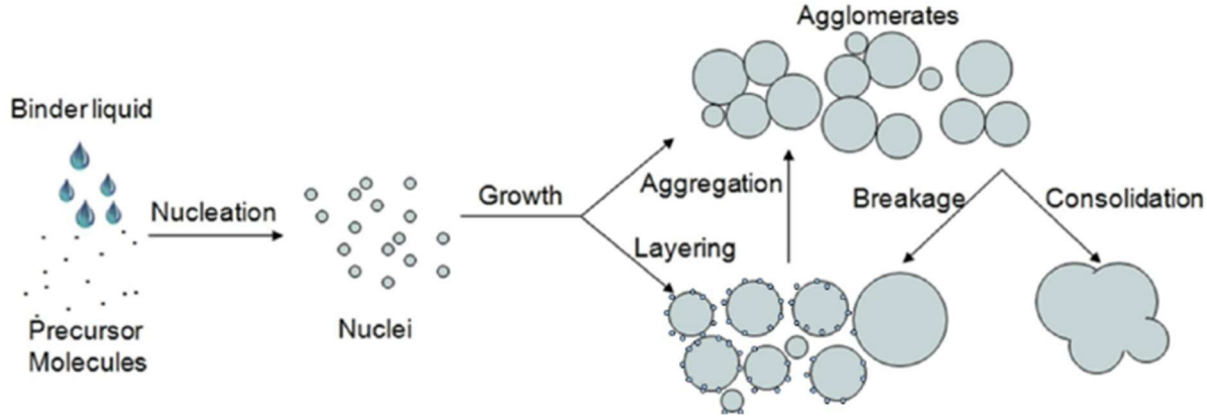

Figure 5.1: *Schematic process of a wet granulation (Kumar, et al., 2014)*

binding with other particles to form larger aggregate. Breakage also occurs during the process when the collision force exceeds a critical value. In Figure 5.1 a schematic representation of the steps of wet granulation are presented.

In the first steps of the wet granulation process, wetting and nucleation, liquid/solid ratio and powder wettability play a crucial role. These parameters have been linked to a broader distribution of nuclei size (Lee, 2012), that, in turn, leads to wider distribution of granule size in the final product, as the process usually retains some sort of "memory" of this step (Hapgood *et al*., 2007). At this stage, nucleation occurs, usually in one of two primary modes depending on the relative dimension of the droplet used to wet the powder blend and the principal particles of the blend. Either the droplets are smaller and distribute on the surface, allowing the wet particle to coalesce with other dry particles, giving a more agglomerated nucleus, or the powder is smaller, and immersion mode nucleation occurs, characterised by nuclei of more compact nature (Lee, 2012). In this stage mixing and thus impeller tip speed has a critical role; one of the most beneficial regimes of operation is the Mechanical dispersion regime, thus defined in Hapgood *et al*., (2003), where powder mixing dominates the liquid dispersion, that depends on the mechanical mixing and agitations only. In the mechanical dispersion regime, nuclei will have a uniform size distribution.

After wetting and nucleation, a further step called granule growth and consolidation takes place. This step is of crucial importance as the final granule properties (size and density) will be decided at this stage. The particle growth is influenced chiefly by the wet massing time. The

two most common types of granule growth that generally occur are steady growth and induction growth. Steady growth is described as when the rate of growth is approximately constant. Induction type of growth is described as when there is an induction time with no growth followed by rapid growth (Lee, 2012). Consolidation takes place due to compressive and shear forces pushing the particles together, this reduces size and density of the resulting particle and influences its porosity. Consolidation influences the growth process, particularly on induced growth, and depends primarily on the solid to liquid ratio (Tardos *et al*., 1997).
Granules breakage and attrition is the last stage in the wet granulation process. Granules breakage refers to the phenomena where wet granules break due to high shear forces exhibited in granulators. Breakage of wet granules can control the maximum and final granules sizes, so a careful balance needs to be struck between the need to agitate the powder blend to obtain mechanical dispersion regime and the added width in the final granule dispersion.

5.2 Results and discussion

In this section results of the DS building exercise for the three case studies presented in the previous section are presented with comparison to the ones obtained with the original methodology will be made. For each case study apart from the first two scenarios are simulated, one with highly noisy measurement, with a variance of the noise of 10% of the mean and a normal error scenario with variance of the noise of 1% of the sample mean.

5.2.1 Results for Case study #1

The assessment of the adherence to the normal distribution of the measurement error is carried out with a q q-plot. A PLS model is then built using the calibration dataset, the number of latent variables to retain in the test is evaluated using an eigenvalue greater than one rule. A number A=2 of latent variables is chosen explaining 95.1% of the total variation on **M**. Note that since A > rank(**y**), a null space exists. The problem of including the uncertainty derived from noisy measurements on a DS building exercise with y_{des}=283 is addressed, with an acceptance probability threshold of 90%. The ability of the proposed methodology to include the measurement error is tested. The real DS of the process is then derived from first principles, in this case, Equation (5.6), and is reported as a black line in Figure (5.3). In the original article by Facco *et al*., (2015) interesting remarks on the representativeness of the PLS in the case that the input is strongly nonlinear are reported. A prior has been generated for the analysis, in this case, a value close to a percentage of the mean has been used in the procedure proposed in chapter §4.3 to compute a prior for the measurement error.
In the case of Figure (5.3) the input matrix is:

$$\boldsymbol{\Sigma}_{\boldsymbol{\Delta X}} = \text{diag}\,[5\ 1\ 315\ 55\ 130] \tag{5.8}$$

Where diag (n) means a matrix with the term n as the diagonal and zero in all off-diagonal terms. The used methodology resulted in a value for $\boldsymbol{\Sigma_M}$, i.e.:

$$\boldsymbol{\Sigma_m} \approx 0.2803. \tag{5.9}$$

In both images the points included in the DS are reported as green diamonds, while the points below the threshold are reported as red diamonds. A thick blue line delimits the boundary of the knowledge space and the calibration dataset is reported as blue circles. In Figure (5.3) the reduced DS is bounded by a thin blue line for a better visualisation. The thin blue line is computed using the MATLAB boundary command using the "shrink-factor" option to control the tightness of the boundary.

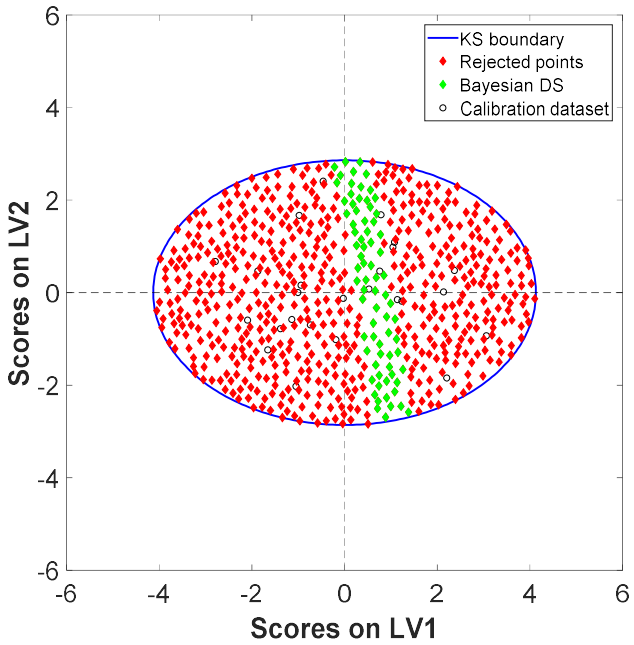

Figure 5.2: *Case study#1; base case, the uncertainty linked to the error has not been accounted for and values near the edge could be below the required 90% assurance. The green diamonds are points with probability greater than the threshold, red*

5.2.2 Results for Case study #2

The assessment of the normality of the measurement uncertainty distribution is carried out with a qq-plot. In this case study the problem at hand is the DS building exercise for a granulate with a median particle size of 0.4099 mm starting from historical data of a known process. The ability the proposed methodology to identify a subset of the original knowledge space within which the probability that the products will be on target or superior, while accounting for measurement error in the calibration dataset, is assessed. A probability of 90% is used as a threshold for acceptance. A PLS model is built using the calibration dataset, the number of latent variables to use is evaluated using an eigenvalue greater than one rule and a number A=2

