UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Causal latent space-based models for scientific learning in Industry 4.0

València, July 2023

Author: Joan Borràs Ferrís

Ph.D. supervisor: Alberto J. Ferrer Riquelme

Abstract

The present Ph.D. thesis is devoted to studying, developing, and applying data-driven methodologies, based on multivariate statistical models of latent variables, to address the scientific learning paradigm in the Industry 4.0 environment. Particular emphasis is placed on causal latent variable-based models using both data coming from a planned design of experiments and, mainly, data coming from the daily production process, namely happenstance data. The dissertation is structured in five parts.

The first part discusses the scientific learning paradigm in the Industry 4.0 environment. The objectives of the thesis are highlighted. In addition to that, a comprehensive description of latent variable-based models is presented, on which the novel methodologies proposed in this thesis are founded.

In the second part, the novel methodological contributions are presented. Firstly, the potential of PLS to analyze data from DOE, with or without missing runs is illustrated. Then, the potential of causal latent variable-based models is concentrated on defining the raw material design space providing assurance of quality with a certain confidence level for the critical to quality attributes, jointly with the development of a novel latent space-based multivariate capability index to rank and select suppliers for a particular raw material used in a manufacturing process.

The third part aims to address novel applications by means of causal latent variable-based models using happenstance data. First, it concerns a health application: the Pandemic COVID-19. In this context, the use of latent variable-

based models is applied to develop an alternative to placebo-controlled clinical trials. Then, latent variable-based models are used to optimize processes within the framework of industrial applications.

The fourth part introduces a graphical user interface developed in Python code that integrates the developed methods with the aim of being self-explanatory and user-friendly.

Finally, the last part discusses the relevance of this dissertation, including proposals that deserve further research.

Resumen

La presente tesis doctoral está dedicada a estudiar, desarrollar y aplicar metodologías basadas en datos, fundamentadas en modelos estadísticos multivariantes de variables latentes, para abordar el paradigma del aprendizaje científico en el entorno de la Industria 4.0. Se pone especial énfasis en los modelos causales basados en variables latentes que utilizan tanto datos provenientes de un diseño de experimentos como, principalmente, datos provenientes del proceso de producción diario, es decir, datos históricos. La tesis está estructurada en cinco partes.

La primera parte discute el paradigma del aprendizaje científico en el entorno de la Industria 4.0. Se destacan los objetivos de la tesis. Además, se presenta una descripción exhaustiva de los modelos basados en variables latentes, sobre los cuales se fundamentan las metodologías novedosas propuestas en esta tesis.

En la segunda parte, se presentan las novedosas aportaciones metodológicas. En primer lugar, se muestra el potencial de PLS para analizar datos del DOE, con o sin datos faltantes. Posteriormente, el potencial de los modelos causales basados en variables latentes se centra en definir el espacio de diseño de la materia prima que proporciona garantía de calidad con un cierto nivel de confianza para los atributos críticos de calidad, junto con el desarrollo de un nuevo índice de capacidad multivariante basado en el espacio latente para clasificar y seleccionar proveedores para una materia prima particular utilizada en un proceso de fabricación.

La tercera parte pretende abordar aplicaciones novedosas mediante modelos causales basados en variables latentes utilizando datos históricos. En primer lugar, se trata de su aplicación en el ámbito sanitario: la Pandemia COVID-19. En este contexto, se utiliza el uso de modelos basados en variables latentes para desarrollar una alternativa a los ensayos clínicos controlados con placebo. Luego, se utilizan modelos basados en variables latentes para optimizar procesos en el marco de aplicaciones industriales.

La cuarta parte presenta una interfaz gráfica de usuario desarrollada en código Python que integra los métodos desarrollados con el objetivo de ser autoexplicativa y fácil de usar.

Finalmente, la última parte discute la relevancia de esta disertación, incluyendo propuestas que merecen mayor investigación.

Resum

Aquesta tesi doctoral està dedicada a estudiar, desenvolupar i aplicar metodologies basades en dades, fonamentades en models estadístics multivariants de variables latents, per abordar el paradigma de l'aprenentatge científic a l'entorn de la Indústria 4.0. Es posa un èmfasi especial en els models causals basats en variables latents que utilitzen tant; dades provinents d'un disseny d'experiments com, principalment, dades provinents del procés de producció diari, és a dir, dades històriques. La tesi està estructurada en cinc parts.

A la primera part es discuteix el paradigma de l'aprenentatge científic a l'entorn de la Indústria 4.0. Es destaquen els objectius de la tesi. A més, es presenta una descripció exhaustiva dels models basats en variables latents, sobre els quals es fonamenten les noves metodologies proposades en aquesta tesi.

A la segona part, es presenten les noves aportacions metodològiques. En primer lloc, es mostra el potencial de PLS per analitzar dades del DOE, amb dades faltants o sense aquestes. Posteriorment, el potencial dels models causals basats en variables latents se centra a definir l'espai de disseny de la matèria prima que proporciona garantia de qualitat amb un cert nivell de confiança per als atributs crítics de qualitat, juntament amb el desenvolupament d'un nou índex de capacitat multivariant basat en l'espai latent per a classificar i seleccionar proveïdors per a una primera matèria particular utilitzada en un procés de fabricació.

La tercera part pretén abordar aplicacions noves mitjançant models causals basats en variables latents utilitzant dades històrques. En primer lloc, es

tracta de la seva aplicació a l'àmbit sanitari: la Pandèmia COVID-19. En aquest context, es fa servir l'ús de models basats en variables latents per desenvolupar una alternativa als assaigs clínics controlats amb placebo. Després s'utilitzen models basats en variables latents per optimitzar processos en el marc d'aplicacions industrials.

La quarta part presenta una interfície gràfica d'usuari desenvolupada en codi Python que integra els mètodes desenvolupats amb l'objectiu de ser autoexplicativa i fàcil d'usar.

Finalment, l'última part discuteix la rellevància d'aquesta dissertació, incloent-hi propostes que mereixen més investigació.

Agraïments

Life is a complex and unpredictable journey, and at the same time, it is genuinely fascinating. In this journey, education represents the best tool to undertake it, as it uncovers new ventures that had remained hidden until that moment. The Ph.D. is part of this, and at this point, it is time to express gratitude to all those who have helped me along the way.

Gràcies a l'educació pública per permetre que tothom tinga almenys l'oportunitat de poder recórrer aquest camí de la manera més aplanada encara que sempre cal lluitar per preservar-la. Als meus pares, Carme i Leo, per aportar des del principi les eines, és a dir, els valors necessaris per emprendre aquest viatge, no amb la seguretat, però sí amb la intenció de fer sempre el que és correcte. I als meus germans, Leo i Lluís, per ser els meus referents des de menut.

Gracias a Alberto por alumbrar y dirigir mis pasos mientras construyo mi propio recorrido, todavía con un objetivo por descubrir, pero guiado siempre por la firmeza de aquello cimentado previamente por él. También me gustaría agradecer al resto del grupo. José Manuel, has sido mi supervisor adoptivo echándome una mano siempre que lo he necesitado. Dani, gracias por no dejarme caer durante los primeros pasos del doctorado. Y en especial a Alba, por hacer que cualquier acontecimiento en este recorrido haya sido impredeciblemente maravilloso.

Moltes gràcies als amics del poble i de la universitat per ser el *comboi* d'aquest camí, en especial a Antón per donar l'últim *toc*.

I entre tots ells, a Noelia, per ser la meua companya en aquesta aventura. La persona capaç de detindre's amb mi a la vora del sender per disfrutar del paisatge que ens envolta, però al mateix temps encoratjar-me, fins i tot amb xicotets gestos quotidians, per continuar quan em flaquejen les forces. Gràcies per ensenyar-me que la fermesa d'aquesta aventura es troba en l'estima.

A tots vosaltres, vos estime.

Contents

Part I

Prologue

Chapter 1

Justification, objectives and contributions

1.1 Justification

For a long time, it has been acknowledged that the process of scientific learning is achieved by a motivated iteration between theory and practice [1]. By practice, it is meant reality in the form of facts and data, and evidence of progress can be achieved through the continuous evolution of a developing theory, in the form of models, as it is exposed to reality, ultimately reaching a currently satisfactory level of understanding. For that, it is critical to determine the cause-and-effect relationships between different phenomena, and hence, causality is a fundamental concept in the scientific learning paradigm. A causal model must explain how changes in input variables relate to changes in the outputs. For this purpose, deterministic (i.e. first principles) models are always desirable. However, the lack of knowledge and the generally ample need for resources required to properly develop such models make their use unfeasible in many cases. In such cases, the inversion of empirical (i.e. data-driven) models, fitted on data from the process, can be carried out instead.

The advent of Industry 4.0 and the growing popularity of the Big Data movement have caused a recent shift in the nature of data. Data is now more abundant than ever before and the rate at which it accumulates is accelerating. This is characterized by the four V's: volume, variety, velocity and veracity. In this new Industry 4.0 environment, a new discipline has emerged: Data Science [2]. In general, data scientists usually apply machine learning models focused on correlation and prediction (i.e., passive use), rather than causation (i.e., active use). The main goal of these models is to find patterns in data and use them to make accurate predictions about new data. Therefore, in many cases, machine learning models focused on passive use can provide a good description of the relationship between different variables and can accurately predict future outcomes. However, these models cannot always identify the underlying causal relationships that explain why certain phenomena occur, and hence, no really new scientific knowledge is acquired in these situations. Therefore, one of the big challenges in the scientific learning paradigm is the development of statistical models that are able to iterate with this new data with the purpose of reaching new scientific knowledge. Although useful, it is well known that all these models will be wrong being not possible to obtain the "true" one [1], even more so, if data come from daily production characterized by a low signal-to-noise ratio. Thus, these statistical models must be useful not only to acquire knowledge but also to identify and comprehend the uncertainty arising from the discrepancy between theory and practice.

The present Ph.D. thesis aims at providing better insight and novel data-driven methodologies, based on multivariate statistical models of latent variables, to address the scientific learning paradigm in the Industry 4.0 environment. As commented, causality is a fundamental concept in the scientific learning paradigm. When using conventional predictive methods that directly relate the registered input variables with the output variables, causality must be inferred from data obtained from a Design of Experiments (DOE). Although Multiple Linear Regression (MLR) is a well-established statistical technique to analyze data from DOE, the analysis of experiments with incomplete data may be difficult for practitioners without a solid training in experimental design. Missing runs in experimental designs lead to aliasing due to correlated regressors and, unlike MLR, this is the environment where Partial Least Squares (PLS) regression, a latent variable-based multivariate statistical technique, performs particularly well. Part of this Ph.D. thesis will be devoted to studying the properties of PLS regression to analyze data from DOE.

Nevertheless, the use of classical DOE techniques is usually not feasible in real processes due to the generally high number of variables involved, which would require an impractically large amount of experimentation. Besides, one must also consider the logistic problems caused by the execution of these experiments, since they would force to stop the production itself in most cases. This leads to a situation in which any potential process improvement would not be enough to justify the high economic costs involved. In addition to this, the complex correlation structure among variables imposes several restrictions that prevent manipulating some factors independently from one another, as is required in a DOE. On the other hand, with the emergence of Industry 4.0 and the Big Data movement [3], it is typical for most companies to have access to large amounts of historical (happenstance) data that usually present certain (unplanned) excitations due to small changes in the operating conditions of the processes during their daily operation. This results from variations of properties and impurities in different batches of raw materials, changes in environmental conditions, equipment wear, process control adjustments made by operators, and so on. However, these data are highly collinear and low rank because variations in the inputs are commonly not independent (i.e., data are not obtained from a DOE that guarantees this independent variation in the inputs). As commented, with observational data not coming from a DOE, classical predictive models (such as linear regression and machine learning models focused on passive use), proven to be very powerful for prediction, cannot be used for extracting interpretable or causal models from historical data for active use. In fact, with historical data, there are an infinite number of models that can arise from any of these linear regression or machine learning meth-

ods, all of which might provide good predictions of the outputs, but none of which is unique or causal [4]. This is the essence of the Box et al. [5] warning: input-output correlation does not mean necessarily causation. Hence, there is a growing body of literature that recognizes the critical role played by causal network structures in order to infer causal relationships between original variables from observational data. Most methods presented so far resort to the existing expert knowledge [6] or use network inference techniques [7, 8] to establish the causal map. However, the aforementioned methodologies suffer with the increase of the process dimensionality [8].

On the contrary, latent variable-based models, such as PLS regression, allow for the analysis of large datasets containing highly correlated data. Since they assume that the input space and the output space are not of full statistical rank, they do not only model the relationship between them (as classical linear regression and machine learning models do) but also provide models for both spaces. This fact gives them a very nice property: uniqueness and causality in the reduced latent space no matter if the data come either from a DOE or daily production process (historical/happenstance data) [9]. Moreover, contrary to causal network structures, this latent variable-based approach does not require relying on expert knowledge or network inference techniques based on the original variables. Nevertheless, it is crucial to highlight that this approach does not prove causation in the way causal inference methods in DOE studies (or even causal network structures) do. Indeed, causality is not inferred in the original space, as it may be hindered by the confounding structure where there is no guarantee that any active change in the original space would respect the correlation structure of the data used to build the model. By contrast, active changes in the original variables can be done along the directions of the latent variables, which is equivalent to implicitly "changing" the latent variables themselves. This, of course, implies that the causality interpretation in the latent space is highly restricted since it will only provide active changes that respect the correlation structure of the latent variable-based model and, consequently, any confounding structure is kept (i.e., it will only allow us to modify the process in specific ways, so that the original variables are not varied independently from each other, and any solution will abide by the correlation structure defined by the subspace of the latent variable-based model).

Therefore, there is tremendous potential in Industry 4.0 to develop causal latent variable-based models using happenstance data (i.e., data not coming from a planned DOE but from historical data). To accomplish this, the rest of the Ph.D. thesis will be devoted to: i) defining the raw material design space, in line with the goals of the Quality by Design (QbD) initiative, ii) developing a

Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$), iii) illustrating the use of PLS for process optimization in some novel applications, and iv) integrating the developed methods by means of a Graphical User Interface (GUI).

1.2 Objectives

This section provides a detailed description of the objectives of this Ph.D. thesis and the proposals to achieve them.

Objective I: To study the properties of Partial Least Squares (PLS) regression to analyze data from Design of Experiments (DOE).

Chapter 4 of this Ph.D. thesis aims to analyze data from DOE, with or without missing runs, with just one method: Partial Least Squares (PLS) regression. This property is very attractive since, to the best of our knowledge, no other statistical tool has comparable versatility. Thus, we challenge the widely held view that PLS is useful only when dealing with non-experimental design (i.e., correlated observational data), but also when dealing with data from experimental designs.

Objective II: To define the raw material design space via latent variable-based models.

Raw materials properties are usually considered as Critical Input Parameters (CIPs) because their variability has an impact on Critical Quality Attributes (CQAs) of the final product. Hence, the development of specification regions for raw materials is crucial to ensure the desired quality of the product. In this context, this thesis focuses on developing novel methodologies based on latent variable-based models using happenstance data for:

- defining multivariate raw material specifications providing assurance of quality with a certain confidence level for the critical to quality attributes (CQAs) of the manufactured product. This corresponds to the estimation of the so-called raw material design space, which is defined as the multidimensional combination and interaction of inputs variables (e.g., raw material properties) and process conditions that have been demonstrated to provide assurance of quality [10] (Chapter 5).
- implementing an effective process control system attenuating most raw material variations. This allows expanding the raw material design space

and, hence, one may potentially be able to accept lower cost raw materials that will yield products with perfectly satisfactory quality properties (Chapter 6).

This objective refers to the robust design advocated long ago by Genichi Taguchi and that, nowadays, it is found in the goals of the QbD initiative nowadays [11].

Objective III: To develop a latent space-based multivariate capability index.

To rank and select suppliers for a particular raw material used in a manufacturing process, Chapter 7 focuses on developing a novel Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$). The novelty of this new index is that, contrary to other multivariate capability indexes proposed in the literature that are defined in the multivariate raw material space, this new $LSb\text{-}MC_{pk}$ is defined in the latent space connecting the raw material properties of a batch with the CQAs of the product manufactured with this raw material batch. This is of great interest as it quantifies the capacity of each supplier of providing assurance of quality with a certain confidence level for the CQAs. All we need is a database with historical information of the several properties measured for a particular raw material along with the CQAs of the corresponding manufactured product, which is usually available in Industry 4.0. Besides, Chapter 7 also aims to carry out the diagnosing assignable causes when a supplier does not score a good capability index.

Objective IV: To illustrate the use of PLS for process optimization using happenstance data.

We cannot know if any statistical technique we develop is useful unless we use it. Major advances in science, and in statistical science in particular, usually occur as the result of the theory-practice iteration [1]. For that reason, one of the most important goals of this thesis is to illustrate the utility of the causal latent variable-based models for process optimization using happenstance data by applying them to different novel applications in Part III (Chapters 8 and 9).

At the same time, the theory-practice iteration also demands us to improve our methods from the discrepancy between theory and practice. Hence, the thesis focuses on developing a reformulation of the process optimization problem with the purpose of addressing these novel applications by means of a creative process converging to a solution [12].

Objective V: To integrate the developed methods by means of a Graphical User Interface (GUI).

In statistical science, the theory-practice iteration requires a closed loop, by contrast, when for any reason the loop is open, the progress stops. Therefore, making the models accessible to new applications favors progress. There are two ways to apply these models: programming the algorithms behind the models (expert users) and using a Graphical User Interface (GUI) (starting users). For that reason, Chapter 10 integrates the developed methods in a GUI with the aim of being self-explanatory and user-friendly.

1.3 Contributions

The following is a comprehensive list of contributions made by the candidate during the course of this Ph.D. thesis.

1.3.1 Papers in peer-reviewed journals

1. [13] D. Palací-López, J. Borràs-Ferrís, and L. Thaise da Silva de Oliveria, "Multivariate Six Sigma: A Case Study in Industry 4.0," *Processes*, vol. 8, pp. 1–20, 2020. DOI: doi:10.3390/pr8091119

2. [14] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Defining multivariate raw material specifications in industry 4.0," *Chemometrics and Intelligent Laboratory Systems*, vol. 225, 2022, ISSN: 18733239. DOI: 10.1016/j.chemolab.2022.104563

3. [15] A. González-Cebrián, J. Borràs-Ferrís, J. P. Ordovás-Baines, M. Hermenegildo-Caudevilla, M. Climente-Marti, S. Tarazona, R. Vitale, D. Palací-López, J. F. Sierra-Sánchez, J. S. de la Fuente, and A. Ferrer, "Machine-learning-derived predictive score for early estimation of COVID-19 mortality risk in hospitalized patients," *PLoS ONE*, vol. 17, no. 9 September, pp. 1–17, 2022. DOI: 10.1371/journal.pone.0274171

4. [16] A. González-Cebrián, J. Borràs-Ferrís, Y. Boada, A. Vignoni, A. Ferrer, and J. Picó, "PLATERO: A calibration protocol for plate reader green fluorescence measurements," *Frontiers in Bioengineering and Biotechnology*, vol. 11, pp. 1–19, 2023, ISSN: 22964185. DOI: 10.3389/fbioe.2023.1104445

5. [17] J. Borràs-Ferrís, C. Duchesne, and A. Ferrer, "Defining Multivariate Raw Material Specifications via SMB-PLS," *Chemometrics and Intelligent Laboratory Systems*, vol. 240, 2023. DOI: 10.1016/j.chemolab.2023.104912

6. [18] J. Borràs-Ferrís, A. Folch-Fortuny, and A. Ferrer, "On the properties of PLS for analyzing Design of Experiments," 2023, **SUBMITTED**

7. [19] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "A latent space-based Multivariate Capability Index: A new paradigm for raw material supplier selection in Industry 4.0," 2023, **SUBMITTED**

1.3.2 Poster conference contributions

1. D. Palací-López, P. Villalba, J. Borràs-Ferrís, and A. Ferrer, "On-line process optimization through latent variable regression model inversion in (big) data environments," in *Symposium on Digitalization and Big Data in Biotech and Pharma*, Zurich, Switzerland, 2019

1.3.3 Oral conference contributions

1. J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Defining multivariate raw materials specifications via PLS model inversion," in *Scandinavian Symposium on Chemometrics (SSC16)*, Oslo, Norway, 2019

2. J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Multivariate specifications in the "big" data era," in *Annual Conference of European Network for Business and Industrial Statistics (ENBIS 2019)*, Budapest, Hungary, 2019

3. J. Borràs-Ferrís, A. Folch-Fortuny, and A. Ferrer, "Analysis of Incomplete Designed Experiments by Partial Least Squares (PLS) Regression," in *Annual Conference of European Network for Business and Industrial Statistics (ENBIS 2019)*, Budapest, Hungary, 2019

4. A. Ferrer, J. Borràs-Ferrís, D. Palací-López, and C. Duchesne, "Multivariate specifications in industry 4.0," in *International Forum Process Analytical Technology (IFPAC 2020)*, North Bethesda, USA, 2020