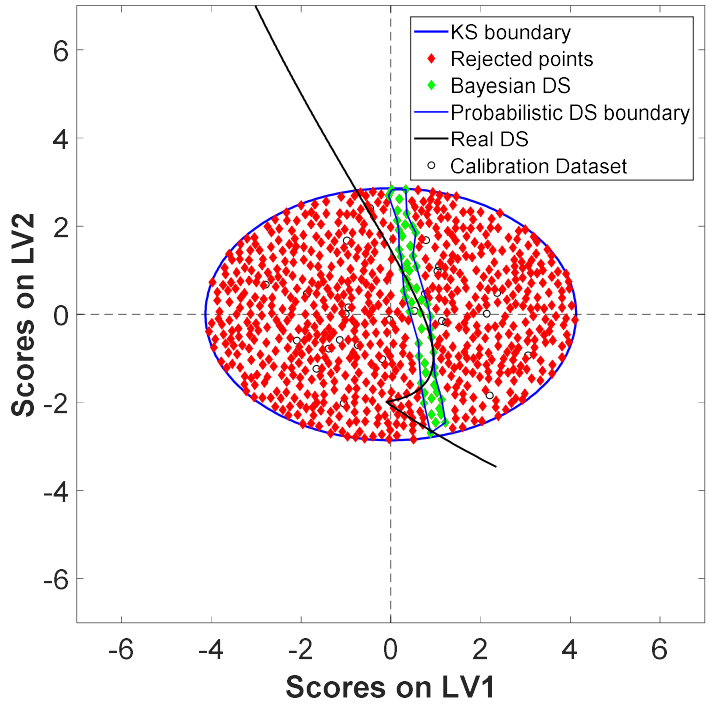

Figure 5.3: *Case study #1: A probabilistic DS constructed accounting for measurement uncertainty. Comparison with Figure (5.2) shows that the DS has shrunk as an effect of the lowered probability of attaining the required specification. The black line is the real DS calculated by first principles, the black circles are the calibration dataset, the thick blue line the knowledge space boundary and the thin blue line is a boundary computed with the MATLAB boundary command with a shrink-factor of 0.1.*

of latent variables is chosen explaining 82.24% of the total variation on **M**. A prior has been generated for the analysis, a percentage of the mean has been used in the procedure proposed in chapter §4.3 to compute a prior for the measurement error.
In the case of Figure (5.5) the input matrix is:

$$\boldsymbol{\Sigma_{\Delta X}} = \text{diag}(0.1 \times [\boldsymbol{\mu_U}]) \tag{5.10}$$

While for the case of Figure (5.6) the input matrix is:

$$\boldsymbol{\Sigma_{\Delta X}} = \text{diag}(0.01 \times [\mu_U]) \tag{5.11}$$

In both Equations (5.10) and (5.11) diag (*n*) means a matrix with *n* in the diagonal and zeros in the off-diagonal. The resulting priors for $\boldsymbol{\Sigma_M}$ are:

$$\boldsymbol{\Sigma_m} \approx 0.154, \tag{5.12}$$

for the case of Equation (5.10) and:

$$\boldsymbol{\Sigma_m} \approx 0.0175, \tag{5.13}$$

for Equation (5.11). A point of known d_{50} equal to the desired one has been set apart and not used in the simulation to validate the results. This point is shown in Figure (5.4), (5.5) and (5.6) with a black square.

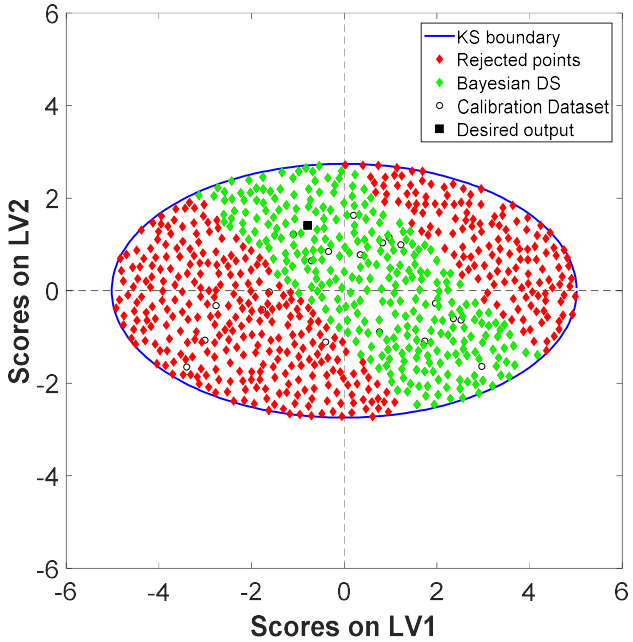

Figure 5.4: *Case study #2: Base case, the green diamond are the point whose probability of attaining the target d_{50} is equal or above 90%. The thick blue line is the knowledge space boundary. The black square is the validation point, a combination of input giving a d_{50} of 0.4099 mm*

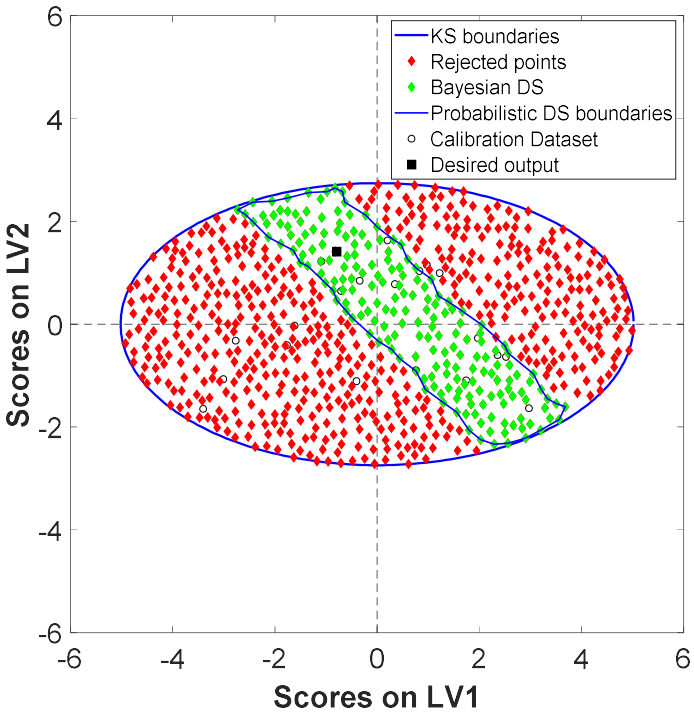

Figure 5.5: *Case study #2: In this figure the green diamonds are point that have a probability of attaining a d_{50} of 0.4099 mm equal or greater than 90%. Red diamonds are rejected points and the black square is the validation point, with known d_{50} equal to the desired one. The assumed measurement error is equal to 10% of the mean.*

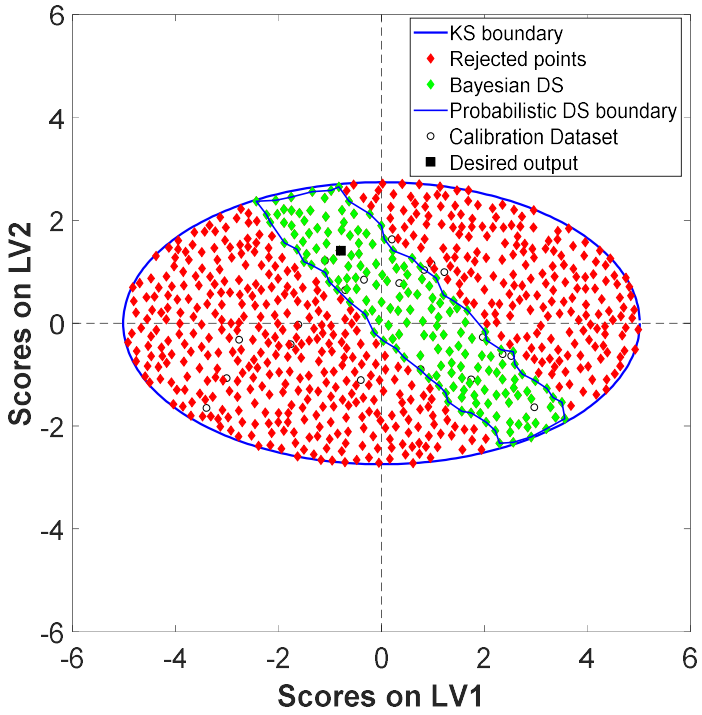

Figure 5.6: *Case study #2: In this figure the green diamonds are point that have a probability of attaining a d_{50} of 0.4099 mm equal or greater than 90%. Red diamonds are rejected points and the black square is the validation point, with known d_{50} equal to the desired one. The assumed measurement error is equal to 1% of the mean.*

5.2.3 Results for Case study #3

The adherence of the measurement error to the normal distribution is assessed using a qq-plot. In this case study the problem at hand is the DS building exercise for a blend of active pharmaceutical ingredient (API) plus excipient starting from data available in the article from Oka *et al.*, (2015). The desired output is a powder with a d_{50} of 1129 μm. A probability of 90% is used as a threshold for acceptance. A validation point has been identified with a response equal to the desired value; this point has been removed from the calibration dataset and used for confirmatory purposes. A PLS model is built using the calibration dataset, the number of latent variables to use is evaluated using an eigenvalue greater than one rule and a number A=2 of latent variables is chosen explaining 66.7% of the total variation on **M**. This value is indeed low compared to the guidelines of chapter §3.4 and in fact the desired output, indicated in Figures (5.6), (5.7) and (5.8) with a black square, is closer to the edge of failure than in the other case studies, the simple structure of the data though make it possible to analyse the process nevertheless. Note that since A > rank(**y**), a null space exists.

The difference in the shape of the knowledge space presented with this PLS model is due to the simple structure of the data and the fact that every variable has the same weight in the representation of the latent structure, as mentioned in chapter §2.2

A prior has been generated for the analysis, in this case, a percentage of the mean has been used in the procedure proposed in chapter §4.3 to compute a prior for the measurement error. In the case of Figure (5.7) the input matrix is:

$$\boldsymbol{\Sigma_{\Delta X}} \approx \begin{bmatrix} 0.0733 & 0 & 0 \\ 0 & 27.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \tag{5.14}$$

While for the case of Figure (5.4) the input matrix is:

$$\boldsymbol{\Sigma_{\Delta X}} \approx \begin{bmatrix} 0.0073 & 0 & 0 \\ 0 & 2.75 & 0 \\ 0 & 0 & 0.05 \end{bmatrix}. \tag{5.15}$$

The resulting priors for $\boldsymbol{\Sigma_M}$ are:

$$\boldsymbol{\Sigma_m} \approx 1.161, \tag{5.16}$$

for the case of Equation (5.14) and:

$$\boldsymbol{\Sigma_m} \approx 0.208, \tag{5.17}$$

for Equation (5.15).

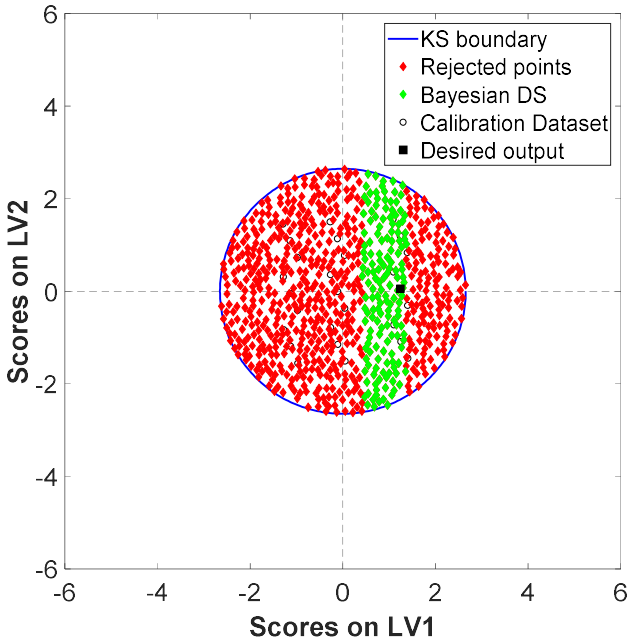

Figure 5.6: *Case study #3: Base case, the measurement error has not been considered. The green diamonds are points whose probability of attaining the target granule median size is equal to or greater than 90%. Red diamonds are points whose probability is lower and the thick blue line is the boundary of the knowledge space. The black square is the validation point whose median diameter equals the desired one. It stands close to the edge of failure due to poor representativeness of the PLS method for this system.*

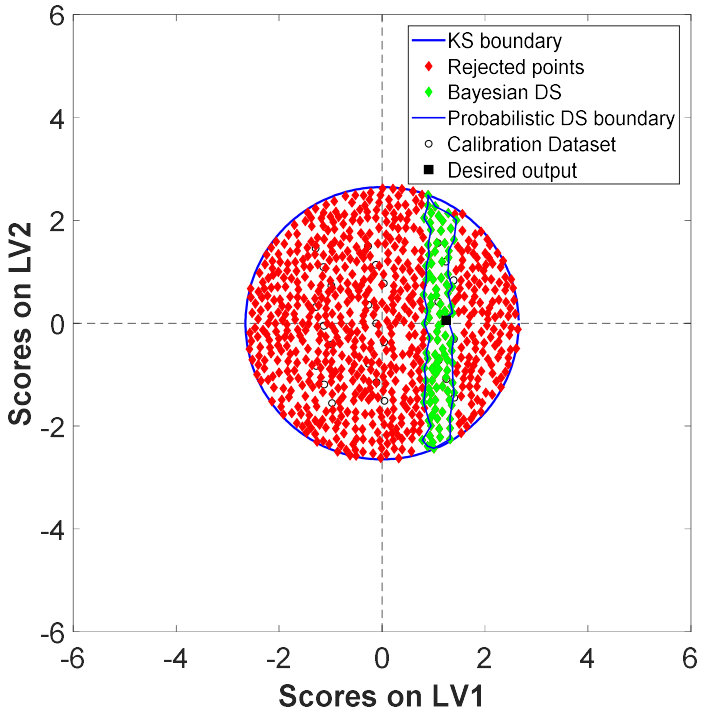

Figure 5.7: *Case study #3: Green diamonds are points with probability greater than 90% of giving a desired product, red ones have lower probability and the black square is the validation point. The thick blue line is the boundary to the knowledge space. In the establishment of this design space a measurement error equal to the ten percent of the mean has been assumed, as theorized the point is still inside the probabilistic boundaries notwithstanding the representativeness of the dataset being in the lower acceptable region.*

5.3 Final remarks

As seen in the case studies the region of the accepted points is shrunk substantially. The shrinking is due to the increased degree of variance removed with the methodology proposed to the PPD of the **y** value. The selection algorithm at step five of the original methodology (chapter §3.3) and step seven of the proposed methodology (chapter §4.3), uses the integral of a portion of the area below the PDF to assess the probability of the calculated **y** to give the desired response.

The values of the error have less impact than suspected, and verification of sensitivity of the methodology to the starting value of $\boldsymbol{\Sigma}_{\boldsymbol{m}}$ in Equation (4.36) could be an interesting point of further investigation.