5. J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Defining Multivariate Raw Materials Specifications in Industry 4.0," in *AIChE Vir-*

tual Spring Meeting and 17th Global Congress on Process Safety, Online, 2021

6. A. Ferrer, J. Borràs-Ferrís, and D. Palací-López, "Are all data analytics techniques equally useful for process optimization in Industry 4.0?" In *Spring Meeting of European Network for Business and Industrial Statistics (ENBIS Spring 2021)*, Online, 2021

7. J. Borràs-Ferrís, C. Duchesne, and A. Ferrer, "Defining multivariate raw material specifications via SMB-PLS," in *Congrès Annuel Chimiométrie*, Brest, France, 2022

8. J. Borràs-Ferrís, A. González-Cebrián, J. Martínez-Minaya, D. Palací-López, and A. Ferrer, "Statistical Machine Learning for defining the Design Space," in *Annual Conference of the European Network for Business and Industrial Statistics (ENBIS 2022)*, Trondheim, Norway, 2022

9. S. García-Carrión, J. Borràs-Ferrís, and A. Ferrer, "Process Optimization from Historical Data in Industry 4.0," in *Annual Conference of the European Network for Business and Industrial Statistics (ENBIS 2022)*, Trondheim, Norway, 2022

10. J. Borràs-Ferrís, C. Duchesne, and A. Ferrer, "Multivariate Statistical Process Control via SMB-PLS," in *Colloquium Chemiometricum Mediterraneum (CCM XI 2023)*, Padova, Italy, 2023

11. S. García-Carrión, J. Borràs-Ferrís, and A. Ferrer, "On the use of retrospective DOE for process optimization from historical data in Industry 4.0," in *Colloquium Chemiometricum Mediterraneum (CCM XI 2023)*, Padova, Italy, 2023

12. A. Ferrer and J. Borràs-Ferrís, "Latent Structures-Based Multivariate Statistical Process Control: An Inevitable Shift in Industry 4.0," in *International Symposium on Statistical Process Monitoring (ISSPM 2023)*, València, Spain, 2023

13. J. Borràs-Ferrís, C. Duchesne, and A. Ferrer, "SMB-PLS for expanding Multivariate Raw Material Specifications in industry 4.0," in *Annual Conference of the European Network for Business and Industrial Statistics (ENBIS 2023)*, **ACCEPTED**, València, Spain, 2023

14. D. Palací-López, J. Borràs-Ferrís, S. García-Carrión, and A. Ferrer, "Multivariate Six Sigma: A case study in a chemical industry," in *Annual*

Conference of the European Network for Business and Industrial Statistics (ENBIS 2023), **ACCEPTED**, València, Spain, 2023

15. L. Pozueta, S. García-Carrión, J. Borràs-Ferrís, and A. Ferrer, "Multivariate Six Sigma: A case study in the automotive sector," in *Annual Conference of the European Network for Business and Industrial Statistics (ENBIS 2023)*, **ACCEPTED**, València, Spain, 2023

16. S. García-Carrión, J. Borràs-Ferrís, P. Goos, and A. Ferrer, "Retrospective DoE Methodology for Guiding Process Optimization from Historical Data," in *Annual Conference of the European Network for Business and Industrial Statistics (ENBIS 2023)*, **ACCEPTED**, València, Spain, 2023

1.3.4 Chapters of books

1. [20] A. Ferrer, D. Palací-López, J. Borràs-Ferrís, M. Barolo, and P. Facco, "Inverse design via PLS model inversion," in *The Digital Transformation of Product Formulation Concepts, Challenges, and Applications for Accelerated Innovation*, Due to 2023, Taylor & Francis, ch. 10

2. [21] A. Ferrer, J. Borràs-Ferrís, D. Palací-López, and C. Duchesne, "Multivariate specifications," in *The Digital Transformation of Product Formulation Concepts, Challenges, and Applications for Accelerated Innovation*, Due to 2023, Taylor & Francis, ch. 12.8

1.3.5 Software

Dragonet. Graphical User Interface (GUI) built in Python.

1.3.6 Awards

The 2019 ENBIS knowledge fund: a competitive grant to take part in the ENBIS 2019 Conference in Budapest.

The 2019 ENBIS Best Student Presentation for the best student presentation at the ENBIS 2019 Conference in Budapest

The 2022 ENBIS knowledge fund: a competitive grant to take part in the ENBIS 2022 Conference in Trondheim.

Best oral communication (jury award) at the VIII Meeting of Doctoral Students of the UPV.

Chapter 2

On latent variable-based regression models

Part of the content of this chapter has been included in:

[13] D. Palací-López, J. Borràs-Ferrís, and L. Thaise da Silva de Oliveria, "Multivariate Six Sigma: A Case Study in Industry 4.0," *Processes*, vol. 8, pp. 1–20, 2020. DOI: doi:10.3390/pr8091119

[14] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Defining multivariate raw material specifications in industry 4.0," *Chemometrics and Intelligent Laboratory Systems*, vol. 225, 2022, ISSN: 18733239. DOI: 10.1016/j.chemolab.2022.104563

2.1 Introduction

Latent variable-based models (LVMs) are statistical models specifically designed to analyze massive amounts of correlated data. The basic idea behind LVMs is that the number of underlying factors acting on a process is much smaller than the number of measurements taken on the system. Indeed, the factors that drive the process leave a similar signature on different measurable variables, which therefore appear correlated. By combining the measured variables, LVMs find new variables (called latent variables (LVs)) that optimally describe the variability in the data and can be useful in the identification of the driving forces acting on the system and responsible for the data variability [22].

LVMs can be used to relate data from different datasets: an input data matrix $\mathbf{X}$, and an output data matrix $\mathbf{Y}$. This is done by means of latent variable-based regression models (LVRMs), such as partial least squares (PLS) regression. Thus, LVRMs find the main driving forces acting on the input space that are more related to the output space by projecting the input ($\mathbf{X}$) and the output variables ($\mathbf{Y}$) onto a common latent space[1]. The number of LVs corresponds to the dimension of the latent space and can be interpreted, from a physical point of view, as the number of driving forces acting on a system [23].

2.2 Partial Least Square (PLS) Regression

PLS regression [24, 25] is a LVRM used not only to model the inner relationships between the matrix of inputs $\mathbf{X}$[2] ($N \times M$) and the matrix of output variables $\mathbf{Y}$ ($N \times L$), but also to provide a model for both. This fact gives them a very nice property: uniqueness and causality in the reduced latent space no matter if the data come either from a DOE or daily production process (historical/happenstance data) typical in Industry 4.0 [26, 27]. The PLS regression model structure can be expressed as follows:

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} \tag{2.1}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2.2}$$

[1] Note that, there are other approximations based on artificial intelligent methodologies (e.g., autoencoders) that use the idea of latent space, but they are not considered in this manuscript.

[2] In the remainder of this Ph.D. thesis (except Chapter 8), the input data matrix can refer to raw material properties, $\mathbf{Z}$, process conditions, $\mathbf{X}$, and known disturbances, $\mathbf{D}$. In Chapter 8, this input data matrix can refer to patient features, $\mathbf{Z}$, and drug therapy, $\mathbf{X}$.

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{2.3}$$

where the columns of the matrix $\mathbf{T}$ $(N \times A)$ are the PLS score vectors $(\boldsymbol{t}_a, a = 1, 2, 3, ..., A)$, containing the first A latent variables (LVs) from PLS. These score vectors explain most of the covariance between $\mathbf{X}$ and $\mathbf{Y}$, and each one of them is estimated as a linear combination of the original variables with the corresponding "weight" vector $\boldsymbol{t}_a = \mathbf{X}\boldsymbol{w}_a^*$ $(a = 1, 2, 3, ..., A)$ (Equation 2.1). These weights vectors, $\boldsymbol{w}_a^*$, are the columns of the weighting matrix $\mathbf{W}^*$ $(M \times A)$.

The PLS scores vectors are also good "summaries" of $\mathbf{X}$ according to the $\mathbf{X}$-loadings, $\mathbf{P}$ $(M \times A)$ (Equation 2.2), and good predictors of $\mathbf{Y}$ according to $\mathbf{Y}$-loadings, $\mathbf{Q}$ $(L \times A)$ (Equation 2.3), where $\mathbf{E}$ $(N \times M)$ and $\mathbf{F}$ $(N \times L)$ are residual matrices of $\mathbf{X}$ space and $\mathbf{Y}$ space, respectively. The sum of squares of $\mathbf{F}$ is an indicator of how good the model is in predicting the $\mathbf{Y}$-space, and the sum of squares of $\mathbf{E}$ is an indicator of how well the model explains the $\mathbf{X}$-space.

In order to evaluate the model performance when projecting the n-th observation $\boldsymbol{x}_n$ onto it, the Hotelling T^2 in the latent space, T_n^2, and the Squared Prediction Error, $SPE_{\boldsymbol{x}_n}$, are calculated [28]:

$$\boldsymbol{\tau}_n = \mathbf{W}^{*\mathrm{T}}\boldsymbol{x}_n \tag{2.4}$$

$$T_n^2 = \boldsymbol{\tau}_n^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\tau}_n \tag{2.5}$$

$$SPE_{\boldsymbol{x}_n} = \left(\boldsymbol{x}_n - \mathbf{P}\boldsymbol{\tau}_n\right)^{\mathrm{T}}\left(\boldsymbol{x}_n - \mathbf{P}\boldsymbol{\tau}_n\right) = \boldsymbol{e}_n^{\mathrm{T}}\boldsymbol{e}_n \tag{2.6}$$

where $\boldsymbol{\tau}_n$ refers to the n-th row extracted from $\mathbf{X}$ being defined as a column vector, $\boldsymbol{e}_n$ is the residual vector associated to the n-th observation (n-th row of $\mathbf{E}$) defined as a column vector, $\boldsymbol{\Lambda}^{-1}$ is defined as the $(A \times A)$ diagonal matrix containing the inverse of the A variances of the scores associated with the LVs, and $\boldsymbol{\tau}_n$ is the column vector of scores corresponding to the projection of the n-th observation $\boldsymbol{x}_n$ onto the latent subspace of the PLS model.

The Hotelling T^2 statistic of an observation (T_n^2) is the estimated squared Mahalanobis distance from the center of the latent subspace to the projection of such observation onto this subspace. The SPE statistic gives a measure of how close (in an Euclidean way) the n-th observation $(\boldsymbol{x}_n)$ is from the A-dimensional latent space. Upper confidence limits (with a specified confidence level) for both statistics, SPE_{lim} and T_{lim}^2, can be calculated for Phase I (model building) and Phase II (model exploiting) based on theoretical distributions

[29, 30]. The normality assumption on which these calculations are based is usually quite reasonable in practice. Alternatively, these confidence limits can be obtained from distribution-free methods by repeated sampling [31]. The only requirement is to have a large reference dataset. Besides, if this large dataset is available (as with historical/happenstance data), confidence limits for Phase II can also be used in Phase I. In the present Ph.D. thesis, SPE and T^2 99% confidence limits are calculated from theoretical distributions.

Once the PLS regression model has been fitted, it can be used directly in order to obtain the prediction vector corresponding to a particular observation, $\boldsymbol{x}^{obs}$, fulfilling that $T^2_{obs} \leq T^2_{lim}$ and $SPE^2_{\boldsymbol{x}^{obs}} \leq SPE^2_{lim}$ for Phase II, as:

$$\hat{\boldsymbol{y}}^{obs} = \mathbf{Q}\boldsymbol{\tau}^{obs} = \mathbf{Q}\mathbf{W}^{*\mathrm{T}}\boldsymbol{x}^{obs} \tag{2.7}$$

2.2.1 Prediction uncertainty

However, predictions are not free from uncertainty, yielding prediction errors. Three different sources of uncertainties can affect the prediction error e_l^{obs} of the l-th CQA $\hat{y}_l^{obs}$ given a new observation $\boldsymbol{x}^{obs}$ [32]: (i) measurement uncertainty in both the regressor matrix ($\mathbf{X}$) and the response matrix ($\mathbf{Y}$) used to calibrate the PLS model, (ii) uncertainty in the estimated model regression parameters, (iii) and uncertainty due to the unmodeled part of the response variable (structural model uncertainty).

Estimation of prediction uncertainty is done by using Ordinary Least Squares (OLS) as Faber and Kowalski [33] suggested. Although this approach is an approximation, it was observed to yield good results in practice [34]. First, it is assumed that the prediction error e_l^{obs} follows a normal distribution with zero mean and variance $\sigma^2_{e_l^{obs}}$ (Equation 2.9).

$$e_l^{obs} = y_l^{obs} - \hat{y}^{obs} \sim N\left(0, \sigma^2_{e_l^{obs}}\right) \tag{2.8}$$

Therefore ${e_l^{obs}}/{s_{e_l^{obs}}}$ follows a t-statistic with $N - df$ degrees of freedom and, consequently, the $(1-\alpha)$ prediction interval ($PI_{y_l^{obs}}$) on y_l^{obs} is calculated as:

$$PI_{y_l^{obs}} = \hat{y}_l^{obs} \pm t_{N-df,\alpha/2} s_{e_l^{obs}} \tag{2.9}$$

where N is the number of the PLS model calibration samples, df the degrees of freedom consumed by the model (it is set equal to the number of LVs of the

model[3]), α the false alarm rate for the prediction interval (i.e., $(1-\alpha)\times 100$ confidence level) and $s_{e_l^{obs}}$ the estimated standard deviation of the prediction error. The latter is calculated using Equation 2.10 when taking into account the second and third sources of uncertainty mentioned above. Note that, to estimate the first source of uncertainty requires explicit knowledge about error variance in $\mathbf{Z}$ and $\boldsymbol{y}$, which is estimated from replications and thus this limits its use in practice. However, it seems to be more practical to assume that the second and third sources of uncertainties dominate and ignore the first one [34].

$$s_{e_l^{obs}} = SE_l\sqrt{1 + h^{obs} + {}^1\!/_N} \tag{2.10}$$

In the above expression, h^{obs} is the leverage of the observation (Equation 2.11) and SE_l the standard error of calibration (Equation 2.12).

$$h^{obs} = \boldsymbol{\tau}^{obs^{\mathrm{T}}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{obs} \tag{2.11}$$

$$SE_l = \sqrt{\frac{\sum_{n=1}^{N}\left(y_{l,n} - \hat{y}_l\right)^2}{N - df}} \tag{2.12}$$

where $y_{l,n}$ is the measured value of the n-th observation for l-th CQA in the calibration dataset, and $\hat{y}_l$ is the estimated value for the l-th CQA in the calibration dataset.

2.2.2 Model inversion

The objective of model inversion is to find (predict) a window of Critical Input Parameters (CIPs) for a desired product quality characterized by the Critical Quality Attributes (CQAs). Jaeckle and MacGregor [9] proposed a framework for the inversion of PLS models using historical data available on the process operating conditions and on the corresponding product quality. Using standard regression or machine learning models, the inversion is inadequate because those models do not contain any information about the covariance structure and, consequently, the inversion solution of the model almost certainly does

[3]Although the derivation of the degrees of freedom for PLS is not straightforward, they are expected to be low in comparison with the number of observations when dealing with historical data, $N - df$ tends to N, thus having a negligible effect on estimating the prediction uncertainty.

not respect previous structural relationships, leading to unfeasible solutions. By contrast, when inverting a PLS model the inversion solution belongs to the latent space (defined by the latent variables) and, therefore, such a solution is constrained to be physically feasible and consistent with the sets of process conditions and correlation structure from the past. In this respect, the PLS model inversion has been demonstrated to be a valid tool to support the development of new products and their manufacturing conditions using historical data in several case studies [13, 23, 35–39].

When considering the inversion of a PLS model (Equations 2.1 and 2.3), the set of CIPs (column vector $\boldsymbol{x}^{new}$) that will yield the desired set of CQAs (column vector $\boldsymbol{y}^{des}$) are obtained by solving the following system of linear equations:

$$\boldsymbol{y}^{des} = \mathbf{Q}\boldsymbol{\tau}^{new} \tag{2.13}$$

where $\boldsymbol{\tau}^{new}$ is the vector of scores corresponding to the projection of the observation $\boldsymbol{x}^{new}$, which is estimated by the inversion of the PLS model:

$$\boldsymbol{\tau}^{new} = f^{-1}\left(\boldsymbol{y}^{des}\right) \tag{2.14}$$

Then, $\boldsymbol{x}^{new}$ is estimated going back from the latent space to the CIPs space as follows:

$$\boldsymbol{x}^{new} = \mathbf{P}\boldsymbol{\tau}^{new} \tag{2.15}$$

Equation 2.15 clearly shows that the solution $\boldsymbol{x}^{new}$, obtained by the PLS model inversion, is a linear combination of the loading vectors $\boldsymbol{p}_a$ (columns of $\mathbf{P}$) and thus belongs to the latent space. Besides, notice that the PLS model inversion involves solving a system of linear equations represented in a matrix form (Equation 2.13), where there are as many linear independent equations as the rank of $\mathbf{Y}$ ($r_{\mathbf{Y}}$), and the number of unknown variables corresponds to the dimensionality of the latent space (A). Thus, three possible cases are considered based on dimensions $r_{\mathbf{Y}}$ and A:

- $r_{\mathbf{Y}} > A$: the most likely case is that no solution provides the desired set of CQAs, but the least squares solution can be obtained as follows:

$$\boldsymbol{\tau}^{new} = \left(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathrm{T}}\boldsymbol{y}^{des}$$

- $r_{\mathbf{Y}} = A$: a single solution exists that provides the desired set of CQAs.

$$\boldsymbol{\tau}^{new} = \mathbf{Q}^{-1}\boldsymbol{y}^{des}$$

- $r_{\mathbf{Y}} < A$: it corresponds to an underdetermined system of linear equations, and has multiple solutions forming a vector space whose dimension is the difference between A and $r_{\mathbf{Y}}$. Hence, multiple solutions $\boldsymbol{\tau}^{new}$ fall into a $(A - r_{\mathbf{Y}})$-dimensional subspace of the A-dimensional space, that theoretically yields the same desired set of CQAs. This subspace is so-called Null Space (NS) and, in such a case, the model inversion requires defining such a space.

The latter situation ($r_{\mathbf{Y}} < A$) corresponds to the most common case and, for that reason, it has been widely studied. Jaeckle and MacGregor [35] defined the hyper-plane related to the NS by both the solution given by the pseudo-inverse with minimal Euclidean norm as a point which belongs to the NS (Equation 2.16), and the orthogonal directions referring to null variations in CQAs ($A - 1$ linearly independent vectors parallel to the NS).

$$\boldsymbol{\tau}^{new} = \mathbf{Q}^{\mathrm{T}}\left(\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\right)^{-1}\boldsymbol{y}^{des} \tag{2.16}$$

García-Muñoz et al. [37] extended this approach proposing a linear equation system where each equation defines the NS for each CQA as proposed by Jaeckle and MacGregor [35] (i.e., by both a point and orthogonal directions of null variations). On the other hand, Palací-López et al. [40] defined the NS for each l-th CQA by the analytical equation of a $(A-1)$-dimensional hyper-plane, which spans the multiple inversion solutions for such l-th CQA. The general form of a hyperplane only requires a constant (v_{0_l}) and a single orthogonal vector to the NS ($\boldsymbol{v}_l$). This vector corresponds to the direction of maximum variation of the l-th CQA. The intersection of all these NS (if they exist) gives the same solution as the one proposed by Jaeckle and MacGregor [35].

In this work, it is assumed that all variables are centered and scaled to unit variance as a pre-treatment. Thus, the l-th NS is defined as follows:

$$\begin{aligned} v_{0_l} + \boldsymbol{v}_l^{\mathrm{T}}\boldsymbol{\tau}^{NS,l} &= 0 \\ v_{0_l} &= -y_l^{des} \\ \boldsymbol{v}_l &= \boldsymbol{q}_l \end{aligned} \tag{2.17}$$

where $\boldsymbol{q}_l$ is the l-th row of $\mathbf{Q}$. When applied to all L CQAs:

$$\begin{gathered} \boldsymbol{v}_0 + \mathbf{V} + \boldsymbol{\tau}^{NS} = \mathbf{0} \\ \boldsymbol{v}_0 = \begin{bmatrix} v_{0_1} \\ v_{0_2} \\ \vdots \\ v_{0_L} \end{bmatrix} = -\boldsymbol{y}^{des} \qquad \mathbf{V} = \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_L^{\mathrm{T}} \end{bmatrix} = \mathbf{Q} \end{gathered} \tag{2.18}$$

Indeed, Equation 2.18 is equivalent to Equation 2.7 but expressed as the intersection of the L NSs (if it exists). However, in most practical cases CQAs are correlated, and this may raise singularity issues upon algebraic model inversion. To overcome this problem, Jaeckle and MacGregor [35] suggested two alternative approaches. The first one is to first build a Principal Component Analysis (PCA) model on the entire set of CQAs, and then use a significant number of columns of the relevant score matrix to build the response matrix. Nevertheless, a drawback of this approach is that some people may feel uncomfortable when using latent variables instead of true variables to represent product quality. The second approach relies on removing a priori some of the CQAs from the model output matrix, and on building the latent variable-based model in such a way that the inputs be related to the remaining CQAs only. By contrast, Arnese-Feffin [39] proposed an algebraic formulation of the latent variable-based model inversion problem, named regularized direct inversion, which can cope with CQA correlation by design. This enables one to retain in the model output matrix all CQAs and addresses output correlation by removing a posteriori only the non-systematic information that would cause singularity issues.