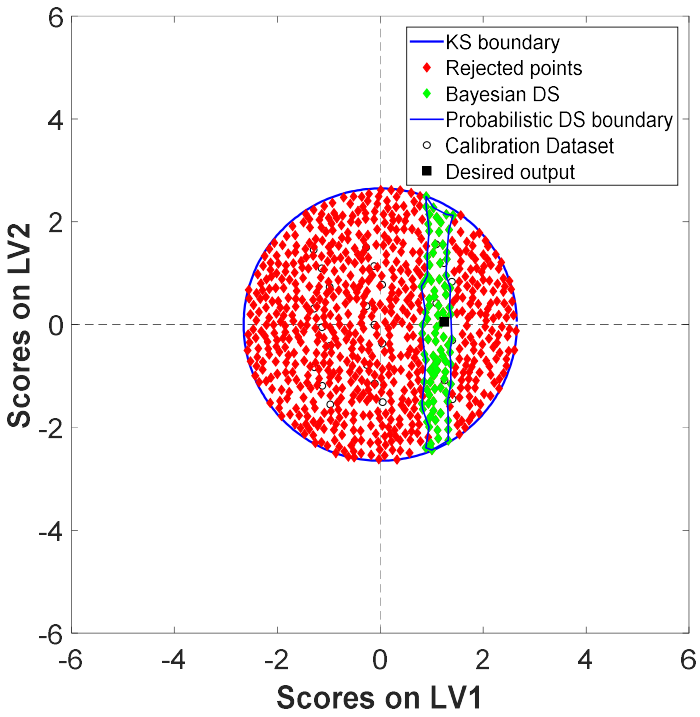

Figure 5.8: *Case Study #3: Green diamonds are points with probability greater than 90% of giving a desired product, red ones have lower probability and the black square is the validation point. The thick blue line is the boundary to the knowledge space. In the establishment of this design space a measurement error equal to the one percent of the mean has been assumed, as theorized the point is still inside the probabilistic boundaries notwithstanding the representativeness of the dataset being in the lower acceptable region.*

The representativeness of the PLS retains its crucial importance from the original methodology, as seen in chapter §3.4. The smaller the amount of cumulative **X**-variability explained by the latent variables, the greater the error on the projection of the KS onto the latent space, and the worse the representation of the original input space by the latent space. In other words, there is a risk of not considering areas of the knowledge space that may be part of the design space.

Conclusions

The objective of this Thesis was the incorporation of measurement uncertainty in the design space determination exercise for a new pharmaceutical product with a risk-based Bayesian posterior predictive approach, while simultaneously reducing the problem dimensionality using PLS modelling. The joint posterior predictive probability of each sampling point was obtained with a multivariate Bayesian linear regression model, and the probability that product quality will meet its specifications for the given point was computed. By performing this analysis for each of the sampling points of the historical knowledge space, the probabilistic (or Bayesian) DS of the product under investigation was determined. To compute the joint posterior predictive probability of each point a Bayesian framework was built. This framework is based on the one coded proposed in the study of Bano *et al*. (2018).

Results are then shown in an intuitive way by two-dimensional latent structure projection. The obtained probabilistic DS represents a way to demonstrate (i.e., quantify) the level of "assurance of quality" the manufacturer can guarantee for a given product, as advocated by the pharmaceutical regulatory agencies. This "assurance of quality" is further refined with the addition of the possibility to account for the uncertainty derived from the measurement error (e.g.: noise) in the calibration input dataset.

With respect to the original methodology, all the cases presented a lower number of accepted points, due to the additional source of uncertainty that has been included. Although all the case studies were conducted on a single vector of responses, the methodology is not limited to a one-dimensional output, despite the increase in computational complexity given that it could require more iterations per chain to converge.

Some points of interest for future studies could be pointed out. For example, an optimization of the acceptance criterion for the MCMC, in order to get results in shorter time, this can be generalized to a further optimization of the whole code to run in online cloud computing or parallel computing. A second possible area of study is the addition of other error models, maintaining the general nature of the methodology the addition of further models should increase the assurance of quality and the possibility to have this design space determination software as customizable as possible. Another important area is the sensitivity analysis to variation to the initial value of the guess for the error of measurement E_m covariance in Equation (4.36). While the effect of the error on the input matrix $\mathbf{X}$ has been thoroughly examined in the context of this Thesis, the effect of that value on the final DS could be of interest.

List of symbols

A	=	Number of selected latent variable (or principal components, depending on the context)
B	=	Linear regression model parameters
$\widehat{\mathbf{B}}_{MLE}$	=	Maximum likelihood estimate of the parameters **B**
cov (x)	=	Covariance of x
c	=	PLS **Y**-weigth vector
$\mathbf{C}^*$	=	PLS **Y**-weigth Matrix
d_{50}	=	Median value of the particle size distribution [μm] or [mm]
d_{90}	=	90-th percentile of the particle size distribution [mm]
d_i	=	Desirability function
D_{roll}	=	Roll diameter [m]
E	=	Linear regression model residuals
e^2_{ija}	=	Residual of variable *j* at observation *i* with *a* principal components
$\mathbf{E}_m$	=	Measurement noise error matrix
$\mathbf{E}_s$	=	Structural error matrix
$\mathbf{E}_{tot}$	=	Total error matrix
$f(\boldsymbol{x}) \otimes g(\boldsymbol{x})$	=	Convolution between function *f* and function *g*
F	=	PLS **Y** - residuals
F	=	Fourier transform
F^{-1}	=	Inverse Fourier transform
F(A, I-A, α)	=	95% percentile of the F distribution with A degrees of freedom in the numerator and (i-A) degrees of freedom in the denominator
F_{roll}	=	Roller pressure force [kN]
F_s	=	Spring-back factor [-]
$\mathbf{G}_m$	=	Scale matrix for the measurement covariance generation via Inverse Wishart distribution
$\mathbf{G}_s$	=	Scale matrix for the structural covariance generation via Inverse Wishart distribution
$g(\mathbf{y}\|\mathbf{x}, \text{data})$	=	Posterior predictive distribution of **y** given the observation **x** and the data

I	=	Rows (observations) of the historical dataset
J	=	Columns (variables) of the historical input dataset
$\mathbf{J_t}(\boldsymbol{\theta}^* \vert \boldsymbol{\theta}^{t-1})$	=	Markov Chain jumping distribution
K	=	Columns (variables) of the historical output dataset
k	=	Compressibility factor
$\mathcal{L}(\mathbf{B}, \boldsymbol{\Sigma_B} \vert \mathbf{Y})$	=	Likelihood function
L(**P**)	=	Lagrange function of **P**
L_i	=	Lower acceptable limit
L/S	=	Liquid to solid ratio
M	=	Input calibration matrix
$\mathrm{N}(\mathbf{BX}, \boldsymbol{\sigma^2 I^N})$	=	matrix-variate normal distribution with mean **XB** and covariance equal to $\boldsymbol{\sigma^2 I^N}$
N_a	=	Total number of discretization samples
N_s	=	Selected number of discretization samples
P	=	PLS (PCA) model loading matrix
$\mathbf{p}_n$	=	PLS model vector of the loadings for the n-th observation
$\mathrm{P}(\mathbf{y} \vert \boldsymbol{\theta})$	=	Probability of an occurrence **y** given the parameters $\boldsymbol{\theta}$
$\mathrm{P}(\mathbf{y} \in \mathrm{AR} \vert \mathbf{x}, \mathbf{X}, \mathbf{Y})$	=	Probability to be inside the acceptable range conditional on the vector **x** and the historical dataset **X**, **Y**
Q	=	Y – scores matrix for the PLS
Q^2	=	Goodness of fit statistic
R^2	=	Goodness of prediction statistic
${}_aR_j^2$	=	Sum of square error for variable j with a principal components
$s_i^2(\mathbf{A})$	=	i-th singular value of matrix **A**
s_{ta}^2	=	Score variance with respect of principal component a
s_{roll}	=	Roller width [m]
T	=	PLS (PCA) model scores matrix
$\mathbf{t}_n$	=	PLS model vector of the loadings for the n-th observation
${T^2}_i$	=	Hotellings's T^2 statistic for observation i
${T^2}_{\mathrm{lim}}$	=	Confidence limit for the Hotelling's T^2
T_i	=	Target state
$\mathbf{T_t}(\boldsymbol{\theta_t} \vert \boldsymbol{\theta_{t-1}})$	=	MCMC transitional distribution
U	=	PLS **Y**-loadings

V_n	=	Mean voltage of the noise
V_s	=	Mean voltage of the signal
$\mathbf{w}$	=	PLS **X**-weight vector
$\mathbf{W}$	=	Matrix of noisy measurement of the true quantity *x*
$\mathbf{W}^*$	=	PLS **X**-weight matrix
$\mathbf{X}$	=	Historical dataset
$\mathbf{X}_{val}$	=	Validation dataset
$\mathbf{X}_{train}$	=	Training dataset
$\mathbf{x}_j^*$	=	A subset of the *j*-th column vector of **X**
$\mathbf{Y}$	=	Historical output dataset
$\mathbf{y} \sim N(0, \boldsymbol{\Sigma})$	=	**y** normally distributed with zero mean and **Σ** covariance matrix
$\mathbf{y}_{des}$	=	Desired quality target

Greek letters

α	=	Scale factor for the canonical form of the normal distribution
γ_{EFR}	=	Effective friction angle [rad]
γ_{FR}	=	Friction angle [rad]
$\boldsymbol{\Delta X}$	=	Matrix of the random measurement error
$\boldsymbol{\eta}$	=	Information vector for the canonical form
θ_{th}	=	Probability threshold
$\boldsymbol{\theta}_{gibbs}$	=	Gibbs sampler parameter vector
$\boldsymbol{\Lambda}$	=	Fisher information matrix for the canonical form
λ	=	Lagrange multiplier
μ	=	Mean
ν_{imp}	=	Impeller speed [rpm]
ν_{roll}	=	Roller speed [rpm]
$\hat{\sigma}^2_{MLE}$	=	Maximum likelihood estimate of the variance
$\boldsymbol{\Sigma}$	=	Variance/covariance matrix
$\boldsymbol{\Sigma}_{ab}$	=	Covariance between matrix **A** and **B**
$\boldsymbol{\Sigma}_{\Delta X}$	=	Covariance of the measurement noise
$\boldsymbol{\Sigma}_{m}$	=	Covariance of the uncertainty due to measurement noise
$\boldsymbol{\Sigma}_{s}$	=	Covariance of the structural uncertainty

$\boldsymbol{\Sigma}_{tot}$	=	Covariance of the total uncertainty
τ_{wet}	=	Wet massing time
υ	=	Degree of freedom for a Wishart or Inverse Wishart distribution
ϕ_s	=	Infravoid fraction
$\mathbf{x} \in \boldsymbol{\chi}$	=	The vector **x** is in a "sweet spot" of the overlapping mean responses approach

Acronyms

API	=	Active Pharmaceutical Ingredient
AR	=	Acceptable Range
CMC	=	Chemistry, Manufacturing and Control
CPP	=	Critical Process Parameter
CQA	=	Critical Quality Attribute
DOE	=	Design Of Experiment
DS	=	Design Space
EMA	=	European Medicine Agency
FDA	=	Food and Drug Administration
GMP	=	Good Manufacturing Practice
ICH	=	International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use
ISPE	=	International Society for Pharmaceutical Engineering
MCMC	=	Markov Chain Monte Carlo
NIR	=	Near Infra-Red
OMR	=	Overlapping Mean Responses
PAT	=	Process Analytical Technology
PAR	=	Proven Acceptable Range
PCA	=	Principal Components Analysis
PDF	=	Probability density function
PLS	=	Partial Least Square
PSE	=	Process System Engineering
PPD	=	Posterior Predictive Distribution
PQLI	=	Product Quality Lifecycle Implementation
PQS	=	Pharmaceutical Quality System
QbD	=	Quality by Design
QbT	=	Quality by Testing
QTPP	=	Quality target Product Profile
SNR_{dB}	=	Signal to Noise Ratio [dB]

Rpm	=	Revolutions Per Minute [1/min]

Appendix

Appendix A: Code

In this appendix bits of the code written for this work are reported, This is the code to approximate and compute the $\mathbf{E}_{m}$ as for step 5) of the methodology of chapter §4.3

```
[LOAD1,ECMdec,LOAD2]=svd(sigma_X);%
tested=diag(ECMdec);
%Eigenvalue greater than one to test the approximation
test=tested(m+1)> 0.95;
if test==1
    disp('first discarded eigenvalue = ');disp(tested(m+1))
    disp('not safe to approximate the covariance')
    return
end
LOAD21=LOAD1(:,1:m);ECMdec2=ECMdec(1:m,1:m);
LOAD22=LOAD2(1:m,:);
f=@(S_X_approx) sum(sum(((sigma_X))-(LOAD21*S_X_approx*LOAD22)).^2 );%
S_X_approx=fminsearch(@(S_X_approx)f(S_X_approx),ECMdec2);
S_TOT= sigma_S+sigma_M;
S_M=pinv( pinv(S_X_approx) + pinv(S_TOT));
f=@(S_M) sum(sum((sigma_S+S_M)-(sigma_S+ inv(inv(S_X_approx)+
inv(sigma_S+S_M)))).^2 );
S_M=fminsearch(@(S_M)f(S_M),ECMdec2);
sigma_M=diag(S_M).*eye(m);
```

In this section, sigma_X is the assumed $\mathbf{\Sigma}_{\Delta X}$, sigma_S is the imposed covariance of the structural error and sigma_M is the error of measurement covariance. Both quantity are set to the identity matrix by default.
In the following section a candidate for the measurement error covariance is generated from an Inverse Wishart distribution and its probability is assessed with the function "*full_conditional_sigma_M*" reported in the next page.

```
if t ==1
    f0 = m ;
    if size(sigma_M)~=1
        sigma_M=Spd_Mat(sigma_M);
    end
    sigma_M_new = iwishrnd(sigma_M,f0);
else
    if size(Gn_M_inv)~= 1
       Gn_M_inv=Spd_Mat(Gn_M_inv);
    end
    sigma_M_new = iwishrnd(Gn_M_inv.*0.2,fn_M);
end
```

```
[pdf_sigma_M_new,fn_M_new,Gn_M_inv_new] = full_conditional_sigma_M
(X,Y,sigma_M_new,theta,sigma_S) ; % full conditional of the proposal
sigma
[pdf_sigma_M,fn_M,Gn_M_inv] =
full_conditional_sigma_M(X,Y,sigma_M,theta,sigma_S); % full conditional
of sigma
r_sigma_M = min (pdf_sigma_M_new/pdf_sigma_M,1);
u_sigma_M = rand; % Seed from U(0,1)

if u_sigma_M < r_sigma_M
   sigma_M = sigma_M_new;
   sigma_M_posterior_samples(:,:,t) = sigma_M;
 pdf_sigma_M_posterior_samples =
[pdf_sigma_M_posterior_samples,pdf_sigma_M_new];
   acceptance_rate_M = [acceptance_rate_M,1];
else
   sigma_M_posterior_samples(:,:,t) = sigma_M;
    pdf_sigma_M_posterior_samples = [pdf_sigma_M_posterior_samples,
pdf_sigma_M];
end
sigma_posterior_samples(:,:,t) = sigma_M+sigma_S;
```

The function Spd_Mat assures that the input matrix is singular positive definite.
Here the code for the full conditional distribution of the covariance of the measurement uncertainty is reported, using this function in the code reported above step 6.c) of the proposed methodology is concluded.

```
function [full_cond_sigma_M,fn_M,Gn_M_inv] =
full_conditional_sigma_M(X,Y,sigma_M,theta,sigma_S)
%-------------------------------------------------------------------------
marcoc
% Full_conditional distribution of the covariance of the measurement
% uncertainty
%
%                              Marco Cattaldo
%
%
% CAPE-Lab - University of Padova (Italy)
%
%-------------------------------------------------------------------------
---

% Dimensions

[m,m] = size (sigma_M); % dimension of the covariance of the residuals
[n,k] = size(X);        % n = number of calibration samples; k = number
of input variables
[n,m] = size(Y);        % n = number of calibration samples; m = number
of response variables

theta = reshape(theta,k,m);  %reshape theta (k*m)

% Initial values of the pdf parameters
f0 = m ;   % to be tuned : it must be >= m
G0_inv = f0.* eye(m,m);
```

```
% Values of the parameters of the posterior pdf

fn_M = f0+ n;
ghi = (Y-X*theta)'*(Y-X*theta);

%Y-Theta*X-Sigma_S=Sigma_M
Gn_M_inv = G0_inv+ (Y-X*theta)'*(Y-X*theta)-(sigma_S);

full_cond_sigma_M =
det(Gn_M_inv)^(fn_M/2)/(det(sigma_M)^((fn_M+m+1)/2))*exp(-
1/2*trace(sigma_M*Gn_M_inv));
end
```

Appendix B: Error in Variables model

The error in variables model is a model to assess noisy measurement in the input and output calibration dataset. The Error in variables model consists in a minimization of every error term that is inserted in the model. While error in variable models have been approached mostly from the frequentist perspective, not much guidance is available on their Bayesian analysis (Lira and Grientschnig, 2017).