Finally, to put it briefly, Figure 2.1 shows the PLS model inversion by means of a simple example. In this example, there are three CIPs ($M = 3$) and the focus is on the l-th CQA, and a PLS model has been previously fitted using two components ($A = 2$). Then, given a desired l-th CQA, multiple solutions are predicted, which will theoretically result in such l-th CQA. These solutions belong to the one-dimensional NS.

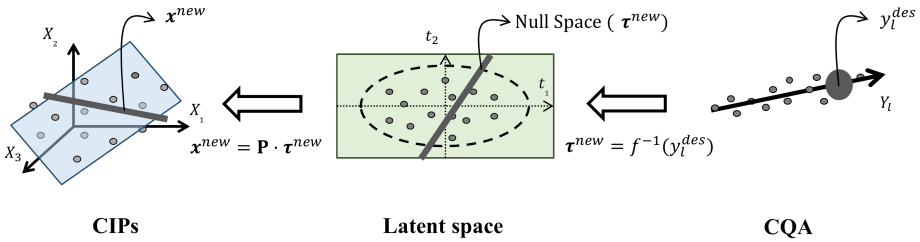


Figure 2.1: Simple example of the model inversion where there are three CIPs and the focus is on the l-th CQA, and a PLS model has been fitted by two components.

2.2.3 Optimization problem formulation

As mentioned in Section 2.2.2, if a null space exists, the solution of the model inversion could theoretically be moved along without affecting the product quality. Hence, the PLS model inversion can be formulated as an optimization problem in order to find the best feasible solution on the null space [23, 36, 41]. In any process, restrictions on the CIPs and CQAs may be imposed, for feasibility reasons. The optimization problem formulation can be formulated as follows based on Palací-López et al. [41]:

$$
\begin{aligned}
&\min_{\boldsymbol{\tau}} \left[g_0 \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau} \right)^{\mathrm{T}} \boldsymbol{\Gamma} \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau} \right) + g_1 \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} \right] \\
&s.t. \\
&\boldsymbol{v}_0 = -\boldsymbol{y}^{des} \\
&\mathbf{V} = \mathbf{Q} \\
&\hat{\boldsymbol{y}}^{new} = \mathbf{Q}\boldsymbol{\tau} \\
&\hat{\boldsymbol{x}}^{new} = \mathbf{P}\boldsymbol{\tau} \\
&T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{\tau} \leq T_{lim}^2 \\
&\mathbf{A}_{\boldsymbol{\tau}} \boldsymbol{\tau} \leq \boldsymbol{d}_{\boldsymbol{\tau}} \\
&\mathbf{F}_{\boldsymbol{\tau}} \boldsymbol{\tau} = \boldsymbol{f}_{\boldsymbol{\tau}}
\end{aligned}
\tag{2.19}
$$

where $\boldsymbol{\tau}$ is the score vector of the solution, composed by A elements, τ_a, $\hat{\boldsymbol{y}}^{new}$ and $\hat{\boldsymbol{x}}^{new}$ are the vectors of CQAs and CIPs, respectively, corresponding to the solution $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}^{-1}$ is the $(A \times A)$ diagonal matrix containing the inverse of the A variances of the scores, s_a^2, associated to the LVs. $\mathbf{A}_{\boldsymbol{\tau}}$ and $\boldsymbol{d}_{\boldsymbol{\tau}}$ ($\mathbf{F}_{\boldsymbol{\tau}}$ and $\boldsymbol{f}_{\boldsymbol{\tau}}$) are a matrix and a vector used to define inequality (and equality) hard constraints

on the LVs, respectively. These hard constraints expressed as restrictions on the LVs are the result of transferring the restrictions on the CPIs and CQAs to the latent space as Palací-López et al. [41] suggested. $\mathbf{\Gamma}$ is a $(L \times L)$ diagonal matrix where the l-th element in the diagonal represents the weight given to achieving the desired value for the l-th CQA. g_0 and g_1 are the weights given to each term in the objective function when solving the optimization problem.

Note that, from Equations 2.18 and 2.19, one can conclude that $\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau} = \hat{\boldsymbol{y}}^{new} - \boldsymbol{y}^{des}$. Therefore, the objective function in Equation 2.19 minimizes the sum of the weighted squared difference between the desired CQAs in $\boldsymbol{y}^{des}$ and those predicted by the model included in $\hat{\boldsymbol{y}}^{new}$ and of the Hotelling's T^2, represented by the second term of the objective function (soft constraint). The soft constraint on T^2 is included to find a solution lying as close as possible to the historical data when multiple solutions exist. In addition to that, a confidence limit, T^2_{lim}, can be also accounted as a hard constraint.

2.3 Sequential Multi-block (SMB) PLS regression

The SMB-PLS is a multi-block latent variable-based regression model [42] that combines the strengths of Multi-block PLS (MB-PLS) [43] and those of the Sequential Orthogonal PLS (SO-PLS) [44] methods as discussed Ref. [45]. Indeed, the SMB-PLS improves the interpretability of between block relationships over the traditional MB-PLS methods by imposing a sequential ordering of the blocks (pathway) and applying stepwise block orthogonalization. Besides, as opposed to the SO-PLS, it models both the orthogonal and correlated information between blocks.

The pseudocode of the SMB-PLS is presented in Appendix 2.A and the algorithm is also shown schematically in Figure 2.2, similarly as in Reference [45]. The algorithm in Figure 2.2 is presented for the two-blocks case ($\mathbf{Z}$ and $\mathbf{X}$) for the sake of simplicity explaining the algorithm, but it can be extended to any number of regressor blocks as it is shown in Appendix 2.A.

Figure 2.2 shows that the SMB-PLS uses a hierarchical structure where the input blocks are ordered according to the process flowsheet with the first block $\mathbf{Z}$, and the second block $\mathbf{X}$. The algorithm computes the block weights and scores from the first block $\mathbf{Z}$. The subsequent block $\mathbf{X}$ is then regressed onto the first block scores to extract the information that is correlated with $\mathbf{Z}$, and their block weights and scores are then calculated. All block scores are combined in the super level score matrix $\mathbf{T}$ and a PLS model is built between $\mathbf{Y}$ and $\mathbf{T}$ to obtain the super weights and super scores. Upon convergence, super-

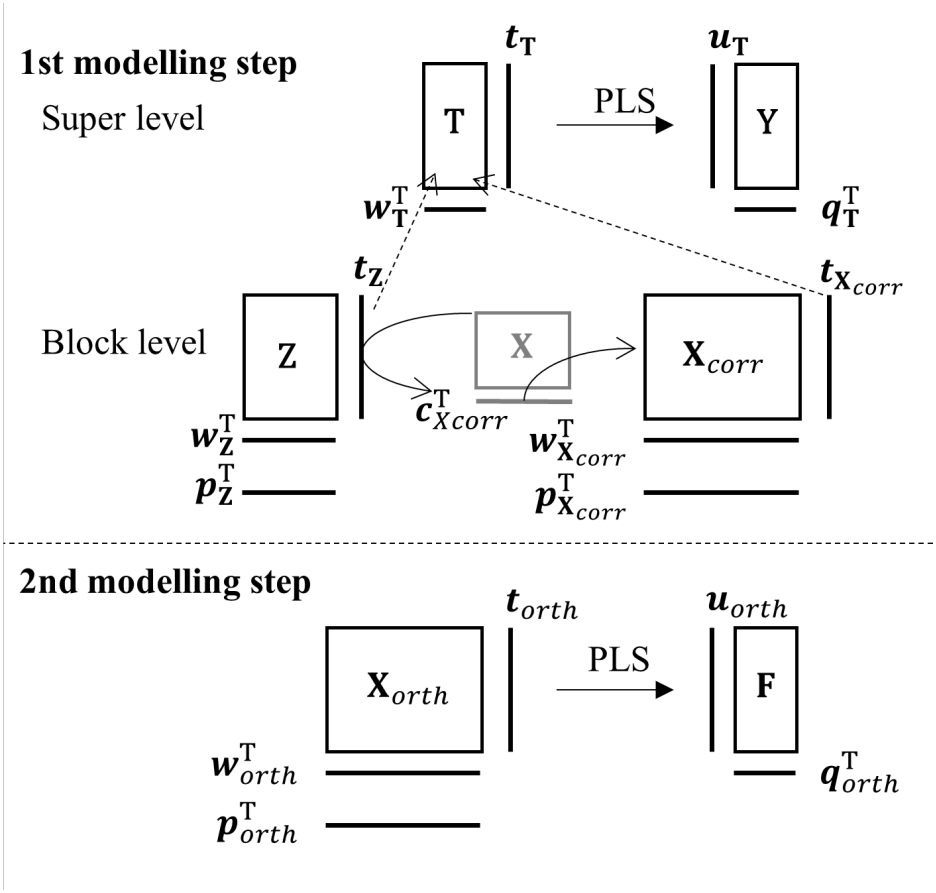


Figure 2.2: Scheme of SMB-PLS algorithm for two input blocks.

score deflation is applied to the input blocks, $\mathbf{Z}$ and $\mathbf{X}$, and the output block, $\mathbf{Y}$, ensuring that the next component will extract orthogonal information to the first one. The procedure is repeated for computing the next component using the residual of all data blocks. It continues to extract components from the first regressor block in the sequence until it has modelled all relevant information from $\mathbf{Y}$. When all relevant information from $\mathbf{Z}$ is extracted in the first modelling step, a regular PLS model is fitted to the $\mathbf{X}$ and $\mathbf{Y}$ residuals (i.e., $\mathbf{X}_{orth}$ and $\mathbf{F}$, respectively) in the second modelling step.

Note that, the sequential order of the blocks is critical. When a priori knowledge exists about the natural ordering of the blocks (e.g., data arising from

sequential operations in a production process), this specification is straightforward. However, in the absence of such knowledge, sequential methods (such as SO-PLS and SMB-PLS) face the problem of having to find the most adequate one. In this sense, Campos, Sousa and Reis [46] proposed a Stepwise SO-PLS as an efficient algorithm for selecting the block ordering when performing SO-PLS with capabilities of block exclusion. A priori, the same approach could be applied to the SMB-PLS. However, in the remainder of this manuscript, it is assumed that the natural order is known due to a priori knowledge.

Appendices

2.A SMB PLS regression

The pseudocode of the SMB-PLS assuming a process with B blocks is presented similarly to Ref. [45].

---

for $b = 1, 2, \ldots, B-1$ **do**
  set $\boldsymbol{u}_\mathbf{T}$ any column of $\mathbf{Y}$ ▷ Initialization
  while there is no convergence on $\boldsymbol{t}_\mathbf{T}$ or $\boldsymbol{u}_\mathbf{T}$ **do**
    $\boldsymbol{w}_b = \mathbf{X}_b^\mathrm{T}\boldsymbol{u}_\mathbf{T} / \left(\boldsymbol{u}_\mathbf{T}^\mathrm{T}\boldsymbol{u}_\mathbf{T}\right)$ ▷ Compute $\mathbf{X}_b$ block weights
    $\boldsymbol{w}_b = \boldsymbol{w}_b / \|\boldsymbol{w}_b\|$ ▷ Normalize weights vectors
    $\boldsymbol{t}_b = \mathbf{X}_b\boldsymbol{w}_b$ ▷ Compute $\mathbf{X}_b$ block scores
    for $k = 1, 2, \ldots, B-b$ **do**
      $\boldsymbol{c}_{(b+k)_{corr}} = \mathbf{X}_{b+k}^\mathrm{T}\boldsymbol{t}_b / \left(\boldsymbol{t}_b^\mathrm{T}\boldsymbol{t}_b\right)$ ▷ Compute correlation coefficients
      $\mathbf{X}_{(b+k)_{corr}} = \boldsymbol{t}_b\boldsymbol{c}_{(b+k)_{corr}}^\mathrm{T}$ ▷ Extract correlated information
      $\boldsymbol{w}_{(b+k)_{corr}} = \mathbf{X}_{(b+k)_{corr}}^\mathrm{T}\boldsymbol{u}_\mathbf{T} / \left(\boldsymbol{u}_\mathbf{T}^\mathrm{T}\boldsymbol{u}_\mathbf{T}\right)$ ▷ Compute weights
      $\boldsymbol{w}_{(b+k)_{corr}} = \boldsymbol{w}_{(b+k)_{corr}} / \|\boldsymbol{w}_{(b+k)_{corr}}\|$ ▷ Normalize weights
      $\boldsymbol{t}_{(b+k)_{corr}} = \mathbf{X}_{(b+k)_{corr}}\boldsymbol{w}_{(b+k)_{corr}}$ ▷ Compute scores
    end for
    $\mathbf{T} = \left[\boldsymbol{t}_b, \boldsymbol{t}_{(b+1)_{corr}}, \ldots, \boldsymbol{t}_B\right]$ ▷ Concatenate block scores in $\mathbf{T}$
    $\boldsymbol{w}_\mathbf{T} = \mathbf{T}^\mathrm{T}\boldsymbol{u}_\mathbf{T} / \left(\boldsymbol{u}_\mathbf{T}^\mathrm{T}\boldsymbol{u}_\mathbf{T}\right)$ ▷ Compute super weights
    $\boldsymbol{w}_\mathbf{T} = \boldsymbol{w}_\mathbf{T} / \|\boldsymbol{w}_\mathbf{T}\|$ ▷ Normalize super weights
    $\boldsymbol{t}_\mathbf{T} = \mathbf{T}\boldsymbol{w}_\mathbf{T}$
    $\boldsymbol{q}_\mathbf{T} = \mathbf{Y}^\mathrm{T}\boldsymbol{t}_\mathrm{T} \left(\boldsymbol{t}_\mathrm{T}^\mathrm{T}\boldsymbol{t}_\mathrm{T}\right)$
    $\boldsymbol{u}_\mathbf{T} = \mathbf{Y}\boldsymbol{q}_\mathrm{T} \left(\boldsymbol{q}_\mathrm{T}^\mathrm{T}\boldsymbol{q}_\mathrm{T}\right)$
  end while
  for $k = 1, 2, \ldots, B-b$ **do**
    $\boldsymbol{p}_k = \mathbf{X}_k^\mathrm{T}\boldsymbol{t}_\mathrm{T} \left(\boldsymbol{t}_\mathrm{T}^\mathrm{T}\boldsymbol{t}_\mathrm{T}\right)$ ▷ Compute $\mathbf{X}_k$ block loadings
    $\mathbf{E}_k = \mathbf{X}_k - \boldsymbol{t}_\mathrm{T}\boldsymbol{p}_k^\mathrm{T}$ ▷ Deflate $\mathbf{X}_k$ block
  end for
  $\mathbf{F} = \mathbf{Y} - \boldsymbol{t}_\mathrm{T}\boldsymbol{q}_\mathrm{T}^\mathrm{T}$ ▷ Deflate $\mathbf{Y}$ block
  Store all vectors at the block and super levels in matrices.
  To compute the next LV, replace $\mathbf{X}_k$ by $\mathbf{E}_k$ ($k \geq b$) and $\mathbf{Y}$ by $\mathbf{F}$, and go back to 2 until the relevant information in block $\mathbf{X}_b$ is depleted.
end for
For $b = B$, fit a regular PLS model to $\mathbf{E}_B$ and $\mathbf{F}$.

Chapter 3

Materials

3.1 Hardware

All the computations for the elaboration of the present Ph.D. were carried out on a DELL Inspiron 7386 equipped with Intel Core i7-8565U, CPU 1.80 GHz, and 16 GB of RAM.

3.2 Software

All functions, algorithms, and scripts used in the present Ph.D. (except Chapter 4) are original code implemented in Python 3.9.13 [47], leveraging the following packages:

- pandas 1.4.4 [48]
- numpy 1.21.5 [49]
- matplotlib 3.5.2 [50]
- scipy 1.9.1 [51]
- PyQt 5.9.2

The software MATLAB 2022a (The MathWorks, Inc.) is used for the development of Chapter 4.

3.3 Datasets

Different datasets are used in the present Ph.D. to evaluate the performance of the novel methods. For the convenience of the reader, the information regarding datasets can be found in the corresponding chapter where they are resorted or applied.

Part II

Novel methodological contributions

Chapter 4

On the properties of PLS for analyzing Design of Experiments

Part of the content of this chapter has been included in:

[18] J. Borràs-Ferrís, A. Folch-Fortuny, and A. Ferrer, "On the properties of PLS for analyzing Design of Experiments," 2023, **SUBMITTED**

4.1 Introduction

Two-level full factorial designs, 2^k, and fractional factorial designs, 2^{k-p}, are well known and widely used experimental designs that can be easily analyzed when there are no missing runs using Multiple Linear Regression (MLR) [5]. Several authors have studied the problem of analyzing factorial designs with missing runs and proposed different solutions. One of the proposals consists of fitting an appropriate model to the incomplete data by MLR. However, the analysis of experiments with incomplete data may be difficult for practitioners without a solid training in experimental design. For this reason, Yates [52] proposed to plug in a fitted value for the response values for the missing runs and analyze the experiment as if there were no missing runs. In this sense, the advice of Cochran and Cox [53] was to estimate the missing values by minimizing the sum of squares for the interactions that are used as error (i.e., sacrificed effects). The minimization of this sum of squares provides a system of equations with as many equations as missing runs as is explained in Section 4.4. However, unless a careful choice is made of effects to be sacrificed, the equations may be dependent, and no solution will be possible. An identical estimate would be obtained by equating to zero the sacrificed effects as Draper and Stoneman proposed [54]. In this case, sacrificed effects must also be chosen providing independent equations in order to estimate the missing runs.

In practice, missing runs occur due to different reasons, such as i) lack of resources to execute all the runs, or ii) problems in the execution of some of the runs. Regarding scenario i), a prior selection of the runs to be omitted can be done to obtain the maximum information in the model estimation. For instance, the D-optimal criterion is one of the most common approaches for generating optimal designs by finding a regression matrix ($\mathbf{X}$) that maximizes the determinant of the information matrix ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) [55] However, as Xampeny et al. [56] pointed out, the provided optimal designs (e.g., D-optimal) often change the factor levels, even if only slightly. Besides, the estimated effects may have different variances making the statistical analysis more complex, especially for practitioners with limited training in DOE. For that reason, they recently proposed a simple and easy-to-understand method being useful for practitioners and experimenters who lack a deep theoretical knowledge of optimal designs and linear models. This method is based on the Draper and Stoneman's [54] method of setting equal to zero the effects that can be considered negligible (i.e. sacrificed effects) a priori in order to estimate the response of the runs to be omitted. Unlike D-optimal designs, Xampeny et al.´s [56] method yields estimates of the effects that not only have similar and small variances, but are also as independent as possible. This has the additional advantage that,

once the missing values have been estimated, the analysis procedure is the same as if there were no missing runs. In the scenario ii), there is no prior selection of the runs to skip. Execution problems yield accidentally missing responses in the design, which do not necessarily follow any optimal criteria. Consequently, estimating the missing runs according to Cochran and Cox [53] or Draper and Stoneman [54] might be difficult, because the alias structure of the design might make prior choice of sacrificed interactions unfeasible in practice. Besides, being X the regressor or contrast matrix, the random missing runs may yield an ill-conditioned information matrix ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) that could create severe problems when using MLR directly. In such a case, a variable selection method would be required (e.g., stepwise regression). In addition, the potential ill-conditioned matrix may hinder causality interpretations in the original space with any predictive method used that directly connects X-space to Y-space, as independent variations in the regressors are not completely satisfied. Box et al. [5] already warned that, due to the confounding, regressor-response correlation does not imply necessarily direct causation.

The goal of this chapter is to study the properties of Latent Variable (LV) models, such as Partial Least Squares (PLS) regression, to analyze incomplete experimental data in both scenarios i) and ii). PLS is well known from its ability to analyze data with many, noisy, collinear, and even incomplete data in both regressors and response spaces [57], typical when dealing with non-experimental design (i.e., correlated observational data). We challenge the widely held view that PLS is useful only when dealing with this kind of data and, hence, we also highlight the potential of PLS to be used to analyze data from design of experiments, especially when some runs are missing (incomplete designs). The rationale for carrying out this proposal is that missing runs in experimental designs lead to aliasing due to correlated regressors and, unlike MLR, this is the environment where PLS performs particularly well.