The error in variables model was subject to investigation for a formulation of another methodology. In addition to the Methodology proposed in chapter §4.3 this one could process a nonlinear relation while accounting for the measurement error directly in the parameter selection step of the methodology without having to set up a MCMC for it.

The Error in variables method can be formulated in a way similar to the classic measurement error shown in chapter §4.1, i.e.:

$$\mathbf{y} = (\mathbf{X} - \mathbf{E}_X)\mathbf{B} + \mathbf{e}_y, \tag{B.1}$$

where $\mathbf{y}$ is the true output, $\mathbf{A}$ the true matrix of inputs, $\mathbf{B}$ the regression parameters and $\mathbf{E}_A$ and $\mathbf{e}_y$ are respectively the input measurement error and the output measurement error.

Furthermore, while not assuming any knowledge on the $\mathbf{y}$ and $\mathbf{X}$ we can nevertheless assume that the error terms have zero mean and a covariance matrix, $\mathbf{\Sigma}$, i.e.:

$$\begin{bmatrix}\mathbf{e}_y \\ \mathbf{E}_X\end{bmatrix} = \begin{bmatrix}\mathbf{e}_y \\ \text{vec}\,\mathbf{E}_X\end{bmatrix} \sim \begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \boldsymbol{\sigma}^2\mathbf{Q}. \tag{B.2}$$

In Equation (A2.2) vec ($\mathbf{X}$) is the operation that vectorizes the matrix $\mathbf{X}$ [I×J] into a column vector [IJ×1], $\boldsymbol{\sigma}^2$ is the variance of unit weight and $\mathbf{Q}$ is the cofactor matrix.

In the above example matrix $\mathbf{Q}$ can be written as:

$$\mathbf{Q} = \begin{bmatrix}\mathbf{Q}_y & \mathbf{Q}_{Xy} \\ \mathbf{Q}_{yX} & \mathbf{Q}_X\end{bmatrix}. \tag{B.3}$$

The variance of unit weight and the cofactor matrix together form the variance/covariance matrix of the model, i.e.:

$$\mathbf{\Sigma} = \boldsymbol{\sigma}^2\mathbf{Q}. \tag{B.4}$$

At this point a model for the closure of this problem is needed. The model that has been considered is the Weighted Total Leas Square model. The weighted total least square problem can be stated as (van Huffel and Lemmerling, 2002):

$$\mathbf{P} = \mathbf{Q}^{-1} = \begin{bmatrix}\mathbf{P}_{11} & \mathbf{P}_{21} \\ \mathbf{P}_{12} & \mathbf{P}_{22}\end{bmatrix}, \tag{B.5}$$

and

$$\begin{cases} \min[\mathbf{e}_y^{\mathbf{T}}\ \mathrm{vec}(\mathbf{E}_X)^{\mathbf{T}}] \begin{bmatrix} \mathbf{P_{11}} & \mathbf{P_{21}} \\ \mathbf{P_{12}} & \mathbf{P_{22}} \end{bmatrix} \begin{bmatrix} \mathbf{e}_y \\ \mathrm{vec}(\mathbf{E}_X) \end{bmatrix}, \\ \mathbf{y} = (\mathbf{X} - \mathbf{E_X})\mathbf{B} + \mathbf{e_y} \end{cases} \tag{B.6}$$

Where the minimization of the first term is to be carried out in respect to the parameters **B**. This minimization is usually carried out by Lagrange multipliers and extensive details on its mathematical derivation can be found in van Huffel and Lemmerling, (2002) or Snow, (2012). This minimization yields:

$$\begin{bmatrix} \mathbf{Q_1} & (\mathbf{X} - \tilde{\mathbf{E}}_X) \\ (\mathbf{X} - \tilde{\mathbf{E}}_X) & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} \mathbf{y} - \tilde{\mathbf{E}}_X\mathbf{B} \\ 0 \end{bmatrix}, \tag{B.7}$$

$$\mathbf{Q_1} = \mathbf{Q}_y + (\mathbf{B}\otimes\mathbf{I}_I)^{\mathbf{T}}\mathbf{Q}_X(\mathbf{B}\otimes\mathbf{I}_I), \tag{B.8}$$

$$\tilde{\mathbf{E}}_X = \mathrm{invec}\ ((\mathbf{Q}_X(\mathbf{B}\otimes\mathbf{I}_I))\boldsymbol{\lambda}, \tag{B.9}$$

Where invec is the inverse of the vec operation of Equation (B.2) and forms a matrix with the original [I×J] dimensions, while ⊗ refers to the Kronecker product (Zehfuss, 1858).
This formulation is true only when the cofactor matrices $\mathbf{Q}_{yX}$ and $\mathbf{Q}_{Xy}$ are zero matrix.
Equations B.7, B.8 and B.9 must be solved iteratively because the parameters **B** are present in both terms of the equation. Several algorithm are available to solve these iteration and a slight modification of "Algorithm 2" from Snow, (2012) was used in the investigation.
The modified algorithm is reported in Table B.1

Table B.1: *the modified version of "Algorithm 2" from Snow, (2012) used in the incomplete methodology*

Step 1	Initialization: calculate $\mathbf{B}^0$ and set $\mathbf{E}_X{}^0 = \mathbf{0}$	
	$\mathbf{B^0} = \left(\mathbf{X^T Q_y^{-1} X} + \boldsymbol{\Sigma}_B^{-1}\right)^{-1}(\mathbf{X^T Q_y^{-1} y} + \boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B)$	(B.10)
Step 2	while $\|\|\mathbf{B}(t) - \mathbf{B}\ (t\text{-}1)\|\| < \delta$	
	$\mathbf{Q}_1^t = \mathbf{Q}_y + (\mathbf{B}^{t-1}\otimes\mathbf{I}_I)^{\mathrm{T}}\mathbf{Q}_X(\mathbf{B}^{t-1}\otimes\mathbf{I}_I)$	(B.11)
	$\boldsymbol{\Sigma}^t = ((\mathbf{X} - \mathbf{E}_X^{t-1})^{\mathrm{T}}(\mathbf{Q}_1^t)^{-1}(\mathbf{X} - \mathbf{E}_X^{t-1}) + \boldsymbol{\Sigma}_B^{-1})^{-1}$	(B.12)
	$\mathbf{B}^t = \boldsymbol{\Sigma}^t \times ((\mathbf{X} - \mathbf{E}_X^{t-1})^{\mathbf{T}}(\mathbf{Q}_1^t)^{-1}(\mathbf{y} - \mathbf{E}_X^{t-1}\mathbf{B}^{t-1}) + \boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B)$	(B.13)
	$\boldsymbol{\lambda}^t = (\mathbf{Q}_1^t)^{-1}[(\mathbf{y} - \mathbf{E}_X^{t-1}\mathbf{B}^{t-1}) - (\mathbf{X} - \mathbf{E}_X^{t-1})\mathbf{B}^t]$	(B.14)
	$\mathbf{e}_X^t = -(\mathbf{Q}_X(\mathbf{B}^t\otimes\mathbf{I}_I))\boldsymbol{\lambda}^t$	(B.15)
	$\mathbf{e}_y^t = \mathbf{Q}_y\boldsymbol{\lambda}^t$	(B.16)
	$\mathbf{E}_X^t = \mathrm{invec}\ (\mathbf{e}_X^t)$	(B.17)