The chapter is structured as follows. Section 4.2 shows the equivalence of one-PLS component model and MLR in the estimated effects and statistical significance analysis of complete two-level full and fractional designs, widely used in practice. In Section 4.3, we present the traditional approaches to the problem of analyzing experimental designs with missing runs and then, in Section 4.4 a novel methodology, based on PLS, is proposed to address this problem. Section 4.5 illustrates the methodology by two illustrative examples. Finally, Section 4.6 gives an easy-to-follow route map useful for practitioners without a solid training in experimental design to efficiently analyze DOE data with missing runs (either complete or incomplete) using PLS, and it sums up the discussion of the findings.

4.2 Two-level factorial designs

4.2.1 Full factorial designs: 2^k

A two-level full factorial design consists of all possible combinations of two levels for k factors yielding a total number of $N = 2^k$ different runs. The resulting design matrix contains in columns the values -1 or $+1$ corresponding to the two levels of the k factors under study.

4.2.1.1 Equivalence in the estimated effects by MLR and PLS

In matrix notation, the MLR model is commonly written as:

$$\boldsymbol{y} = \mathbf{X}\hat{\boldsymbol{b}}_{MLR} + \boldsymbol{f} \tag{4.1}$$

where $\boldsymbol{y}$ is the centered vector of observed values of the response variable, $\mathbf{X}$ is the regressor or contrast matrix[1] of dimensions $N \times M$, being $M = 2^k - 1$ the number of total estimable effects in the full factorial design, $\boldsymbol{f}$ represents the residual vector expressing the deviation between measured and predicted response values, and $\hat{\boldsymbol{b}}_{MLR}$ is the estimation of the vector of population regression coefficients $(\boldsymbol{b}_{MLR})$[2].

Since in these designs $\mathbf{X}$ is an orthogonal matrix (i.e., full rank), the effects can be estimated by the standard least squares solution as follows:

$$\hat{\boldsymbol{b}}_{MLR} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} \tag{4.2}$$

having the regression variables (columns of $\mathbf{X}$) the same variance $s^2_{\boldsymbol{x}} = {}^{N}/_{(N-1)}$. Thus, the information matrix can be expressed as:

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) = s^2_{\boldsymbol{x}}\,(N-1)\,\mathbf{I}^{M\times M} = N\mathbf{I}^{M\times M} \tag{4.3}$$

[1] $\mathbf{X}$ matrix is the 2^k design matrix augmented with new columns obtained by multiplying the original k columns. This way, the columns of the $\mathbf{X}$ matrix contain the regressors to estimate the different contrasts: main and interaction effects.

[2] The estimated regression coefficients are equal to one-half of the respective estimated effects of the two-level factorial designs obtained as the scalar product of the response variable vector, $\boldsymbol{y}$, and the k-th column of $\mathbf{X}$ matrix corresponding to the effect of interest, divided by half the number of runs $N/2$.

where $\mathbf{I}^{M\times M}$ is the identity matrix ($M \times M$). Substituting Equation 4.3 in Equation 4.2:

$$\hat{\boldsymbol{b}}_{MLR} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{N} \tag{4.4}$$

On the other hand, by substituting Equation 2.1 in Equation 2.3, PLS can be rewritten as an MLR-like model:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^{*}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \mathbf{X}\hat{\mathbf{B}}_{PLS} + \mathbf{F} \tag{4.5}$$

where $\hat{\mathbf{B}}_{PLS}$ is the matrix of estimated PLS regression coefficients ($\hat{\mathbf{B}}_{PLS} = \mathbf{W}^{*}\mathbf{Q}^{\mathrm{T}}$). This PLS regression, often termed PLS2, allows modelling a set of different responses jointly. A special version of it, the PLS1 algorithm, used to model a single variable $\boldsymbol{y}$, can be expressed from Equation 4.6 where matrices $\mathbf{Q}^{\mathrm{T}}$, $\hat{\mathbf{B}}_{PLS}$ and $\mathbf{F}$ degenerate to vectors $\boldsymbol{q}$, $\hat{\boldsymbol{b}}_{PLS}$ and $\boldsymbol{f}$, respectively.

$$\boldsymbol{y} = \mathbf{X}\mathbf{W}^{*}\boldsymbol{q} + \boldsymbol{f} = \mathbf{X}\hat{\boldsymbol{b}}_{PLS} + \boldsymbol{f} \tag{4.6}$$

It is important to remark the PLS space has up to C relevant components for prediction if and only if there are C different eigenvalues of the regressor covariance matrix. Besides, the PLS regression coefficients, $\hat{\boldsymbol{b}}_{PLS}$, will be equal to the MLR regression coefficients, $\hat{\boldsymbol{b}}_{MLR}$, if and only if the number of the latent variables of the PLS model, A, is equal to the C relevant components [58]. Therefore, MLR is a particular case of PLS when extracting the maximum number of C relevant components.

When PLS is applied to data from a full factorial design the regression matrix is orthogonal having only $C = 1$ relevant component (i.e., all eigenvalues are equal). Therefore, only the first PLS component has predictive ability, and the one-PLS component model matches the MLR solution [59]. An analytical demonstration is shown below.

When considering data from a two-level full factorial design, the one-PLS component model $\boldsymbol{y}$-loading q is obtained as follows:

$$q = \frac{\parallel \mathbf{X}^{\mathrm{T}}\boldsymbol{y} \parallel}{N} \tag{4.7}$$

where $\| \mathbf{X}^{\mathrm{T}}\boldsymbol{y} \|$ is the 2-norm of the vector $\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$, and the elements of the weight vector $\boldsymbol{w}^{*}$ are calculated according to Equation 4.8.

$$\boldsymbol{w}^{*} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\| \mathbf{X}^{\mathrm{T}}\boldsymbol{y} \|} \tag{4.8}$$

More details about Equations 4.7 and 4.8 can be found in Appendix 4.A.

The last equivalence can be also deduced when considering that the first PLS component maximizes the covariance between the first latent variable $\boldsymbol{t}$ and the response vector $\boldsymbol{y}$ [60] (see Appendix 4.B).

Then, when having a one-PLS component model, the regression vector of PLS coefficients is estimated from Equation 4.6 as:

$$\hat{\boldsymbol{b}}_{PLS} = \boldsymbol{w}^{*}q \tag{4.9}$$

Hence, by substituting Equations 4.7 and 4.8 in Equation 4.9:

$$\hat{\boldsymbol{b}}_{PLS} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\| \mathbf{X}^{\mathrm{T}}\boldsymbol{y} \|}\frac{\| \mathbf{X}^{\mathrm{T}}\boldsymbol{y} \|}{N} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{N} \tag{4.10}$$

Equation 4.10 is equivalent to Equation 4.4, demonstrating that the solution to the one-PLS component model corresponds to the MLR solution. The reason why only the first PLS component has predictive ability is that the $\boldsymbol{y}$ residual vector after extracting the first PLS component, $\boldsymbol{f}$, is orthogonal to $\mathbf{E}$ (see Appendix 4.C), so when analyzing orthogonal arrays there is no point in extracting more than one PLS component. Note that, this is consistent with the Helland criterion [58] previously commented in the same section, as in a two-level full factorial there is only one different eigenvalue of the contrast matrix and, consequently, there is only one relevant component for prediction.

Finally, in Appendix 4.D we present a simple case to illustrate the latent space in a full factorial design.

4.2.1.2 Equivalence in the statistical significance analysis by MLR and one-PLS component model

In MLR, it is assumed that the estimation of the m-th regression coefficient follows a normal distribution with mean $b_{MLR,m}$ and standard deviation $\sigma_{b,m}$. Since $\sigma_{b,m}$ is almost always unknown, an estimate of $\sigma_{b,m}$, $s_{b,m}$, is used instead. Thus, when the null hypothesis (the m-th effect is zero) is true ($b_{MLR,m} = 0$), the t-statistic follows a Student's distribution with $N - M^{*3} - 1$ degrees of freedom:

$$t = \frac{\hat{b}_{MLR,m}}{s_{b,m}} \sim t_{N-M^*-1} \tag{4.11}$$

where $s_{b,m}$ is calculated as:

$$s_{b,m} = \sqrt{MS_E \left[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right]_{m,m}} \tag{4.12}$$

where MS_E is the mean square error of the model and $\left[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right]_{m,m}$ is the m-th element of the diagonal of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. In a two-level full factorial design, the following equivalence is deduced from Equation 4.3:

$$\left[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right]_{m,m} = \frac{1}{N} \tag{4.13}$$

and Equation 4.11 is expressed as:

$$t = \frac{\hat{b}_{MLR,m}}{\sqrt{MS_E \left[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right]_{m,m}}} = \frac{\hat{b}_{MLR,m}}{\sqrt{MS_E/N}} \sim t_{N-M^*-1} \tag{4.14}$$

The uncertainty of the model parameters from two block regression modelling by PLS has been already discussed [62, 63] when using happenstance data (i.e., data not coming from an experimental design). In this case, the use of cross-validation/jack-knife (CV/JK) resampling approach is widely recommended

[3] Without replicates, it is necessary to include negligible effects on the residual in order to have degrees of freedom to estimate random noise. Negligible effects can be detected, for instance, by the Normal Probability Plot (NPP) or the Lenth's method [61], as illustrated later.

[62]. Cross-validation consists of re-estimating all the parameters of the model several times, each time keeping out one or more of the available samples (a cross-validation segment) during the estimation. In each cross-validation segment, a set of the model's parameters estimates is obtained. Then, statistical significance is evaluated by calculating the JK confidence intervals [63].

Nevertheless, when data come from a DOE without replicates, there is insufficient redundancy between observed samples and the resampling strategy is meaningless [64]. In such a case, we propose a statistical significance analysis in the PLS solution, which is equivalent to MLR's. This approach is based on the acceptance region for the estimated regression weights $\boldsymbol{w}_m$ proposed by Martens et al. [59], which contains those weights being consistent with the null hypothesis (i.e., $\boldsymbol{b}_{PLS,m} = 0$) given a particular false alarm rate. Thus, having data from a two-level full factorial design:

$$SS_m = {w_m^*}^2 q^2 N \tag{4.15}$$

$$SS_E = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - q^2 N \tag{4.16}$$

where SS_m is the sum of the squares of the m-th effect and SS_E the residual sum of the squares (see Appendix 4.E for more details). Then, by assuming that the null hypothesis is true ($\boldsymbol{b}_{PLS,m} = 0$), the ratio between the mean square of the m-th effect (MS_m) and the residual mean square (MS_E) follows a F-distribution with 1 and $N - (A^\star + 1)$ degrees of freedom:

$$\frac{MS_m}{MS_E} = \frac{{w_m^*}^2 q^2 N}{\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - q^2 N}{N-(A^\star+1)}} \sim F_{1,N-(A^\star+1)} \tag{4.17}$$

where $(A^* + 1)$ denotes the degrees of freedom used for parameter estimation. For PLS, the most common value used for A^* corresponds to the number of the latent variables of the PLS model, A (i.e., PLS model dimensionality). However, when having a full factorial design only the first PLS component is used to estimate all the effects included in the model (M^*), hence, A^* is equal to M^* instead of 1.

Combining Equation 4.9 for the m-th effect with Equation 4.17, and reorganizing terms, it is deduced that:

$$\frac{\hat{b}_{PLS,m}^2}{MS_E/N} \sim F_{1,N-M^*-1} \tag{4.18}$$

Since the square root of the $F_{1,\nu}$ distribution, with 1 and ν degrees of freedom, is equivalent the t_ν distribution, with ν degrees of freedom, Equation 4.18 can be rewritten as follows:

$$\frac{\hat{b}_{PLS,m}}{\sqrt{MS_E/N}} \sim t_{N-M^*-1} \tag{4.19}$$

Equation 4.19 is equivalent to Equation 4.14, demonstrating that the statistical significance analysis of the one-PLS component estimates corresponds to the MLR case.

4.2.2 Fractional factorial designs: 2^{k-p}

A fractional factorial design is an experimental design in which only a selected subset or fraction of the runs in the full factorial design are carried out. They require fewer samples than the full design without becoming unbalanced and spurious, like a design with missing values at random. In a fractional factorial design, some effects cannot be distinguished from others due to the confounding. Consequently, one single regression variable might be representing different confounded effects. The estimate associated to that regression variable refers to the sum of its confounded effects. In such a case, the regression variables present the same properties as a full factorial design (centered, equal variance (i.e., $N/(N-1)$) and orthogonality). Therefore, as discussed above, the solution given by the one-PLS component model is equivalent to the MLR solution, including the effect estimates and their statistical significance analysis as discussed above.

In addition to that, we suggest another option by augmenting the regression matrix with new columns allocating the effects to be estimated despite their confounding. This yields an augmented regression matrix, $\boldsymbol{X}^{aug}$. While estimation of fully confounded effects is not possible in MLR, it is possible with PLS, as it can handle rank-deficient data matrices (i.e., not full rank). This leads to a scenario where there may be C different eigenvalues, so more than one PLS component may need to be selected in order to explain the response variability related to regression variables. Besides, the statistical significance analysis according to Equation 4.17 will no longer be possible, but the estimation of the effects will be. In such a case, PLS will evenly distribute the value of the block of fully confounded effects among all of them. Therefore, an experimenter with lack of knowledge on experimental designs can obviate the details of the generator and the aliasing structure of the design. For that, they

will directly use the $\boldsymbol{X}^{aug}$ to estimate all effects by PLS, and then, they can detect the aliasing structure of the design by looking at the effects with the same estimate. Following that, the identical estimates will be pooled[4], and the statistical analysis of the effect groups will be carried out in a similar way as in a full factorial design, giving the same results as if they had used the $\boldsymbol{X}^{red}$ in MLR.

4.3 Traditional approaches applied to two-level factorial designs with missing runs

A common difficulty in using designed experiments is that there might be missing runs with respect to factorial designs. Hence, Cochran and Cox [53] proposed to estimate the missing values by minimizing the sum of squares for the sacrificed effects.

If R runs are missing, only $(N - R)$ effects (including the mean) can be estimated from $(N - R)$ remaining runs. Therefore, the user must choose which $(N - R - 1)$ of the original $(N - 1)$ effects, apart from the mean, are to be estimated, and which R effects are to be sacrificed. Suppose we sacrifice the $(N - R)$-th to $(N - 1)$-th effects. Minimization of the sum of squares of these R effects with respect to the R missing runs leads to the following system of R equations:

$$\sum_{i=N-R}^{N-1} a_{ij} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{y} = 0 \qquad j = 1, 2, ..., R \tag{4.20}$$

where $\boldsymbol{x}_i$ is the column vector of the contrast matrix $\mathbf{X}$ related to the i-th effect to be sacrificed, and $a_{i,j}$ is ± 1 according as the coefficient of the j-th missing observation in the i-th sacrificed effect is positive or negative. Thus, if the R by R matrix $\mathbf{A} = \{a_{i,j}\}$ is non-singular (i.e., the $a_{i,j}$ are such that the above R equations (Equation 4.20) are independent) then it is equivalent to:

$$\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{y} = 0 \qquad i = (N - R), ..., (N - 1) \tag{4.21}$$

So, an identical estimate would be obtained by equating to zero the sacrificed effects as Draper and Stoneman proposed [54]. The correct choice of the effect

[4]Note that, it is extremely unlikely to get exactly the same estimate for two effects that are not confounded.

to be sacrificed is, therefore, equivalent to finding effects for which matrix **A** is non-singular. Note that, solving for the system of equations (Equation 4.21) is not needed as the method will give the same estimates as a least squares solution (Equation 4.2) from a contrast matrix **X** obtained from the original one after deleting all rows corresponding to missing runs and all columns related to the effects to be sacrificed.

On the other hand, Kenett, Rahav, and Steinberg [65] proposed another approach based on bootstrapping to analyze designed experiments handling missing runs. The findings suggest that bootstrapping can contribute significantly to the design of experiments methodology in the presence of missing runs, however, the minimal requirement for applying the bootstrap is the presence of replicate observations (or a suitable model for generating replicates) at all levels of the experiment. Note that, replicates are not particularly common in most industrial DOE, and hence, this requirement is not assumed in the remainder of this chapter.

4.4 PLS applied to two-level factorial designs with missing runs

The approaches of Cochran and Cox [53] and Draper and Stoneman [54] have a good performance when missing runs are selected in advance in an optimal sense (scenario i)), however, they might lead to difficulties when having missing runs due to problems in their execution (scenario ii)) because the prior choice of sacrificed interactions might be unfeasible in practice. This situation could also create severe problems when using MLR directly.

For that reason, we propose a simple procedure addressing scenario i) (see Section 4.4.1) and scenario ii) (see Section 4.4.2) with just one method: PLS.

4.4.1 Lack of resources to execute a factorial design (scenario i))

4.4.1.1 Selection of runs to be omitted

In industrial applications of design of experiments, practical constraints in resources such as budget, time, material, etc. could lead to difficulties in running a complete factorial design [66]. In the line of Xampeny et al. [56], that is, with the aim of designing a simple and easy-to-understand method, we propose using an optimal design, but constraining the solution to a subset of the complete factorial design. The latter provides designs that do not change

factor levels (they are set at ± 1) and are, therefore, easily implementable for practitioners. We use the K-optimal criterion [67] based on the condition number of the information matrix, κ:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \tag{4.22}$$

where λ_{max} and λ_{min} are, respectively, the maximum and minimum eigenvalues of the information matrix. For this step, we recommend selecting only the main and two-factor interaction effects in the first model to not yield a rank-deficient information matrix (i.e., infinite condition number). The higher the condition number the closer a matrix is to being singular. Therefore, the combination of experiments yielding an information matrix with minimum condition number is recommended.

Anyway, one must be careful if estimating the resulting missing runs with respect to a complete factorial design because, as Box [68] warned, such estimation is a convenient computational approach that does not of course recover the information that has been lost. Hence, the collinearity, when exists, will be artificially removed. Indeed, Box [68] pointed out that one could start to feel uncomfortable with the estimation, for instance, in the case of having more than two observations missing from a sixteen run experiment. However, instead of considering the number of missing runs, we propose the experimenter to be warned by the condition number of the information matrix, since the same number of missing runs could yield information matrices with an unequal degree of collinearity. In this sense, Belsey [69] adopted a threshold of 30 for this condition number, from which one would expect moderate relations among regression variables being problematic in practice. Despite being a heuristic threshold, it performs well discerning ill-conditioned information matrices according to the authors' experience.

Table 4.1 presents a practical guide for quickly and simply selecting which runs to skip for the most popular designs according to the minimum condition number criterion. All these designs present a condition number less than 30. Grey cells indicate that for that particular combination of full or fractional design and number of runs to skip there is no resulting design with a condition number less than 30 and, therefore, there is no recommended design.

The lists of recommended combinations of missing runs when skipping more than one missing run are shown in Appendix 4.F.

Table 4.1: Practical guide for selecting the runs to skip for the most popular full and fractional factorial designs (2^k and 2^{k-p}). Condition number κ is shown in each case.

	Design					
Runs to skip	2^3	2^4	2^{4-1}_{IV}	2^{5-1}_{V}	2^{6-2}_{IV}	2^{7-3}_{IV}
1	Any ($\kappa = 7$)	Any ($\kappa = 3$)			Any ($\kappa = 7.5$)	Any ($\kappa = 15$)
2		Any ($\kappa = 3$)			($\kappa = 8$)	
3		($\kappa = 4$)				
4		($\kappa = 4$)				
5		($\kappa = 4$)				

4.4.1.2 How to estimate missing runs by PLS

After carrying out the recommended runs, regression variables are not orthogonal due to the missing runs, hence the one-PLS component model is no longer equivalent to the MLR solution. More than one component could be extracted in a PLS model to improve prediction. In fact, the PLS space has more than one relevant components ($C > 1$) for prediction [58]. However, PLS might run into overfitting issues when too many PLS components are used. For that reason, when the purpose of the model is prediction (e.g., estimating missing data), the cross-validation approach is widely used for determining the number of components based on checking the model's predictive ability [70].

Regarding the estimation, one could replace missing runs with the predictions from the fitted model, however, these estimated response values might yield to an overestimation of the correlation between $\mathbf{X}$ and $\boldsymbol{y}$. Thus, we suggest adding random noise to the predictions to overcome this problem according to Equation 4.23.

$$y^p = \boldsymbol{x}^{obs\,\mathrm{T}} \hat{\boldsymbol{b}}_{PLS} + e^{obs} \tag{4.23}$$

where y^p is the response prediction of a missing run ($\boldsymbol{x}^{mr}$) after adding random noise (e^{mr}). The noise is the prediction error obtained from a normal distribution with zero mean and variance $\sigma^2_{e^{obs}}$. The estimation of $\sigma_{e^{mr}}$, $s_{e^{mr}}$, is calculated as $s_{e_l^{obs}}$ in Equation 2.10. This is feasible if there are degrees of freedom to estimate the random noise. For that reason, although the PLS model can be fitted incorporating all effects, at this point it is advisable to select only main and two-factor interaction effects in the first model in order to have enough degrees of freedom to estimate random noise properly. Once missing

runs are estimated, one may fit a one-PLS component model by using data referring to the complete full or fractional factorial design as in Section 4.2.