The modifications from the original algorithm are mainly in Equations (B.12) and (B.13) to include prior knowledge in the calculations of the parameters.
In this methodology, still under investigation, the calculation of a prior for the $\mathbf{E}_m$ as of step 5 of the one proposed in chapter §4.3 is no longer necessary; furthermore, if data with different level of confidence are present their influence on the value of the parameters can be adjusted by using weighting within matrix **Q**. In the present investigation a variance of unit value of 1 was used, such as $\mathbf{Q} = \mathbf{\Sigma}$. The methodology discussed in this appendix proposed to change the generation step for the parameters **B** to use the algorithm proposed in Table A2.1 to simultaneously describe the measurement error on **X** and **y**, while generating the parameters.
The parameters are then assessed as per original methodology with a metropolis random walk algorithm and accepted with the criterion in the original methodology of Bano *et al*, (2018).
The problem with this methodology and the area of current study is the acceptance of the error term $\mathbf{E}_X$; whether to use it to simulate a distribution for the error as part of the data pre-processing and subtract it from the original data or using it to modify the **X** data matrix during the MCMC convergence process. The second approach has been tested and the results were found lacking. Having the error modelled explicitly by the methodology is surely an added benefit of the discussed methodology and as such this approach could be subject of further study along with the refinement of the methodology proposed in chapter §4.3

A2.1 Code

Following some of the code for the implementation of this method is reported; it is worth noting that this method as much of the error literature consider the output to be just a vector. A solution applied in this code is to iterate on each variable of the output matrix considering it a single vector. Starting from the hypothesis of independent response this can hold without problems, but other cases have to be investigated.

```
function [full_cond_theta,Vxx] =
full_conditional_theta_EIV(X,Y,theta,theta_mean,sigma,sigma_X)
%-------------------------------------------------------------------------marcoc
% Full conditional distribution of the model parameters
%
%                            Marco Cattaldo
%
% CAPE-Lab - University of Padova (Italy)
%--------------------------------------------------------------------------

% Dimensions
```

```
[~,k] = size(X);     % n = number of calibration samples; k = number of
input variables
[~,m] = size(Y);     % n = number of calibration samples; m = number of
response variables

% Mean and covariance (to be tuned)

u0 = reshape (theta_mean',k*m,1);
c = 1e3;   %the higher the coefficient the lesser the impact of the priors
V0 = kron(eye(m,m),c*eye(k,k));
[un,Vn,Vxx]=EIVMAPSnow2012V3(X,Y,u0,V0,sigma_X);
Vn = Spd_Mat(Vn.*eye(k*m));
full_cond_theta = mvnpdf (theta,un,Vn);          %full conditional of theta

end


function [theta,stdth,Vxx]=EIVMAPSnow2012V3(X,Y,u0,V0,~)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Simple algorithm without singular matrices, adapted from Snow, (2012)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[~,u]=size(Y);
[~,m]=size(X);

n=size(Y,1);
PYY=kron(eye(u),eye(n));
PXX=kron((eye(m)),eye(n));
X=[ones(n,1),X];

[~,m]=size(X);
PXX=[zeros(n,n*m-n);PXX];
PXX=[zeros(n*m,n),PXX];
QXX=pinv(PXX);
QYY=pinv(PYY);
sigma1=pinv(V0);
sigma_t=zeros([m,m,u]);
for i=1:u
    h=i-1;
```

```
    sigma_t(2:m,2:m,i)=sigma1(h*(m-1)+1:i*(m-1),h*(m-1)+1:i*(m-1));
end
mean1=reshape(u0,m-1,u);
mean_t=[ones(1,u);mean1];
theta=[];
for i=1:u
    h=i-1;

theta(:,i)=pinv(X'*PYY(h*n+1:i*n,h*n+1:i*n)*X)*(X'*PYY(h*n+1:i*n,h*n+1:i*n)
*Y(:,i));
end
% % % % INITIALIZATION
controlTH=1;
thit=0;
Vxx=zeros(n,m);
stdth1=zeros([m,m,u*u]);
% tic

while controlTH>=5e-7
    prevtheta=theta;
    for i=1:u
        h=i-1;

        %save iteration i-1
        thit=thit+1;
        Xi=X-Vxx;

Q1=QYY(h*n+1:i*n,h*n+1:i*n)+kron(prevtheta(:,i),eye(n))'*QXX*kron(prevtheta
(:,i),eye(n));
        Q1=pinv(Q1);
        stdth1(:,:,i*i)=pinv((Xi)'*(Q1)*(Xi)+(sigma_t(:,:,i)));
        stdth1(:,:,i*i)=(stdth1(:,:,i*i)+stdth1(:,:,i*i)')./2;
        theta(:,i)=(stdth1(:,:,i*i)*((Xi)'*(Q1)*((Y(:,i)-
Vxx*prevtheta(:,i)))+ (sigma_t(:,:,i))*mean_t(:,i)));
        lambda(:,i)=(Q1)*((Y(:,i)-Vxx*prevtheta(:,i))-(Xi*theta(:,i)));
        V(:,i)=(-QXX*kron(theta(:,i),eye(n)))*lambda(:,i);

    end
```

```
    Vx=sum(V,2);
    Vxx=reshape(Vx,n,m);
    controlTH=norm(prevtheta-theta,2) ;%%break condition
    if thit>=30
        controlTH=0;
    end

end
theta=reshape(theta(2:end,:),u*m-u,1);
Vxx=Vxx(:,2:end);

for H=1:u
    VAR{H}=matlab.lang.makeValidName(num2str(H))    ;
    stdth2.(VAR{H})=[];
    h=H-1;
    for K=1:u
        k=K-1;
        stdth2.(VAR{H})=[stdth2.(VAR{H}) ; stdth1(2:end,2:end,h*K+1+k)];
    end
end
stdth=[];
for i=1:u
    stdth=[stdth,stdth2.(VAR{i})];
end
```

References

Abboud, L., & Hensley, S. (2003, september 12). New prescription for drug makers: update the plants. *Wall Street Journal*.

Abramowitz, G., Gupta, H., Pitman, A., Wang, Y., Leuning, R., Cleugh, H., & Hsu, K.-L. (2005). Neural Error Regression Diagnosis (NERD): A Tool for Model Bias Identification and Prognostic Data Assimilation. *J. Hydrometeorol.,* **7**, 160-177.

Allegrini, F., Braga, J. W., Moreira, A. C., & Olivieri, A. C. (2018). Error Covariance Penalized Regression: A novel multivariate modelcombining penalized regression with multivariate error structure. *Anal. Chim. Acta*, 20-27. doi:10.1016/j.aca.2018.02.002

Anderson, M. J., & Whitcomb, P. J. (1993). *Optimizing Formulation Performance with Desirability Functions.* Minneapolis: Stat-Ease, inc. Retrieved from www.statease.com.

Aoki, R., Bolfarine, H., & Singer, J. M. (2001). Null intercept measurement error regression models. *Test,* **10**(*2*), 441-452.

Bano, G., Facco, P., Bezzo, F., & Barolo, M. (2018). Probabilistic design space determination in pharmaceutical product development: a Bayesian/latent variable approach. *AIChE J.* doi:10.1002/aic.16133

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *The Philosophical Transactions,* **53**, 370-418.

Bromley, P. A. (2014). Products and Convolutions of Gaussian Probability Density. Manchester: Imaging Sciences Research Group, University of Manchester.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* Boca Raton, FL, U.S.A.: Chapman & Hall/CRC, Taylor & Francis Group.

Cotter, S. L., Roberts, G. O., Stuart, A. M., & White, D. (2013). MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Stat. Sci.,* **28**(*3*), 424-446.

Debrus, B., Lebrun, P., Kindenge, J., Lecomte, F., Ceccato, A., Caliaro, G., . . . Hubert, P. (2011). Innovative high-performance liquid chromatography method development for the screening of 19 antimalarial drugs based on a generic approach, using design of experiments, independent component analysis and design space. *J Chromatogr A,* **1218**(*31*), 5205-5215.

Derringer, G., & Suich, R. (1980). Simultaneous Optimization of Several Response Variables. *J QUAL TECHNOL,* **12**(*4*), 214-219.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). *Multi- and Megavariate Data Analysis: Principles and Applications.* Umeå: UMETRICS AB.

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics* (4th ed.). Cambridge: Cambridge University Press.