4.4.2 Unexpected problems in the execution of some runs (scenario ii))

Once an experimental design has been planned, it may face unexpected problems such as: running out of resources during the experiments, having problems collecting data, not being able to reproduce the same conditions between experiments, having unfeasible conditions in some runs, outliers, etc. For these kinds of problems, it is assumed that the missingness mechanism is ignorable, what refers to the Missing At Random (MAR) case [71].

Note that, these missing runs are not selected in advance in an optimal sense and, hence, these designs may result in either well-conditioned matrix (i.e., $\kappa \leq 30$) or ill-conditioned information matrices (i.e., $\kappa > 30$). Regarding the first situation, the authors recommend following the missing runs estimation strategy as in Section 4.4.1.2 In the second situation, the missing estimation itself is no longer recommended because the estimation would remove the collinearity artificially. At this point, the user may consider carrying out some of the missing runs to overcome the confusion[5]. If possible, it will be preferable to carry out those experiments that minimize the condition number of the resulting information matrix according to Equation 4.22 yielding a well-conditioned matrix. If it is not possible, note that the potential ill-conditioned information matrix is no longer a problem with PLS due to its ability to handle correlated variables and, thereby, potential crucial effects can be all considered instead of having to select them (in contrast to stepwise MLR). Then, the interpretation and decision-making could be carried out with caution. Indeed, the confounding map is crucial to reveal the potential risk in the analysis due to the confusion. However, the recommendation for a practitioner without a solid training in experimental design would be not to proceed to the analysis in such a case.

[5]If there is a suspicion that something has not remained constant, it would be wise to define a new factor that designates whether a given run was in the original or in the new experiment.

4.5 Illustrative examples

4.5.1 First illustrative example: 2^4

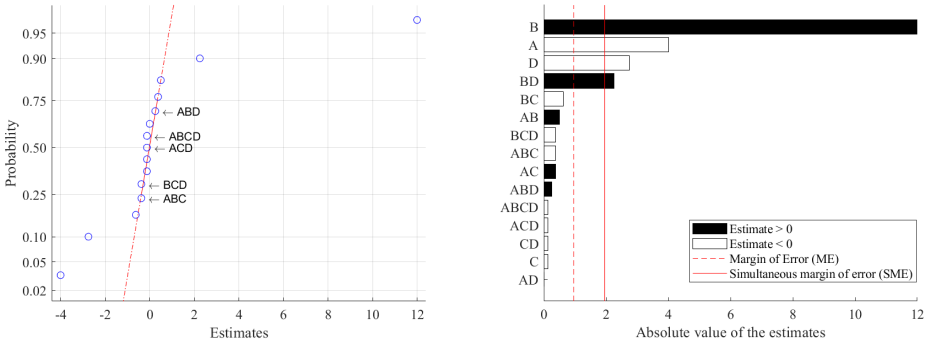
To illustrate how to use the proposed methodology, we use a 2^4 two-level full factorial design from Box et al. [5]. The four quantitative design variables are: amount of catalyst charge (A), temperature (B), pressure (C) and concentration of one of the reactants (D). The response variable $\boldsymbol{y}$ is the conversion rate at each of the 16 reaction conditions being centered in this work.

4.5.1.1 Full factorial design

Fitting one-PLS model and detecting negligible effects

First, a one-PLS model is fitted considering all effects. Since there are not replicates, it is necessary to include negligible effects on the residual in order to have degrees of freedom to estimate random noise. Thus, the Normal Probability Plot (NPP) and the Lenth's method [61] are used to determine the effects that are negligible (Figure 4.1).

In this example, interactions between three or more factors are considered negligible. This can be checked by representing all estimated effects on a NPP, showing that three and four level interactions take values close to zero and lie on the straight line of the negligible effects (see Figure 4.1a). To complement the NPP, a Pareto chart with the Lenth's method [61] can be also used to analyze the statistical significance of the effects (see Figure 4.1b). An effect whose bar extends beyond the simultaneous margin of error (SME) line is clearly active, one which does not extend beyond the margin of error (ME) line cannot be deemed active, and one in between is in a zone of uncertainty where a good argument can be made both for its being active and for its being a happenstance result of an inactive contrast. Figure 4.1b shows that three and four level interactions do not extend beyond the ME line.



(a) Normal probability plot of the estimated effects.

(b) Pareto chart for the absolute values of the estimated effects by the Lenth's method [61].

Figure 4.1

Refitting one-PLS model and getting p-values

The linear model refitted in the following will only contain main and two-factor interaction effects. Table 4.2 compares the estimates and the statistical significance of both MLR and PLS approaches. P-values for PLS solution are calculated based on both F-distribution (Equation 4.17) and CV/JK resampling approaches.

Table 4.2: Estimates (MLR and PLS) and p-values (MLR, PLS based on both the F-distribution and CV/JK resampling) for full factorial design. P-values<0.05 in bold.

		A	B	C	D	AB	AC	AD	BC	BD	CD
Estimates	MLR	-4.00	12.00	-0.13	-2.75	0.50	0.38	0.00	-0.63	2.25	-0.13
	PLS	-4.00	12.00	-0.13	-2.75	0.50	0.38	0.00	-0.63	2.25	-0.13
p-values	MLR	**0.00**	**0.00**	0.67	**0.00**	0.13	0.23	1.00	0.07	**0.00**	0.67
	$\text{PLS}_{F\text{-dist}}$	**0.00**	**0.00**	0.67	**0.00**	0.13	0.23	1.00	0.07	**0.00**	0.67
	$\text{MLR}_{\text{CV/JK}}$	0.22	**0.00**	0.97	0.40	0.88	0.91	1.00	0.85	0.49	0.97

Table 4.2 shows that not only the one-PLS component model gives the same estimates as MLR, but also p-values when considering the F-distribution approach. However, by CV/JK resampling approach no statistically significant relevance is detected in A, D and BD effects. Since there are no replicates in this kind of designs, the regression matrix in each perturbed model spans different spaces. For that reason, the uncertainty assessment of the individual model parameters estimated by jack-knifing is greater than expected, reducing the statistical power of the approach.

4.5.1.2 Lack of resources to execute a factorial design (scenario i))

The purpose of this section is to compare the proposed PLS-based approach (based on the condition number of the information matrix, followed by PLS estimation) with respect to the Xampeny et al.'s [56] approach (based on the Draper and Stoneman [54] estimation), when skipping runs due to a lack of resources to carry out a factorial design.

Selection of the runs to omit

Table 4.3 shows the number of recommended combinations in each case when skipping up to 5 runs in a 2^4 two-level full factorial design.

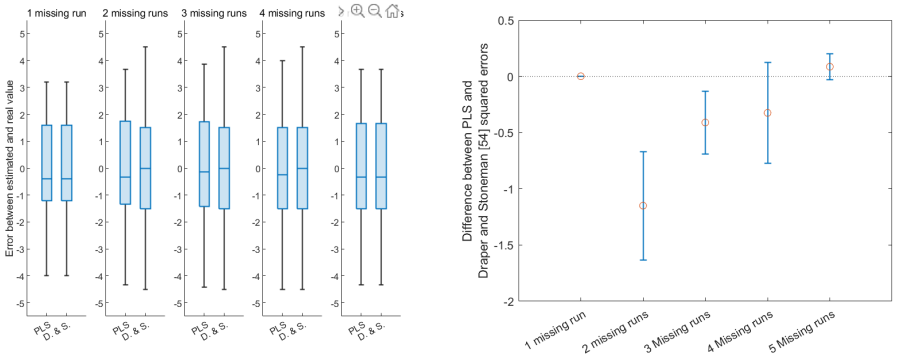
Table 4.3: Number of combinations and recommended combinations of missing runs when skipping up to 5 runs from 2^4 full factorial design based on the Xampeny et al.'s [56] and the condition number κ of the information matrix approaches. The latter is based on the lists of recommended combinations shown in Table 4.1 and Appendix 4.F.

Number of missing runs	Number of combinations of missing runs	Number of recommended combinations (Xampeny et al.'s [56] approach)	Number of recommended combinations (minimum κ)
1	16	16	16 ($\kappa = 3$)
2	120	80	80 ($\kappa = 4$)
3	560	160	160 ($\kappa = 4$)
4	1820	40	120 ($\kappa = 4$)
5	4368	16	16 ($\kappa = 4$)

Regardless the approach, Table 4.3 shows that the number of recommended combinations is the same for both approaches, except for the four missing runs case. Indeed, Xampeny et al. [56] highlighted this case as peculiar, because it does not follow their general rule and, finally, they ended up proposing 40 combinations instead of 120. Note that, the D-optimal criterion would give the same recommended combinations as the K-optimal criterion, as both require all the eigenvalues of the information matrix to be as equal as possible [72].

Estimation of missing runs

After selecting the appropriate combinations of missing runs for the different cases up to 5 missing runs from Table 4.3, the skipped runs were estimated by PLS (without adding random noise to make the results comparable), and the Draper and Stoneman's [54] approach. The first PLS model was fitted by CV including only main and two-factor interaction effects. Figure 4.2a shows



(a) Multiple boxplots of the error (i.e., the difference between estimated and real values) by approaches.

(b) 95% Confidence intervals for the difference in the mean of the squared errors of both approaches.

Figure 4.2

multiple boxplots of the difference between the estimated and the real values (i.e., the errors) of the missing runs in both approaches for the different cases. Figure 4.2b shows the 95% confidence intervals for the difference in the mean of the squared errors between PLS, and Draper and Stoneman [54] approaches for the different cases.

As Figure 4.2a shows, for any number of missing runs, no relevant discrepancy is found between both approaches in the distribution of differences between real and estimated values. However, Figure 4.2b does show statistically significant differences (p-values$<$ 0.05) for two and three missing runs cases, being the mean of the squared errors lower in both cases for the PLS approach. Thus, PLS performs equal or even slightly better than Draper and Stoneman [54] approach in the estimation of the missing runs in this example. Note that, once missing runs are estimated the user could analyze the analysis as a full factorial design as in Section 4.5.1.1.

4.5.1.3 Unexpected problems in the execution of some runs (scenario ii)

Assessment of the conditioning of the information matrix

As commented above, when missing runs are not planned in advance, different degrees of collinearity can be obtained, being the best case that obtained in scenario i). When the result of unplanned missing runs yields a well-conditioned matrix ($\kappa \leq 30$), the authors recommend following the missing runs estimation strategy as in Section 4.4.1.2. Nevertheless, when $\kappa > 30$, missing estimation is no longer recommended because the estimation would remove the collinearity artificially. Besides, collinearity may hinder the estimates as shown in the example below. Let us assume that, for example, runs 1, 2, 13 and 14 from the 2^4 full factorial design are missing. This design yields a rank-deficient information matrix (i.e., infinite condition number). Figure 4.3 shows the confounding map considering main and two-factor interaction effects. In this case, a variable selection method is required if using MLR (e.g., stepwise regression). In contrast, PLS allows considering all variables due to its ability to handle correlated variables. Table 4.4 shows the outcome of stepwise regression and PLS analysis of the incomplete factorial design.

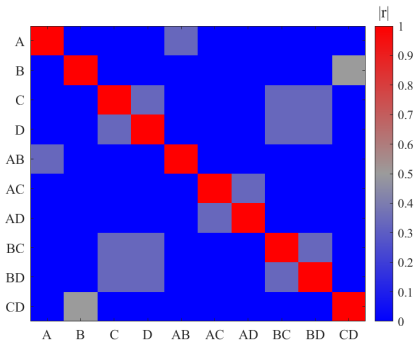


Figure 4.3: Confounding map for the full factorial design with four missing runs (1, 2, 13 and 14) and considering main and two-factor interaction effects.

What stands out in Table 4.4 is that BD and D effects are not selected by stepwise MLR even though they are statistically significant (p-value< 0.05) in the full factorial design. Note that, since the regression matrix is not full of rank, the stepwise regression method cannot select all effects. Indeed, Figure 4.3 shows that effects C, D, BC and BD are partially aliased, and therefore, not all of them are selected (in this case only C and BC are selected). Alterna-

Table 4.4: Estimated effects by the stepwise regression and PLS by CV for the full factorial design with four missing runs (1, 2, 13 and 14). Full design results are also shown for comparison. "$*$" means that such design variable is not selected.

	A	B	C	D	AB	AC	AD	BC	BD	CD
Stepwise MLR	-3.83	12.13	2.38	$*$	$*$	$*$	$*$	-3.13	$*$	$*$
PLS	-4.00	12.37	1.00	-1.63	0.50	0.31	-0.06	-1.75	1.13	-0.50
Full design	-4.00	12.00	-0.13	-2.75	0.50	0.38	0.00	-0.63	2.25	-0.13

tively, PLS is able to estimate all effects but, due to collinearity, the effects could not be estimated correctly either. For that reason, we propose using the confounding map (Figure 4.3) to improve interpretation and decision-making. A closer look at Figure 4.3 shows that C, D, BC and BD effects are moderately correlated with each other, being impossible to separate their direct causation with the response and, hence, their estimates differ from those obtained from the full design. If the confusion would have involved not two but only one main effect, the interpretation and decision-making could be carried out taking the same risk as in a fractional factorial design of resolution III (i.e., give more credit to the main effect than to the interaction effects). However, in this example, two main effects are involved making the interpretation of the results extremely risky.

Selection of the runs to carry out and following analysis as in the scenario i)

At this point, the user may consider carrying out some of the missing runs to overcome the confusion[6]. If possible, it will be preferable to carry out those experiments that minimize the condition number of the information matrix. In this example, if a new experiment could be carried out, all four possibilities would result in the same condition number, $\kappa = 10.53$, yielding a well-condition matrix and thus the missing runs estimation strategy as in the scenario i) could be applied. After carrying out the missing run 1, Table 4.5 shows the estimates and p-values after fitting one-PLS component model of the complete design, which has been previously filled in by means of a first PLS model by CV estimation (using only main and two-factor interaction effects) with or without random noise addition (Equation 4.23).

Table 4.5 shows that, after carrying out the missing run 1, the performance of the analysis improves significantly after estimating the remaining missing

[6]If there is a suspicion that something has not remained constant, it would be wise to define a new factor that designates whether a given run was in the original or in the new experiment.

Table 4.5: Estimates and p-values after estimating the three missing runs (2, 13 and 14) by means of the PLS strategy (without adding random noise and adding random noise). P-values < 0.05 in bold). Full design results are also shown for comparison.

		A	B	C	D	AB	AC	AD	BC	BD	CD
Estimates	Without noise	-4.00	12.38	-0.56	-3.19	0.50	0.31	-0.06	-0.19	2.69	-0.50
	Adding noise	-4.75	12.37	-0.74	-3.37	1.25	-0.62	-0.99	-0.01	2.87	-0.50
	Full Design	-4.00	12.00	-0.13	-2.75	0.50	0.38	0.00	-0.63	2.25	-0.13
p-values	Without noise	**0.00**	**0.00**	0.05	**0.00**	0.07	0.20	0.78	0.42	**0.00**	0.07
	Adding noise	**0.00**	**0.00**	0.39	**0.00**	0.17	0.47	0.26	0.99	**0.02**	0.56
	Full Design	**0.00**	**0.00**	0.67	**0.00**	0.13	0.23	1.00	0.07	**0.00**	0.67

data by means of PLS both adding and not adding random noise. When random noise is not added, some non-significant effects (C, AB, CD) have low p-values near 0.05. This inconsistency may be due to an overestimation of the correlation between $\mathbf{X}$ and $\boldsymbol{y}$ when estimating the missing runs. For that reason, adding random noise seems to be a more conservative approach despite the fact that the randomness in the imputation slightly affects the estimates.

4.5.2 Second illustrative example: 2^{6-2}

This is a simulated example corresponding to a 2^{6-2} two-level fractional factorial design. This is a popular design, being widely used in practice, since it allows estimating the 6 main effects by a 16-run design of resolution IV, and generators E = ABC and F = BCD. The simulator assumes a linear model according to the known values from Table 4.6, and random noise is added to the responses assuming a standardized normal distribution. This illustrative example has the advantage over the previous one that we can compare the effects estimates with respect to the known values. These estimates are made on the assumption that all interactions between three or more factors are negligible. However, two-factor interaction effects cannot be distinguished from others due to the confounding.

4.5.2.1 Fractional factorial design

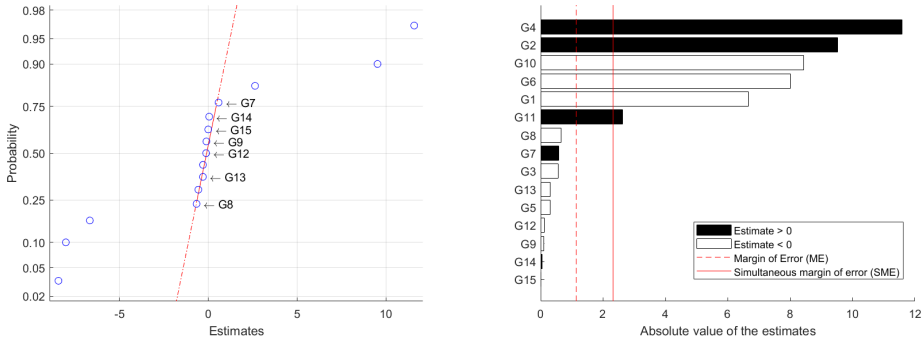
A priori, confounded effects should be represented by one single regression variable yielding to the $\mathbf{X}^{red}$. As commented, the PLS tool allows the experimenter to obtain the same results as MLR when using the same orthogonal regression matrix ($\mathbf{X}^{red}$). In addition to that, PLS also allows estimating fully confounded effects yielding to the $\mathbf{X}^{aug}$. Table 4.6 shows the estimates and the statistical significance of the proposed based-PLS approach.

Table 4.6: Known population effects and their estimates for the 2_{IV}^{6-2} fractional factorial design using the augmented design matrix. P-values < 0.05 in bold. "$*$" means that such design variable is not included in the statistical analysis.

2_{IV}^{6-2}	Known value	$\mathbf{X}^{aug}$ (PLS by CV)	Poleed efects	p-value
G1: A + BCE + DEF + ABCDF	-7, 0, 0, 0	-1.665 (x4)	-6.66	**0.00**
G2: B + ACE + CDF + ABDEF	10, 0, 0, 0	2.38 (x4)	9.52	**0.00**
G3: C + ABE + BDF + ACDEF	0, 0, 0, 0	-0.14 (x4)	-0.56	0.16
G4: D + AEF + BCF + ABCDE	11, 0, 0, 0	2.896 (x4)	11.584	**0.00**
G5: E + ABC + ADF + BCDEF	0, 0, 0, 0	-0.074 (x4)	-0.296	0.43
G6: F + ADE + BCD + ABCEF	-8, 0, 0, 0	2.002 (x4)	-8.008	**0.00**
G7: AB + CE + ACDF + BDEF	0, 0, 0, 0	0.145 (x4)	0.58	$*$
G8: AC + BE + ABDF + CDEF	0, 0, 0, 0	-0.165 (x4)	-0.66	$*$
G9: AD + EF + ABCF + BCDE	0, 0, 0, 0	-0.027 (x4)	-0.108	$*$
G10: AE + BC + DF + ABCDEF	-5, -4, 0, 0	-2.809 (x3)	-8.427	**0.00**
G11: AF + DE + ABCD + BCEF	0, 3, 0, 0	0.658 (x4)	2.632	**0.00**
G12: BD + CF + ABEF + ACDE	0, 0, 0, 0	-0.03 (x4)	-0.12	$*$
G13: BF + CD + ABDE + ACEF	0, 0, 0, 0	-0.075 (x4)	-0.3	$*$
G14: ABD + ACF + BEF + CDE	0, 0, 0, 0	0.013 (x4)	0.052	$*$
G15: ABF + ACD + BDE + CEF	0, 0, 0, 0	0 (x4)	0	$*$

Table 4.6 shows that PLS allows estimating all effects. Note that, as already commented in Section 4.2.2, fully confounded effects have exactly the same estimation. Thus, the aliasing structure of the design can be detected by looking at those effects with the same estimate. Following that, the identical estimates are pooled, and the statistical analysis of the pooled effects is carried out by means of either the NPP or Lenth's method [61] shown in Figure 4.4. This figure shows that the pooled effects G7, G8 G9, G12, G13, G14 and G5 take values close to zero and lie on the straight line of the negligible effects (see Figure 4.4a), and none of them extends beyond the margin of error (ME) (see Figure 4.4b). Finally, p-values of Table 4.6 are calculated based on F-distribution (Equation 4.17) after deleting the negligible pooled effects according to Figure 4.4 and refitting a one-PLS component model.

Note that, the pooled effects G3 and G5 are included in the statistical analysis for calculating p-values of Table 4.6 because they present main effects involved in potentially significant interactions.



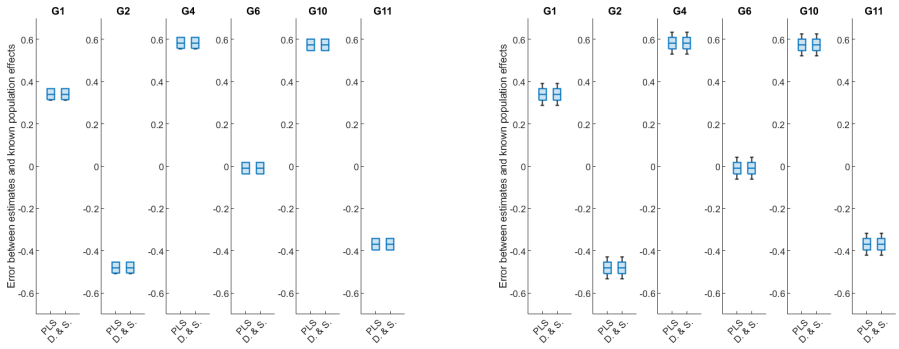
(a) Normal probability plot of the estimates. **(b)** Pareto chart for the absolute values of the estimates by the Lenth's method [61].

Figure 4.4

4.5.2.2 Lack of resources to execute a factorial design (scenario i))

In the case of having lack of resources to execute a factorial design (i.e., scenario i)), Xampeny et al.'s [56] addresses the cases of having one or two missing values. In the case of having one missing value, any of the 16 runs are recommended to be skipped. In the case of having two missing runs, only 64 pairs of the 120 possible of runs are recommended to be skipped (see Appendix 4.F). In both cases, the recommended list corresponds to the minimum condition number of the information matrix, which is less than 30. After selecting all recommended combinations up to 2 missing runs, those skipped runs were estimated and, then, the effects were estimated by means of the proposed PLS-based approach (without adding random noise to make the results comparable), and the Draper and Stoneman [54] approach. Figure 4.5 shows multiple boxplots of the difference between estimates and known population effects (i.e., the errors) for all active effects, distinguishing between one and two missing runs. For the PLS-based approach confounding effects are pooled to make results comparable.

Figures 4.5a and 4.5b shows that for any number of missing runs, the distribution of differences between estimates and known population effects are very similar for both approaches (no statistical significance discrepancies are found, p-values < 0.05).



(a) one missing run. **(b)** two missing runs.

Figure 4.5: Multiple boxplots of the difference between estimates and known population effects for all active effects with PLS, and Draper and Storeman [54] (D. & S.) approaches when having:

4.6 Discussion and conclusions

A novel framework to analyze two-level full and fractional factorial designs with one single technique, PLS, is proposed. This property is very attractive for practitioners since, to the best of our knowledge, no other statistical tool has comparable versatility. To provide an easy-to-follow route map for practitioners interested in using PLS to analyze design of experiments no matter their completeness, Figure 4.6 shows the proposed scheme.

In the case of a full factorial design, the one-PLS component yields the same analytical solution as MLR, not only in the estimation of the effects, but also in their statistical significance analysis. Besides, when data from a fractional factorial design is analyzed, PLS also allows the possibility of including and estimating straightforwardly all effects in the model despite their confounding, in contrast to MLR. When dealing with lack of resources to execute a factorial design (scenario i)) we propose an alternative method to Xampeny's et al [56] approach in order to decide which runs to omit based on the condition number of the information matrix. The potential ill-conditioned information matrix in scenario ii) is no longer a problem with PLS due to its ability to handle correlated data. In both scenarios, one could use a PLS model to estimate the value of the missed runs (by getting the prediction from the PLS model and adding random noise as in Equation 4.23) yielding an orthogonal factorial design which can be analyzed by the one-PLS component model. This proposed PLS-based approach does not require selecting a priori the effects to

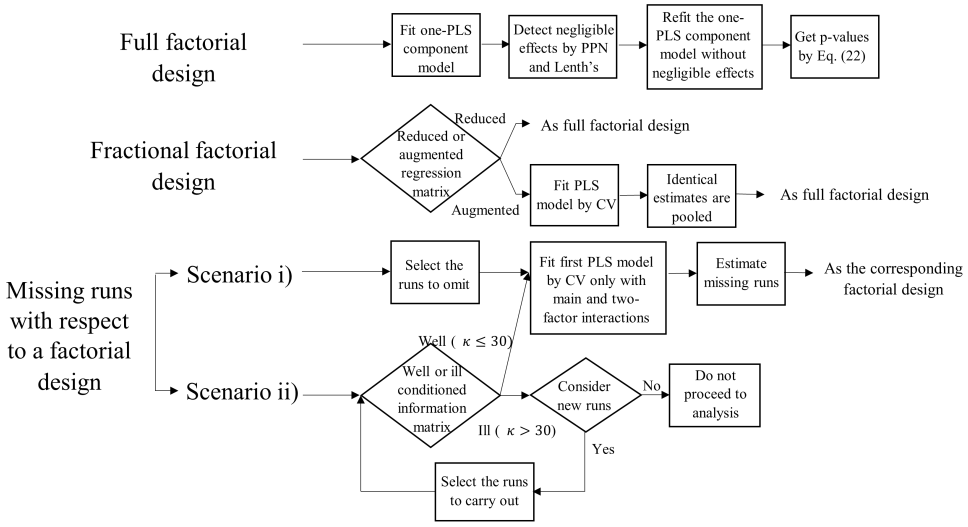


Figure 4.6: Route map summarizing how to use PLS when dealing with data from design of experiments.

be sacrificed, and its performance is similar or slightly better to the Draper and Stoneman [54] approach. Nevertheless, in the case of having ill-conditioned information matrix we do not recommend using any of the approaches discussed in this chapter to avoid obtaining severe estimation biases. Alternatively, the user may consider carrying out some of the missing runs until getting a well-conditioned information matrix.

Practitioners can resort to using only PLS instead of using different methods (MLR, Xampeny's et al [56]) and Draper and Stoneman [54]) without sacrificing accuracy, reliability, or interpretability. Indeed, as shown in this chapter, PLS can be used to estimate the effects no matter of the completeness of experimental data, being MLR a particular case of it. In addition to that, another type of DOE application concerns the mixture designs, where data analysis becomes more challenging as the mixture factors are correlated due to the restrictions that must be fulfilled. Hence, the MLR. method is not directly applicable, but a special model form needs to be used [73]. By contrast, PLS regression works well because analyzing correlated mixture variables is not a problem [74–76]. Therefore, PLS is not only a powerful tool when dealing with non-experimental data (i.e., observational data), but also when dealing with data from experimental designs.

Appendices

4.A Definition of PLS coefficients from NIPALS algorithm in a full factorial design

Let us consider two centered data arrays $\mathbf{X}$ ($N \times M$) and $\boldsymbol{y}$ ($N \times 1$), where columns of $\mathbf{X}$ have the same variance and are orthogonal to each other. To calculate the parameters of the PLS model in a sequential manner, the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm can be used [77]. The algorithm is as follows:

1:	Start: set $\boldsymbol{u}$ to $\boldsymbol{y}$	▷ Initialization
2:	**while** there is no convergence on $\boldsymbol{t}$ or $\boldsymbol{u}$ **do**	
3:	$\boldsymbol{w}^{old} = \mathbf{X}^{\mathrm{T}}\boldsymbol{u} / (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})$	▷ Compute $\mathbf{X}$ block weights
4:	$\boldsymbol{w}^{new} = \boldsymbol{w}^{old} / \|\boldsymbol{w}^{old}\|$	▷ Normalize weights vectors
5:	$\boldsymbol{t} = \mathbf{X}\boldsymbol{w}^{new}$	▷ Compute $\mathbf{X}$ block scores
6:	$q = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{t} / (\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t})$	▷ Compute $\boldsymbol{y}$ weights
7:	$\boldsymbol{u} = \boldsymbol{y}/q$	▷ Compute $\boldsymbol{y}$ scores
8:	**end while**	
9:	$\boldsymbol{p} = (\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t})^{-1} \mathbf{X}^{\mathrm{T}}\boldsymbol{t}$	▷ Compute $\mathbf{X}$ block loadings
10:	$\mathbf{E} = \mathbf{X} - \boldsymbol{t}\boldsymbol{p}^{\mathrm{T}}$ and $\boldsymbol{f} = \boldsymbol{y} - \boldsymbol{t}q$	▷ Compute residual matrices

The next set of iterations starts with the new $\mathbf{X}$ and $\boldsymbol{y}$ arrays as the residual arrays, $\mathbf{E}$ and $\boldsymbol{f}$, respectively.

For the first component (note that q is a scalar), one can substitute the equation of step 3 to the equation of the step 4:

$$\boldsymbol{w}^{new} = \frac{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})^{-1} \mathbf{X}^{\mathrm{T}}\boldsymbol{u}}{\sqrt{(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})^{-1} \boldsymbol{u}^{\mathrm{T}}\mathbf{X} (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})^{-1} \mathbf{X}^{\mathrm{T}}\boldsymbol{u}}} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{u}}{\sqrt{\boldsymbol{u}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{u}}} \tag{4.A.1}$$

and by substituting the equation of the step 7 in Equation 4.A.1:

$$\boldsymbol{w}^{new} = \frac{\mathbf{X}^{\mathrm{T}} (\boldsymbol{y}/q)}{\sqrt{(\boldsymbol{y}/q)^{\mathrm{T}} \mathbf{X}\mathbf{X}^{\mathrm{T}} (\boldsymbol{y}/q)}} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\sqrt{\boldsymbol{y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|} \tag{4.A.2}$$

Since $\boldsymbol{w}^{new}$ is of length one ($\boldsymbol{w}^{new\mathrm{T}}\boldsymbol{w}^{new} = 1$), it is deduced from Equation 4.A.2:

$$\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\| = \boldsymbol{w}^{new\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} \tag{4.A.3}$$

On the other hand, one can substitute the equation of step 5 to the equation of the step 6:

$$q = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}^{new}}{\boldsymbol{w}^{new\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}^{new}} \tag{4.A.4}$$

and by substituting Equation 4.3 in Equation 4.A.4:

$$q = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}^{new}}{\boldsymbol{w}^{new\mathrm{T}}s_{\boldsymbol{x}}^2\,(N-1)\,\mathbf{I}^{M\times M}\boldsymbol{w}^{new}} \tag{4.A.5}$$

Since $\boldsymbol{w}^{new}$ is of length one ($\boldsymbol{w}^{new\mathrm{T}}\mathbf{I}^{M\times M}\boldsymbol{w}^{new} = 1$), it is deduced from Equation 4.A.5:

$$q = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}^{new}}{s_{\boldsymbol{x}}^2\,(N-1)} = \frac{\boldsymbol{w}^{new\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{s_{\boldsymbol{x}}^2\,(N-1)} \tag{4.A.6}$$

Substituting Equation 4.A.3 in Eq. Equation 4.A.6:

$$q = \frac{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|}{s_{\boldsymbol{x}}^2\,(N-1)} \tag{4.A.7}$$

Note that, if $\mathbf{X}$ comes from a two-level full factorial design coded by minus and plus values, $s_{\boldsymbol{x}}^2 = N/(N-1)$ and Equation 4.A.7 is expressed as:

$$q = \frac{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|}{N} \tag{4.A.8}$$

Besides, after each PLS component is calculated the $\mathbf{X}$-matrix is deflated according to step 10, making the PLS model alternatively be expressed in weights $\boldsymbol{w}^{new}$ referring to the residuals after previous dimension, instead of the $\mathbf{X}$-variables themselves (as $\boldsymbol{w}^{*}$ does according to Equation 2.1). However, for

the first component, $\mathbf{X}$ has not yet deflated and, therefore, both $\boldsymbol{w}^{new}$ and $\boldsymbol{w}^{*}$ relate directly to $\mathbf{X}$ (i.e., $\boldsymbol{w}^{new} = \boldsymbol{w}^{*}$). Thus, for the first component, Equation 4.A.7 can be expressed as follows:

$$\boldsymbol{w}^{*} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|} \tag{4.A.9}$$

4.B Definition of PLS coefficients from the criterion of maximum variance in a full factorial design

Let us consider two centered data arrays $\mathbf{X}$ and $\boldsymbol{y}$, where columns of $\mathbf{X}$ have the same variance. From NIPALS algorithm (Appendix 4.A), it follows that:

$$\boldsymbol{t}^{\mathrm{T}} = \boldsymbol{w}^{new\,\mathrm{T}}\mathbf{X}^{\mathrm{T}} \tag{4.B.1}$$

Besides, since PLS maximizes the covariance between vector $\boldsymbol{t}$ and $\boldsymbol{y}$, the following expressions are obtained for the first PLS component:

$$\max cov\left(\boldsymbol{t},\boldsymbol{y}\right) \propto \max\left(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{y}\right) = \max\left(\boldsymbol{w}^{new\,\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\right) \tag{4.B.2}$$

The last equivalence is the scalar product of two vectors, the unitary vector $\boldsymbol{w}^{new}$ and $(\mathbf{X}^{\mathrm{T}}\boldsymbol{y})$, being maximum if both vectors are parallel (Equation 4.B.3).

$$max\frac{\boldsymbol{w}^{new}}{\|\boldsymbol{w}^{new}\|} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|} \tag{4.B.3}$$

Since $\boldsymbol{w}^{new}$ is of length one ($\|\boldsymbol{w}^{new}\| = 1$):

$$max\boldsymbol{w}^{new} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|} \tag{4.B.4}$$

As commented in Appendix 4.A, for the first PLS component: ($\boldsymbol{w}^{new} = \boldsymbol{w}^{*}$) and, therefore, Equation 4.B.4 is expressed as:

$$\boldsymbol{w}^{*} = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|} \tag{4.B.5}$$

4.C The second PLS component lacks predictive ability in a full factorial design

Let us consider two centered data arrays $\mathbf{X}$ ($N \times M$) and $\boldsymbol{y}$ ($N \times 1$), where columns of $\mathbf{X}$ have the same variance and are orthogonal to each other. For the first component, it is deduced from NIPALS algorithm (Appendix 4.A) that residual arrays are obtained as:

$$\mathbf{E} = \mathbf{X} - \boldsymbol{t}\boldsymbol{p}^{\mathrm{T}} \tag{4.C.1}$$

$$\boldsymbol{f} = \mathbf{X} - \boldsymbol{t}q \tag{4.C.2}$$

Then, the iteration of the second component starts with the new $\mathbf{X}$ and $\boldsymbol{y}$ arrays as $\mathbf{E}$ and $\boldsymbol{f}$, respectively. Thus, the second PLS component lacks predictive ability if the vector $\boldsymbol{f}$ is orthogonal to the subspace $\mathbf{E}$, i.e., $\mathbf{E}^{\mathrm{T}}\boldsymbol{f} = 0$.

From Equation 4.C.1, $\mathbf{E}^{\mathrm{T}}\boldsymbol{f}$ can be expressed as:

$$\mathbf{E}^{\mathrm{T}}\boldsymbol{f} = \mathbf{X}^{\mathrm{T}}\boldsymbol{f} - \boldsymbol{p}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{f} \tag{4.C.3}$$

and substituting the equation of the step 5 (Appendix 4.A) in Equation 4.C.3:

$$\mathbf{E}^{\mathrm{T}}\boldsymbol{f} = \mathbf{X}^{\mathrm{T}}\boldsymbol{f} - \boldsymbol{p}\boldsymbol{w}^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{f} \tag{4.C.4}$$

On the other hand, Equations 4.4 and 4.10 demonstrate that the solution to the one-PLS component model corresponds to the MLR solution (i.e., $\hat{\boldsymbol{b}}_{PLS} = \hat{\boldsymbol{b}}_{MLR}$) and, hence, the residual vector expressing the deviation between measured and predicted response values are also equivalents (i.e., $\boldsymbol{f} = \boldsymbol{e}$). Thus, since in MLR $\boldsymbol{e}$ is orthogonal to the subspace $\mathbf{X}$, it can be deduced that $\mathbf{X}^{\mathrm{T}}\boldsymbol{f} = 0$. Therefore, Equation 4.C.4 is expressed as:

$$\mathbf{E}^{\mathrm{T}}\boldsymbol{f} = \mathbf{0}\boldsymbol{p}\boldsymbol{w}^{*\mathrm{T}}\mathbf{0} = \mathbf{0} \tag{4.C.5}$$

proving that $\boldsymbol{f}$ is orthogonal to the subspace $\mathbf{E}$ and, therefore, the second PLS component lacks predictive ability.

4.D Illustrating the latent space in a full factorial design

Consider a two-level full factorial design of experiment with three factors A, B and C ($N = 8$), and suppose that the population regression coefficient vector for the main effects is $\boldsymbol{b}_{MLR}^{\mathrm{T}} = (5, -2, 4)^{\mathrm{T}}$. To ease the graphical representation interactions effects and random noise are not considered. Thus, the response variable $\boldsymbol{y}$ is calculated as $\mathrm{X}\boldsymbol{b}_{MLR}$, where X is the (8×3) contrast matrix coded by -1 and $+1$ values. The parameters of the first PLS component, $\boldsymbol{p}$, $\boldsymbol{w}^*$ and $\boldsymbol{t}$, are calculated according to NIPALS algorithm (Appendix 4.A).

In this simple case, the X can be represented as 8 points in the 3-dimensional regressor space where each column of X defines one coordinate axis. The PLS model defines a 1-dimensional hyper-plane (i.e., it is defined by one line). The direction coefficients of this line are $\boldsymbol{p}$. The coordinates of each run when its data are projected down on this line are defined by $\boldsymbol{t}$ [25]. These positions can be represented in the 3-dimensional regressor space as $\hat{\mathbf{X}} = \boldsymbol{t}\boldsymbol{p}^{\mathrm{T}}$. This is illustrated in Figure 4.D.1.

Note that, according to the equation of the step 9 (Appendix 4.A), $\boldsymbol{p}$ can be expressed as:

$$\boldsymbol{p} = \left(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t}\right)\mathbf{X}^{\mathrm{T}}\boldsymbol{t} \tag{4.D.1}$$

and substituting equation of the step 5 (Appendix 4.A) in Equation 4.D.1:

$$\boldsymbol{p} = \left(\boldsymbol{w}^{new\,\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}^{new}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}^{new} \tag{4.D.2}$$

Besides, in a two-level full factorial design $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is equivalent to $8\mathbf{I}^{3\times 3}$ (from Equation 4.3) and, hence, Equation 4.D.2 can be expressed as:

$$\boldsymbol{p} = \left(\boldsymbol{w}^{new\,\mathrm{T}}8\mathbf{I}^{3\times 3}\boldsymbol{w}^{new}\right)^{-1}8\mathbf{I}^{3\times 3}\boldsymbol{w}^{new} = \left(\boldsymbol{w}^{new\,\mathrm{T}}\boldsymbol{w}^{new}\right)^{-1}\boldsymbol{w}^{new} \tag{4.D.3}$$

Since $\boldsymbol{w}^{new}$ is of length one ($\boldsymbol{w}^{new\,\mathrm{T}}\boldsymbol{w}^{new} = 1$), $\boldsymbol{p}$ is equivalent to $\boldsymbol{w}^{new}$ and, consequently, equivalent to $\boldsymbol{w}^*$ (see Appendix 4.A), where $\boldsymbol{w}^*$ corresponds to the regression coefficients multiplied by a scalar (Equation 4.9) . For that reason, the direction of the latent space, $\boldsymbol{p}$, is consistent with the direction of the maximum response gradient as is shown in Figure 4.D.1.

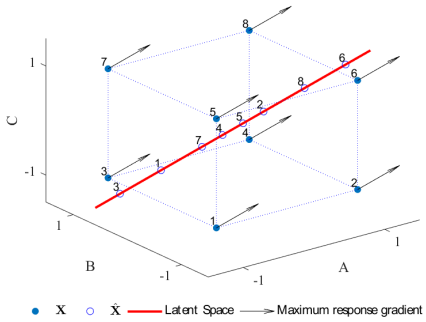


Figure 4.D.1: The geometric representation of PLS in the regressor space for a two-level full factorial design with three factors A, B and C and vector for the main effects $\boldsymbol{b}_{MLR}^{\mathrm{T}} = (5, -2, 4))^{\mathrm{T}}$.