Facco, P., Del Pastro, F., Meneghetti, N., Bezzo, F., & Barolo, M. (2015). Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development. *Ind. Eng. Chem. Res,* **54**(*18*), 5128–5138.

FDA. (2004). *Guidance for industry. PAT – A framework for innovative pharmaceutical development, manufacturing and quality assurance.* Rockville (MD), USA: Center for Drug Evaluation and Research, U.S. Food and Drug Administration.

Garcia, T., Cook, G., & Nosal, R. (2008). PQLI Key Topics - Criticality, Design Space,. *J Pharm Innov,* **3**, 60-68.

García-Muñoz, S., & Oksanen, C. A. (2010). Process modeling and control in drug development and manufacturing. *Comput. Chem. Eng.,* **34**, 1007-1008.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton: CRC Press.

Gilks, W., & Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *J. Royal Stat. Soc. Series C (Applied Statistics),* **41**(*2*), 337-348.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* **82**(*4*), 711-732.

Hapgood, K. P. (2007). Chapter 20: Granulation Rate Processes. In *Granulation.* Elsevier Science.

Hapgood, K. P., Lister, J. D., & Smith, R. (2003). Nucleation regime map for liquid bound granules. *AIChE J.,***49**(*2*), 350-361.

Hoskuldsson, A. (1988). PLS REGRESSION METHODS. *J. Chemom,* **2**, 211-228.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika,* **28**(*3-4*), 321-377.

Hughes, B. (2009). 2008 FDA drug approvals. *Nat Rev Drug Discov,* **8**, 93-96.

IBM Business Consulting Services. (2005). *Transforming industrialization: A new paradigm for pharmaceutical development.* Retrieved 05 18, 2018, from Transforming Industrialization: http://www-935.ibm.com/services/us/imc/pdf/ge510-3997-transforming-industrialization.pdf

ICH. (1999, October 6). *ICH harmonised tripartite guideline, specifications: test procedures and acceptance criteria for new drug substances and new drug producproducts: chemical substances Q6A, Step 4, 6-October-1999.*

ICH. (2008, june 4). *PHARMACEUTICAL QUALITY SYSTEM Q10.*

ICH. (2009, November). *The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, Quality Guideline Q8 (R2) Pharmaceutical Development.*

ICH. (2009, November). *The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, Quality Guideline Q8 (R2) Pharmaceutical Development.*

ICH. (2011). *ICH quality implementation working group. Points to consider (R2). ICH-endorsed guide for ICH Q8/Q9/Q10 implementation.* Geneva: ICH secretariat.

Jaeckle, C. M., & Macgregor, J. F. (1998). Product design through multivariate statistical analysis of process data. *AIChE J.,* **44**(*5*), 1105-1118.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge: Cambridge University press.

JR, J. (1965). A rolling theory for granular solids. *J Appl Mech, 32*(4), 842-848.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ Psychol Meas,* **20**(*1*), 141–151.

Kumar, A., Vercruysse, J., Belllandi, G., Gernaey, K. V., Vervaet, C., Remon, J., . . . Nopens, I. (2014). Experimental investigation of granule size and shape dynamics in twin-screw granulation. *Int.J.Pharm,* **1-2**, 485-495.

Lee, K. (2012). Continuous granulation of pharmaceutical powder using a twin screw granulator. University of Birmingham Research Archive . Retrieved from http://etheses.bham.ac.uk/4002/1/Lee13PhD.pdf

Lenk, P. (2001). Bayesian Inference and Markov Chain Monte Carlo. Ann Arbor: The University of Michigan Business School.

Lira, I., & Grientschnig, D. (2017). Error-in-variables models in calibration. *Metrologia,* **54**(*6*). doi:10.1088/1681-7575/aa8f02

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* New York: Springer-Verlag.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. *Control Eng. Pract,* **3**(*3*), 403-414.

Maguire, J., & Peng, D. (2015, October 6). How to Identify Critical Quality Attributes and Critical Process Parameters. *FDA/PQRI 2nd Conference*. North Bethesda, Maryland.

Mallick, R., Fung, K., & Krewski, D. (2002). Adjusting for measurement error in the Cox proportional hazards regression model. *J Cancer Epidemiol Prev*, **7**(*4*) 155-164.

Morris, A. S., & Langari, R. (2016). *Measurement and Instrumentation: Theory and Application* (II ed.). San Diego, CA: Academic Press.

Oka, S., Kašpar, O., Tokárová, V., Sowrirajan, K., Wu, H., Khan, M., . . . Ramachandran, R. (2015). A quantitative study of the effect of process parameters on key granule characteristics in a high shear wet granulation process involving a two component pharmaceutical blend. *Adv. Powder Technol,* **26**, 315-322.

Peterson, J. J. (2008). A Bayesian Approach to the ICH Q8 Definition of Design Space. *Journal of Biopharmaceutical Statistics,* **18**(*5*), 959-975.

Peterson, J. J., & Lief, K. (2010). The ICH Q8 definition of design space: A comparison of the overlapping means and the Bayesian predictive approaches. *Stat Biopharm Res.*,**2**, 249–259.

Peterson, J. J., Yahyah, M., Lief, K., & Hodnett, N. (2017). Predictive Distributions for Constructing the ICH Q8 Design Space. In *Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture* (pp. 55-70). Hoboken, New Jersey: John Wiley & Sons, inc.

Pfender, F., & Ziegler, G. M. (2004). Kissing numbers, sphere packings, and some unexpected proof. *Notices Am. Math. Soc.,* **51**, 873-883.

Rao, R. C. (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *J. Multivar. Anal,* **9**(*3*), 362-377.

Rosipal, R., & Krämer, N. (2006). Overview and Recent Advances in Partial Least squares. In S. C., G. M., G. S., & S.-T. J. (Eds.), *Subspace, Latent Structure and Feature Selection. Lecture Notes in Computer Science* (**3940**). Berlin, Heidelberg: Springer.

Snow, K. (2012). *Topics in Total Least-Squares Adjustment within the Errors-In-Variables Model: Singular Cofactor Matrices and Prior Information.* Retrieved from Ohio State University knowledge bank: https://kb.osu.edu/dspace/bitstream/handle/1811/78619/1/SES_GeodeticScience_Report_502.pdf

Souihi, N. (2014). *Multivariate Synergies in Pharmaceutical Roll Compaction: the quality influence of raw materials and process parameters by design of experiments.* Umeå: VMC-KBC Umeå.

Stockdale, G. W., & Cheng, A. (2009). Finding Design Space and a Reliable Operating Region Using a Multivariate Bayesian Approach with Experimental Design. *QUAL TECHNOL QUANT M,* **6**(*4*), 391-408.

Sun, Z. (2010). QbD for Generic Drugs: A Case Study for Immediate-Release Products. *2010 AIChE Annual Meeting.* Salt Lake City.

Tardos, G. I., Khan, I. M., & Mort, P. R. (1997). Critical parameters and limiting conditions in binder granulation of fine powders. *Powder Technol.,* **94**(*3*), 245-258.

Tomba , E. (2013, 10 14). *Latent variable modeling approaches to assist the implementation of quality-by-design paradigms in pharmaceutical development and manufacturing.* Retrieved from Padua Digital University Archive: http://paduaresearch.cab.unipd.it/5847/1/tomba_emanuele_tesi.pdf

Trefethen, L. N., & Bau, D. (1997). *Numerical Linear Algebra* (1st ed.). Philadelphia: Siam.

van Huffel, S., & Lemmerling, P. (2002). *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics* (1st ed.). Boca Raton: CRC Press.

Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Nordahl, Å., . . . Wold, S. (1998). Multivariate process and quality monitoring applied to an electrolysis process: Part I. Process supervision with multivariate control charts. *Chemometrics and Intelligent Laboratory Systems,* **42**(*1-2*), 221-231.

Yu, L. X. (2008, Apr). Pharmaceutical quality by design: product and process development, understanding, and control. *Pharm Res,* **25**(*4*), 781-791.

Zehfuss, G. (1858). Über eine gewisse Determinante. *Zeitschrift für Mathematik und Physik, 3*, 298-301. Retrieved from https://gdz.sub.uni-goettingen.de/id/PPN599415665_0003