4.E Partitioning of the sum of squares for PLS in a full factorial design

Let us consider a two-level full factorial design 2^k, the sum of squares of a particular m effect SS_m is expressed as:

$$SS_m = I\left[\left(\overline{y}^{+1} - \overline{\overline{y}}\right)^2 \left(\overline{y}^{-1} - \overline{\overline{y}}\right)^2\right] \tag{4.E.1}$$

where I is the number of samples at each level (i.e., $N/2$), $\overline{y}^{+1}$ is the average for level $+1$, $\overline{y}^{-1}$ is the average for level -1 and $\overline{\overline{y}}$ is the total average. Using $\overline{y}^{+1} = A$ and $\overline{y}^{-1} = B$, and replacing the value $\overline{\overline{y}}$ in Equation 4.E.1 by $A+B/2$ (mean of two averages) gives:

$$\begin{aligned} SS_m &= I\{\,[A - (A+B)/2]^2 + [B - (A+B)/2]^2\,\} \\ &= I\{\,[(A-B)/2]^2 + [(B-A)/2]^2\,\} \\ &= I\{\,[(A^2 + B^2 - 2AB)/4] + [(A^2 + B^2 - 2AB)/4]\,\} \\ &= (I/2)(A^2 + B^2 - 2AB) \end{aligned} \tag{4.E.2}$$

Besides, using $\boldsymbol{a} = \mathbf{X}^{\mathrm{T}}\boldsymbol{y}$, it is deduced for the m-th effect that:

$$a_m = I\left(\overline{y}^{+1} - \overline{y}^{-1}\right) = I\,(A - B) \tag{4.E.3}$$

The square of Equation 4.E.3 is $a_m^2 = I^2 \left(A - B\right)^2$, and dividing by $2I$ gives:

$$\frac{a_m^2}{2I} = (I/2)(A-B)^2 = (I/2)(A^2 + B^2 - 2AB) \tag{4.E.4}$$

Comparing Equation 4.E.2 and Equation 4.E.3:

$$SS_m = \frac{a_m^2}{2I} = \frac{a_m^2}{N} \tag{4.E.5}$$

On the other hand, from Appendix 4.A (Equation 4.A.7) is deduced that:

$$q = \frac{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|}{s_{\boldsymbol{x}}^2\,(N-1)} = \frac{a}{\frac{N}{N-1}(N-1)} = \frac{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}}}{N} \tag{4.E.6}$$

and from Appendix 4.B (Equation 4.B.5):

$$\boldsymbol{w}^* = \frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}{\|\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\|} = \frac{\boldsymbol{a}}{\|a\|} = \frac{\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}}} \tag{4.E.7}$$

Hence, for the m-th effect:

$$w_m^* = \frac{a_m}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}}} = \frac{a_m}{qN} \tag{4.E.8}$$

Thus, by relating Equation 4.E.5 and Equation 4.E.8:

$$SS_m = \frac{{w_m^*}^2 q^2 N^2}{N} = {w_m^*}^2 q^2 N \tag{4.E.9}$$

The residual sum of squares SS_E can be easily calculated by subtracting the explained sum of squares ($\sum_{m=1}^{M} SS_m$) from the total sum of squares SS_T as Equation 4.E.10.

$$SS_E = SS_T - \sum_{m=1}^{M} SS_m = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - q^2 N \tag{4.E.10}$$

Note that, $\boldsymbol{w}^*$ is of length one: $(\sum_{m=1}^{M} {w_m^*}^2 = 1)$.

4.F Lists of recommended combinations of missing runs for the most popular designs

This appendix presents the lists of recommended combinations of missing runs for the most popular designs.

Table 4.F.1: List of the 80 recommended combinations when skipping two runs in a 2^4.

1, 4	2, 3	3, 5	4, 6	5, 10	6, 15	8, 12	10, 15
1, 6	2, 5	3, 6	4, 7	5, 11	6, 16	8, 13	10, 16
1, 7	2, 7	3, 8	4, 9	5, 14	7, 9	8, 14	11, 13
1, 8	2, 8	3, 9	4, 10	5, 15	7, 11	8, 15	11, 14
1, 10	2, 9	3, 10	4, 11	5, 16	7, 12	9, 12	11, 16
1, 11	2, 11	3, 12	4, 14	6, 7	7, 13	9, 14	12, 13
1, 12	2, 12	3, 13	4, 15	6, 9	7, 14	9, 15	12, 14
1, 13	2, 13	3, 15	4, 16	6, 10	7, 16	9, 16	12, 15
1, 14	2, 14	3, 16	5, 8	6, 12	8, 10	10, 11	13, 16
1, 15	2, 16	4, 5	5, 9	6, 13	8, 11	10, 13	14, 15

Table 4.F.2: List of the 160 recommended combinations when skipping three runs in a 2^4.

1, 4, 6	1, 8, 15	2, 5, 16	3, 5, 8	3, 10, 16	4, 9, 15	5, 14, 15	7, 12, 13
1, 4, 7	1, 10, 11	2, 7, 9	3, 5, 9	3, 12, 13	4, 9, 16	6, 7, 9	7, 12, 14
1, 4, 10	1, 10, 13	2, 7, 11	3, 5, 10	3, 12, 15	4, 10, 11	6, 7, 12	7, 13, 16
1, 4, 11	1, 10, 15	2, 7, 12	3, 5, 15	3, 13, 16	4, 10, 15	6, 7, 13	8, 10, 11
1, 4, 14	1, 11, 13	2, 7, 13	3, 5, 16	4, 5, 9	4, 10, 16	6, 7, 16	8, 10, 13
1, 4, 15	1, 11, 14	2, 7, 14	3, 6, 9	4, 5, 10	4, 11, 14	6, 9, 12	8, 10, 15
1, 6, 7	1, 12, 13	2, 7, 16	3, 6, 10	4, 5, 11	4, 11, 16	6, 9, 15	8, 11, 13
1, 6, 10	1, 12, 14	2, 8, 11	3, 6, 12	4, 5, 14	4, 14, 15	6, 9, 16	8, 11, 14
1, 6, 12	1, 12, 15	2, 8, 12	3, 6, 13	4, 5, 15	5, 8, 10	6, 10, 13	8, 12, 13
1, 6, 13	1, 14, 15	2, 8, 13	3, 6, 15	4, 5, 16	5, 8, 11	6, 10, 15	8, 12, 14
1, 6, 15	2, 3, 5	2, 8, 14	3, 6, 16	4, 6, 7	5, 8, 14	6, 10, 16	8, 12, 15
1, 7, 11	2, 3, 8	2, 9, 12	3, 8, 10	4, 6, 9	5, 8, 15	6, 12, 13	8, 14, 15
1, 7, 12	2, 3, 9	2, 9, 14	3, 8, 12	4, 6, 10	5, 9, 14	6, 12, 15	9, 12, 14
1, 7, 13	2, 3, 12	2, 9, 16	3, 8, 13	4, 6, 15	5, 9, 15	6, 13, 16	9, 12, 15
1, 7, 14	2, 3, 13	2, 11, 13	3, 8, 15	4, 6, 16	5, 9, 16	7, 9, 12	9, 14, 15
1, 8, 10	2, 3, 16	2, 11, 14	3, 9, 12	4, 7, 9	5, 10, 11	7, 9, 14	10, 11, 13
1, 8, 11	2, 5, 8	2, 11, 16	3, 9, 15	4, 7, 11	5, 10, 15	7, 9, 16	10, 11, 16
1, 8, 12	2, 5, 9	2, 12, 13	3, 9, 16	4, 7, 14	5, 10, 16	7, 11, 13	10, 13, 16
1, 8, 13	2, 5, 11	2, 12, 14	3, 10, 13	4, 7, 16	5, 11, 14	7, 11, 14	11, 13, 16
1, 8, 14	2, 5, 14	2, 13, 16	3, 10, 15	4, 9, 14	5, 11, 16	7, 11, 16	12, 14, 15

Table 4.F.3: List of the 120 recommended combinations when skipping four runs in a 2^4.

1, 4, 6, 7	1, 8, 10, 13	2, 5, 8, 14	3, 5, 8, 10	3, 10, 13, 16	4, 9, 14, 15
1, 4, 6, 10	1, 8, 10, 15	2, 5, 9, 14	3, 5, 8, 15	4, 5, 9, 14	4, 10, 11, 16
1, 4, 6, 15	1, 8, 11, 13	2, 5, 9, 16	3, 5, 9, 15	4, 5, 9, 15	5, 8, 10, 11
1, 4, 7, 11	1, 8, 11, 14	2, 5, 11, 14	3, 5, 9, 16	4, 5, 9, 16	5, 8, 10, 15
1, 4, 7, 14	1, 8, 12, 13	2, 5, 11, 16	3, 5, 10, 15	4, 5, 10, 11	5, 8, 11, 14
1, 4, 10, 11	1, 8, 12, 14	2, 7, 9, 12	3, 5, 10, 16	4, 5, 10, 15	5, 8, 14, 15
1, 4, 10, 15	1, 8, 12, 15	2, 7, 9, 14	3, 6, 9, 12	4, 5, 10, 16	5, 9, 14, 15
1, 4, 11, 14	1, 8, 14, 15	2, 7, 9, 16	3, 6, 9, 15	4, 5, 11, 14	5, 10, 11, 16
1, 4, 14, 15	1, 10, 11, 13	2, 7, 11, 13	3, 6, 9, 16	4, 5, 11, 16	6, 7, 9, 12
1, 6, 7, 12	1, 12, 14, 15	2, 7, 11, 14	3, 6, 10, 13	4, 5, 14, 15	6, 7, 9, 16
1, 6, 7, 13	2, 3, 5, 8	2, 7, 11, 16	3, 6, 10, 15	4, 6, 7, 9	6, 7, 12, 13
1, 6, 10, 13	2, 3, 5, 9	2, 7, 12, 13	3, 6, 10, 16	4, 6, 7, 16	6, 7, 13, 16
1, 6, 10, 15	2, 3, 5, 16	2, 7, 12, 14	3, 6, 12, 13	4, 6, 9, 15	6, 9, 12, 15
1, 6, 12, 13	2, 3, 8, 12	2, 7, 13, 16	3, 6, 12, 15	4, 6, 9, 16	6, 10, 13, 16
1, 6, 12, 15	2, 3, 8, 13	2, 8, 11, 13	3, 6, 13, 16	4, 6, 10, 15	7, 9, 12, 14
1, 7, 11, 13	2, 3, 9, 12	2, 8, 11, 14	3, 8, 10, 13	4, 6, 10, 16	7, 11, 13, 16
1, 7, 11, 14	2, 3, 9, 16	2, 8, 12, 13	3, 8, 10, 15	4, 7, 9, 14	8, 10, 11, 13
1, 7, 12, 13	2, 3, 12, 13	2, 8, 12, 14	3, 8, 12, 13	4, 7, 9, 16	8, 12, 14, 15
1, 7, 12, 14	2, 3, 13, 16	2, 9, 12, 14	3, 8, 12, 15	4, 7, 11, 14	9, 12, 14, 15
1, 8, 10, 11	2, 5, 8, 11	2, 11, 13, 16	3, 9, 12, 15	4, 7, 11, 16	10, 11, 13, 16

Table 4.F.4: List of the 16 recommended combinations when skipping five runs in a 2^4.

1, 4, 6, 10, 15	1, 8, 12, 14, 15	2, 7, 9, 12, 14	3, 6, 10, 13, 16
1, 4, 7, 11, 14	2, 3, 5, 9, 16	2, 7, 11, 13, 16	4, 5, 9, 14, 15
1, 6, 7, 12, 13	2, 3, 8, 12, 13	3, 5, 8, 10, 15	4, 5, 10, 11, 16
1, 8, 10, 11, 13	2, 5, 8, 11, 14	3, 6, 9, 12, 15	4, 6, 7, 9, 16

Table 4.F.5: List of the 64 recommended combinations when skipping two runs in a 2^{6-2}_{IV}

1,3	2,3	3,7	4,9	5,15	7,11	9,11	11,15
1,4	2,4	3,8	4,1	5,16	7,12	9,12	11,16
1,5	2,5	3,9	4,15	6,7	7,13	9,13	12,15
1,6	2,6	3,1	4,16	6,8	7,14	9,14	12,16
1,11	2,11	3,15	5,7	6,9	8,11	10,11	13,15
1,12	2,12	3,16	5,8	6,1	8,12	10,12	13,16
1,13	2,13	4,7	5,9	6,15	8,13	10,13	14,15
1,14	2,14	4,8	5,1	6,16	8,14	10,14	14,16

Chapter 5

Defining multivariate raw materials specifications

Part of the content of this chapter has been included in:

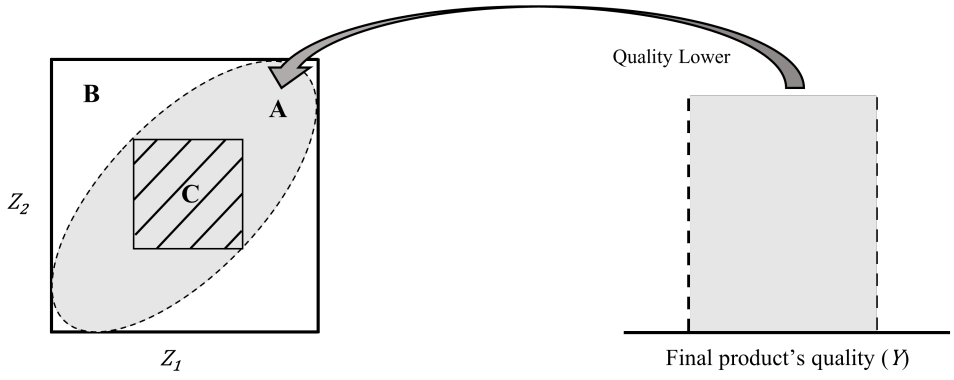
[14] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Defining multivariate raw material specifications in industry 4.0," *Chemometrics and Intelligent Laboratory Systems*, vol. 225, 2022, ISSN: 18733239. DOI: 10.1016/j.chemolab.2022.104563

5.1 Introduction

Raw materials properties are usually considered as Critical Input Parameters (CIPs) because their variability has an impact on Critical Quality Attributes (CQAs) of the final product. Thus, as commented by Duchesne and MacGregor [78], the development of specification regions for raw materials is crucial to ensure the desired quality of the product. In this chapter, we propose a novel method to define meaningful raw material specifications, namely a region that is expected to provide assurance of quality with a certain confidence level for the CQAs. Our approach overcomes the drawbacks of the current industrial practice of setting univariate specifications for each property of raw material and allows the producer to make a decision on accepting or rejecting a raw material batch based on the confidence of producing good product quality prior to starting the manufacturing process.

Despite their importance, specifications are usually defined in an arbitrary way based mostly on subjective past experience, instead of using a quantitative objective description of their impact on CQAs. Furthermore, in many cases, univariate specifications on each property are designated, with the implicit assumption that these properties are independent from one another. As a consequence, significant amounts of raw materials whose properties are correlated may be misclassified, as appropriate or otherwise, when univariate specifications are considered, as it is shown in Figure 5.1.

Let us consider a raw material with two correlated properties, Z_1 and Z_2 (see Figure 5.1) used in the manufacturing of a particular product with final product quality Y. The elliptical region "A" is the true multivariate region in Z_1 and Z_2 such that any batch of raw material used with Z_1 and Z_2 properties falling within it will provide good product quality (i.e., within the Y quality specification limits). On the contrary, raw material batches with properties outside this elliptical region correspond to unacceptable raw material batches, as they lead to poor product quality (i.e., outside the Y quality specification limits). The square region "B" corresponds to the univariate specification region when accepting the same variance on each individual property as the multivariate region. In this case, accepting raw material batches with properties outside region "A" and inside region "B" leads to manufacturing products with final product quality Y outside its specification limits. To avoid this, companies are forced to shrink the univariate specifications from region "B" to the region "C", at the cost of rejecting acceptable raw material batches (i.e., those outside region "C" but inside region "A"). Another consequence of setting these more



A. True multivariate region.

B. Univariate specification region containing the true multivariate region.

C. Univariate specification region containing only acceptable raw material.

Figure 5.1: Problem of using univariate specifications on correlated raw material properties (Z_1 and Z_2).

restrictive univariate specifications is an increase in costs in the acquisition of raw material batches with tighter variations in their properties.

Multivariate specifications provide, therefore, much insight into what constitutes acceptable raw material batches when their properties are correlated (as usually happens). In order to cope with this correlation several authors suggest using multivariate approaches, such as Partial Least Squares (PLS) regression, to improve the definition of raw materials specifications.

The first systematic study was reported by De Smet [79], where PLS regression is used first to build a model between raw materials properties and CQAs by using historical data. Then, a boundary in the model subspace is defined within which most of the values for the raw materials properties associated with good CQAs can be found. This multivariate region (in the latent space) can then be used to accept or reject new batches of raw materials. The key assumption of this method is that variability in the CQAs results exclusively from variations in the raw materials properties of a single material. Duchesne and MacGregor [78] generalized this method by assuming that both variation in raw materials properties and in process operating conditions are responsible for CQAs variations. Uncontrolled variability in the operating conditions will increase the variability of the CQAs and require tightening specifications on the

raw material properties to make up for it. On the other hand, properly tuned feedback and feedforward controllers may compensate for CQAs variations allowing for wider raw material properties specifications [80]. Later on, García-Muñoz [81] extended the Duchesne-MacGregor method to combine data from multiple scales (e.g. lab or pilot scale and commercial scale) with different processing conditions and control strategies.

These approaches, however, focused on defining multivariate specification regions on the multiple properties of a single raw material. To overcome this limitation, MacGregor et al. [82] extended them to determine the acceptability of new raw materials from multiple suppliers and with multiple measured properties, as well as to assess the suitability of combining specific batches of raw materials currently in inventory to minimize the risk of manufacturing a poor quality product. Finally, Azari et al. [42] proposed a sequential multiblock PLS algorithm to better sort the contribution of raw materials and process operating conditions on CQA variations, considering two types of raw materials.

In the aforementioned references, the aim was to determine the boundary in the latent space of the historical data that best separates acceptable from unacceptable raw materials by direct mapping (i.e., those leading to good and poor CQAs, respectively). Nonetheless, the general shape (e.g., an ellipsoid or a straight line) and locus of such boundary was decided based on subjective criteria, trying to best balance out the type I and type II risks[1]. In contrast to this, García-Muñoz, Dolph, and Ward [80] emphasized the use of mathematical and statistical models as an objective way to define such specifications by linking them with specification limits for CQAs. Thus, given a desired set of CQAs, and in order to predict an appropriate set of raw materials properties, it is necessary to carry out the inversion of the model relating inputs (raw materials properties) with outputs (CQAs). Recently, Paris, Duchesne and Poulin [83] carried out a comparison between direct mapping and model inversion stating their advantages and drawbacks.

However, when inverting PLS models, their prediction uncertainty is also backpropagated [40, 84]. This issue has not been addressed in the past when defining multivariate raw materials specifications and, thereby, all the methods commented above are considered as descriptive approaches focused on historical data, lacking a probabilistic interpretation. For that, uncertainty is accounted

[1]Type I risk is defined as the proportion of truly acceptable batches of raw materials that is rejected by the customer under a given specification region; type II risk consists of the proportion of truly unacceptable batches of raw materials that is accepted by the customer under a given specification region [78].

in the form of prediction intervals, with a certain confidence level, finding a window within which any batch with raw material properties is expected to produce product with CQAs within specification limits with at least the predefined confidence level. In this regard, this window refers to the estimation of the so-called Raw Material Design Space (DS). The DS is defined as the multidimensional combination and interaction of input variables (e.g., raw material properties) that have been demonstrated to provide assurance of quality [10].

Bayesian approaches [85–87] can be used to include the model-parameter uncertainty and estimate the probability map of meeting the specifications imposed on the CQAs being used to identify the DS [88]. However, these methodologies define the DS by means of a predictive (forward) approach instead of carrying out the model inversion (backward). Therefore, the representation of the DS a priori requires the discretization of the multidimensional input domain by sampling algorithms. Then, simulation methods, such as Markov-Chain Monte Carlo techniques, are required for each discretization point to determine if it is within the DS. Hence, these approaches do not represent analytically the DS in the input domain, with the additional drawback of being computationally costly.

The novelty of the methodology presented in this chapter is the implementation of the frequentist probabilistic interpretation in the definition of the Raw Material DS in the latent space. For that, we propose a method to define analytically a window in the latent space of the raw material properties that is expected to provide assurance of quality for the CQAs with at least a certain confidence level. Besides, it can be used with historical data (i.e., daily production data not coming from any experimental design but with varying raw material properties, typical from Industry 4.0 environment) since, when fitting PLS models, causality can be inferred in the latent space, which allows the meaningful inversion of the model as discussed in Part I.

The chapter is organized as follows. Data requirements for defining multivariate specification are first discussed in Section 5.2. How PLS inversion addresses the definition of multivariate specifications by considering a probabilistic approach is then presented (Section 5.3), followed by a description of the exploitation of those specifications. Finally, the methodology is illustrated by means of two industrial case studies (Section 5.5).

5.2 Data requirements

The data required for developing raw materials multivariate specifications following the methodology proposed in this chapter involves two blocks, $\mathbf{Z}$ and $\mathbf{Y}$. $\mathbf{Z}$ $(N \times M)$ is a matrix of inputs which includes a total of M measurements characterizing the properties of each of the N batches of a particular raw material. Finally, the $\mathbf{Y}$ $(N \times L)$ output matrix consists of L measurements of the CQAs of the final product obtained for each one of the N corresponding batches.

Furthermore, process conditions may be under tight control to attenuate some raw material variations, whenever the eventual effect of such variability on the CQAs can be compensated by control systems. Specifications for incoming raw materials are nonetheless required, however, to account for variations in raw materials whose effect on the CQAs cannot be compensated by control systems. Therefore, if this situation prevails in the future there is no need to consider process data to establish the specification regions associated to the latter source of variation.

5.3 Defining the design space in the latent space by means of PLS

Defining the multivariate raw material specification region in the latent space is equivalent to defining the multidimensional combination and interaction of raw materials properties that have been demonstrated to provide assurance of quality (i.e., the Raw Material DS). Hence, both terms (multivariate specifications region and DS) are used interchangeably in the remainder of the chapter.

5.3.1 Design space with no uncertainty

If there is no prediction uncertainty, the DS must be defined as a region in the latent space associated with raw materials properties such that these properties yield an expected value of CQAs, according to Equation 2.7, within their specification limits.

Besides, since PLS is an empirical model based on historical data, any new set of raw materials properties must respect the correlation structure and range of this historical data [35]. Regarding the correlation structure: since the DS is defined in the latent space, it ensures new observations to behave in the same

way as the ones used to create the model, in the sense that the correlation structure of the model is respected. Regarding the historical range: when considering the Hotelling T^2 confidence limit as a raw material specification limit, the new set of raw material properties are constrained to be within historical ranges by a multivariate approach. Additionally, historical univariate ranges for each property (and other constraints) might be included. In this study, we initially focus on the l-th CQA and, hence, vector $\boldsymbol{y}^{des}$ degenerates to scalar y^{des}, and matrix $\mathbf{Q}$ degenerates to vector $\boldsymbol{q}_l^{\mathrm{T}}$ (l-th row of matrix $\mathbf{Q}$). Besides, one might face three scenarios depending on the kind of specifications for it:

1. $y_l = y^{des}$. In this first scenario, a specific value of the l-th CQA is required.
2. $y_l^{LSL} \leq y_l \leq y_l^{USL}$. In the second scenario, it is desired that the l-th CQA is between a lower specification limit (y_l^{LSL}) and an upper specification limit (y_l^{USL}).
3. In the third scenario, only one specification limit is considered, which might be lower ($y_l^{LSL} \leq y_l$) (scenario 3i) or upper ($y_l \leq y_l^{USL}$) (scenario 3ii).

Following the same framework as in Figure 2.1, there are three raw material properties ($M = 3$) and the focus is on the l-th CQA, and a PLS model has been previously fitted using two components ($A = 2$). Figure 5.2 shows the DS in the latent space for the latter three scenarios assuming a PLS model with no uncertainty.

In the first scenario, the desired specific value for the l-th CQA yields a one-dimensional NS and, the DS is defined by the intersection of this NS and the Hotelling's T^2 confidence region. In the same way, in the second and third scenarios, each specification limit is defined in the latent space by its associated NS. Thus, the DS in the latent space is defined by the intersection of the scores fulfilling the specifications' NSs and the Hotelling T^2 confidence region.

Until now, the DS has been defined without taking into account the prediction uncertainty. However, since empirical models are subject to uncertainty, when a PLS model is inverted, the uncertainty is backpropagated to the calculated inputs (i.e., the DS calculation is probabilistic) [40, 84].

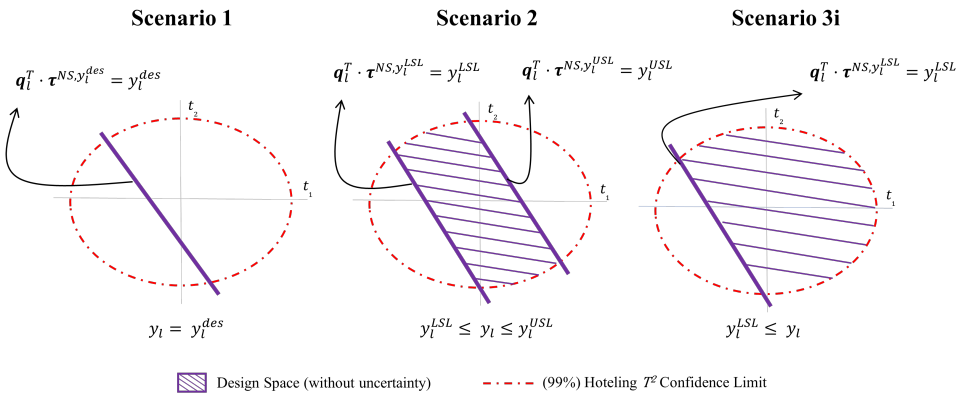


Figure 5.2: The Design Space in the latent space, for the three scenarios, assuming a PLS model with no uncertainty. NS: null space.

5.3.2 High confidence design space

5.3.2.1 Bracketing the design space

When prediction uncertainties are present, the DS without uncertainty shown in Figure 5.2 does not correspond to the true DS. Therefore, it might be possible to improve the estimation of the DS by running a set of experiments designed within the input domain that have already been used in the past (i.e., the so-called Knowledge Space (KS)). However, exploring the entire KS may be impractical due to the high number of experiments that may be needed to account for the variability in all accessible inputs [84]. For that reason, several approaches have already been proposed in order to define a subspace of the historical KS where the true DS is likely to lie with a predefined confidence level. This subspace is called the Experimental Space (ES).

In particular, Facco et al. [84] present a methodology to account for the back-propagation of the prediction uncertainty in model inversion to bracket the DS. This methodology resorts to the calculation of the prediction interval considering only the inversion solution by means of the pseudo-inverse (Equation 2.16). However, this approach does not consider the difference in the amplitude of the confidence region due to the leverage of different sets of scores along the NS. A proposed solution was given by Palací-López et al. [40] leading to non-linear confidence limits.

A graphical interpretation of the methodology proposed by Palací-López et al. [40] is shown in Figure 5.3 assuming the first scenario ($y_l = y^{des}$).

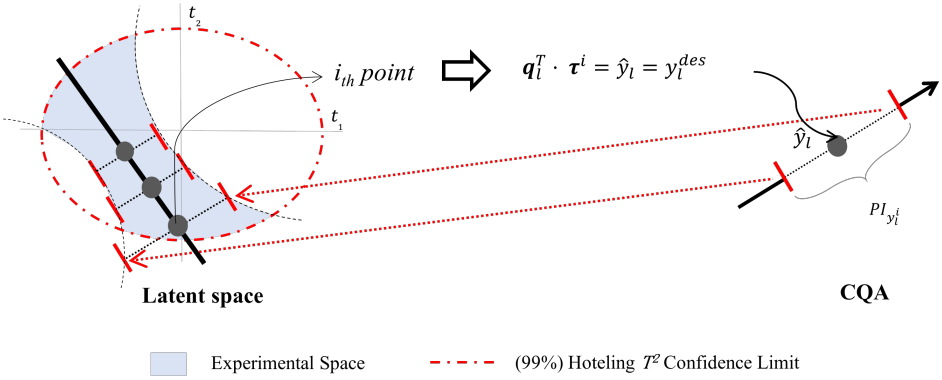


Figure 5.3: FIRST SCENARIO: The methodology proposed by Palací-López et al. [40].

Figure 5.3 shows that, as expected, moving along the NS one would obtain the same prediction of the l-th CQA. Nevertheless, due to model uncertainty, it does not guarantee to obtain exactly such a prediction. When considering the prediction uncertainty, a prediction interval which is expected to contain the true value of an individual value with a predefined confidence level can be calculated. Note that, since the prediction interval depends on the leverage of the observation (Equations 2.9 2.10 and 2.11), its amplitude is expected to be lower for observations close to the centre of projection (small leverage) than for those far away from it (high leverage) [40]. Then, the prediction intervals for the multiple solutions are backpropagated when the model is inverted. Thus, the KS is restricted in such a way as to identify an experimental space in the latent space, which has a high probability of containing the true DS. However, this does not mean high probability of providing assurance of quality, which is what we are interested in when defining multivariate specifications.

5.3.2.2 Proposed definition of the design space

The proposed methodology for defining multivariate raw material specifications is motivated by Facco et al. [84] and Palací-López et al. [40] ideas when back-propagating the uncertainty, but framing the knowledge space with a different purpose. The ES has a high probability of containing the true DS at the expense of including unacceptable raw material batches. By contrast, in this chapter we propose considering the prediction uncertainty in a different

way, when the model is inverted, in order to define a subspace of the KS where there is assurance of quality with a certain confidence level. For ease of understanding of the proposed methodology, we illustrate the second scenario (Figure 5.4) where it is desired that l-th CQA is between y_l^{LSL} and y_l^{USL}.

As discussed above, even though working in the NS associated with the specification limit leads to a predicted value between specifications, it might yield out of specifications values for the l-th CQA due to prediction uncertainties. For that reason, focusing on the y_l^{LSL}, one should accept raw materials properties such that its projection in the latent space leads to a lower endpoint, which is equal or higher than the y_l^{LSL}, thus delimiting a lower confidence region (Equation 5.1.

$$y_l^{LSL} \leq \boldsymbol{q}_l^{\mathrm{T}} \boldsymbol{\tau}^{new} - t_{N-df,\alpha/2} s_{e_l^{new}} \tag{5.1}$$

When calculating this confidence limit for the multiple solutions along the NS of y_l^{LSL}, a non-linear boundary is obtained for the y_l^{LSL} as is shown in Figure 5.4a. Such boundary in the latent space refers to the Lower Specification Confidence Limit (LSCL). If working in the LSCL there will be a high probability to obtain the l-th CQA higher than the y_l^{LSL}.

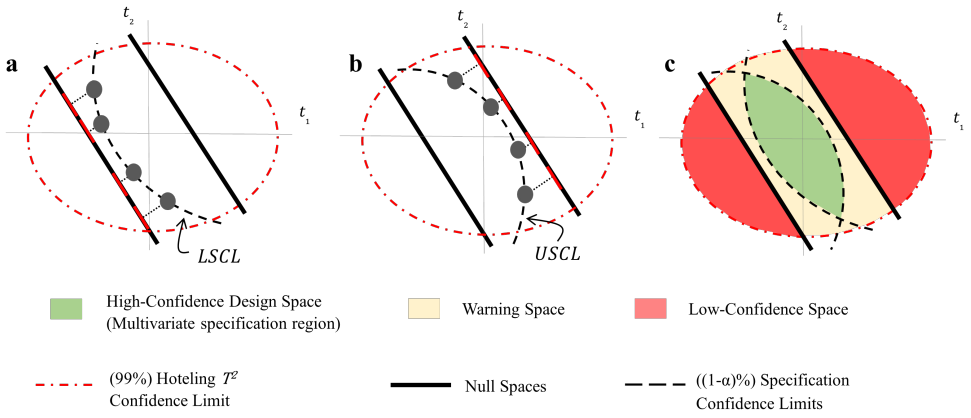


Figure 5.4: SECOND SCENARIO: Graphical interpretation of the proposed definition of the High-Confidence Design Space. (a) lower specification confidence limit (LSCL). (b) upper specification confidence limit (USCL). (c) Splitting the KS into High-Confidence Design Space, Warning Space and Low-Confidence Space.

In the same way, considering the y_l^{USL}, one should accept raw materials properties such that its projection in the latent space leads to an upper endpoint which is equal or lower than the y_l^{LSL}, thus delimiting an upper confidence region (Equation 5.2).

$$y_l^{USL} \geq \boldsymbol{q}_l^{\mathrm{T}} \boldsymbol{\tau}^{new} + t_{N-df,\alpha/2} s_{e_l^{new}} \tag{5.2}$$

Following an analogous reasoning as before, another non-linear boundary, called Upper Specification Confidence Limit (USCL), is obtained for the y_l^{USL} (see Figure 5.4b). If working in the USCL there will be a high probability to obtain the l-th CQA lower than the y_l^{USL}.

Appendix 5.A shows the analytical expression, which allows calculating the score belonging to both the lower and upper specification confidence limits given its respective score in the NS for the l-th CQA. Although Equations 5.1 and 5.2 refer to one-sided prediction intervals, the t-statistic is calculated at the $\alpha/2$ significance level because two specifications limits are considered. In the case of having one specification limit (i.e., third scenario), Equations 5.1 or 5.2, as appropriate, would be used at α significance level.

The intersection regions delimited by the LSCL, USCL and the Hotelling T^2 confidence ellipsoid, delimits the so-called High-Confidence Design Space, where any batch of raw material properties results in a prediction interval for the CQA within specifications. Therefore, from a frequentist probabilistic interpretation, these batches are expected to produce product with CQAs within specification limits with a confidence level equal or higher than $1 - \alpha$. In other words, this definition of the High-Confidence DS has been demonstrated to provide assurance of quality with at least a certain confidence level (Figure 5.4c). The High-Confidence DS is a potential opportunity to establish Real-Time Release (RTR), which is defined as the ability to evaluate and ensure the acceptable quality of the final product based on inputs variables (e.g., raw material properties) without using end-product testing [10].

Additionally, the intersection between the region bounded by the two NSs corresponding to the y_l^{LSL} and y_l^{USL}, and the Hotelling's T^2 confidence region, but outside the High-Confidence DS, defines the so-called Warning Space (Figure 5.4c). Note that, although this space does not belong to the multivariate raw material specification region as defined, it does not necessarily imply the rejection of batches. In fact, batches lying within the Warning Space lead to predicted values between specifications, but they result in prediction intervals for the CQA partially outside of specifications given the predefined confidence

level $1 - \alpha$. Namely, there is no assurance of quality due to the prediction uncertainty and, hence, RTR testing is not feasible. Instead of that, end-product testing may be employed, which usually involves undertaking specific lab-testing procedures on samples of the final product. This could be interesting when rejecting all batches in the Warning Space is not affordable. Finally, the Low-Confidence Space (Figure 5.4c) leads to predicted values outside specifications. Although batches lying within this subspace may lead to response values between specifications, most of the time such values are expected to be outside.

Therefore, following the proposed approach, the KS is split into three regions: High-Confidence DS, Warning Space and Low-Confidence Space, providing a strategy where RTR or end-product testing, can be used as needed.

Note that, the High-Confidence DS is more restrictive that the unknown true DS, and the less uncertainty there is, the more similar the High-Confidence DS and the true DS are, as it is graphically shown in Figure 5.5.

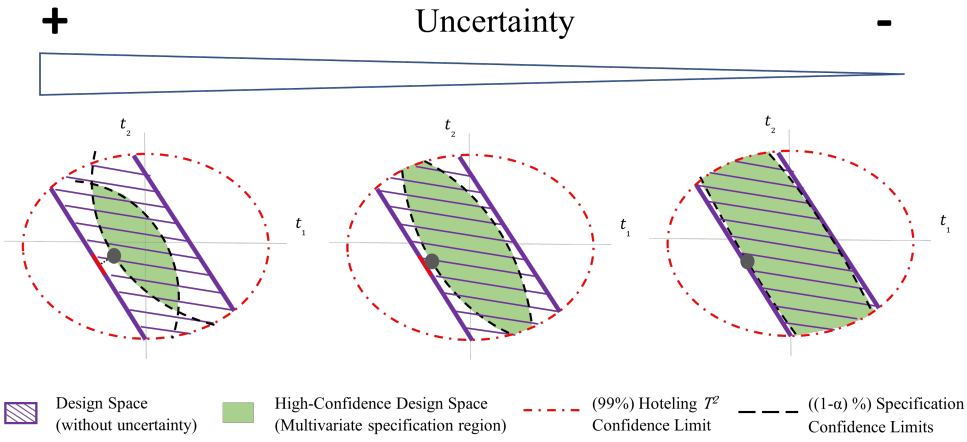


Figure 5.5: SECOND SCENARIO: Effect of the uncertainty on the High-Confidence DS related to the DS without uncertainty.

High uncertainty in the data is reflected in a low goodness of prediction model. But this does not limit the proposed methodology, indeed, the lower goodness of prediction, the more crucial it is to take uncertainties into account if product quality is to be guaranteed. In that point, the authors would like to challenge the widely held view that a low goodness of prediction model is useless and point out that low goodness of prediction model, typical from the industry

4.0 environment, can be useful if being cautious. In this sense, García-Muñoz and Mercado [89] already worked in a real process under control where a LV regression model, that had the ability to systematically predict 21% of the variability in the quality attribute, was used with a great potential for improvement. However, in certain situations a low goodness of prediction model may be a warning of non-linearities in the original dataset that is not captured adequately by the linear PLS model [86].

To summarize, Figure 5.6 shows the DS (if there is not uncertainty in the model), the experimental space and our proposed High-Confidence DS for all scenarios.

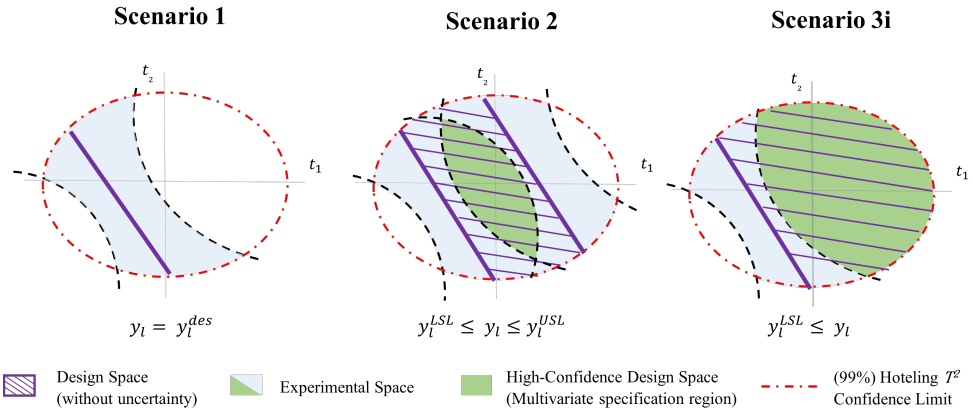


Figure 5.6: Comparison of the DS (without uncertainty), ES and High-Confidence DS for the three scenarios.

The first scenario is a particular case of the second scenario where $y_l^{LSL} = y_l^{USL}$. In this case, there is no intersection between the LSCL and USCL and, therefore, the High-Confidence DS does not exist. Up to this point, we have defined the High-Confidence DS for the l-th CQA. The joint High-Confidence DS for the L CQAs will be obtained as the intersection of the L High-Confidence DSs for each CQA.

5.4 Exploiting the model

Once the High-Confidence DS have been defined as discussed above, the model can be used to inspect every new batch of raw material, $\boldsymbol{z}^{obs}$ (Phase II). This allows the user to predict if the CQAs of the product, that would be manufactured using any new raw material batch, would be within specifications, and consequently accept or reject the raw material batch prior to introducing it into the production process. The procedure for that is as follows:

1. Mean-center and scale $\boldsymbol{z}^{obs}$ using the same mean and scaling factor used on the calibration data when the PLS model was developed in Phase I.

2. The scores $\boldsymbol{\tau}^{obs}$ are obtained from the linear combinations of mean-centered and scaled raw materials properties according to Equation2.4, and the $SPE_{\boldsymbol{z}^{obs}}$ is obtained according to Equation 2.6.

3. The final decision on whether to accept or reject a new raw material batch is up to the user based on the values of $SPE_{\boldsymbol{z}^{obs}}$ and $\boldsymbol{z}^{obs}$. When the $SPE_{\boldsymbol{z}^{obs}}$ is higher than SPE_{lim}, this suggests that their properties reflect a different correlation structure than that of the raw material batches from the historical dataset used to build the PLS model. It is then impossible to predict with the fitted PLS model the impact of this raw material batch on CQAs of the final product. Besides, in such a case, one could use the SPE contribution plot in order to examine which raw material properties contribute the most to this high SPE value, providing the supplier with useful information about deviations in the batch raw material properties. Regarding the projection in the latent space $\boldsymbol{z}^{obs}$, if these scores fall within the High-Confidence DS, this batch will be expected to produce product with CQAs within specification limits with at least a certain confidence level. Note that, instead of rejecting all the high $SPE_{\boldsymbol{z}^{obs}}$ and high T^2_{obs} raw material batches, one may also process some of them (when deviations are not too important), incorporate them as new design points to augment the historical data matrices $\mathbf{Z}$ and $\mathbf{Y}$, and fit a new PLS model in order to better define sequentially the multivariate specification region.

5.5 Industrial case studies

5.5.1 First industrial case study: cereal extraction process

Description of the dataset

Historical data collected from a maize cereal extraction process is used to illustrate the proposed methodology. The maize is fed to the production process where, initially, it is cleaned to free the maize of all kinds of impurities and then it is steeped. Subsequently, a grinding process takes place to grind the harder parts of the maize, followed by a degerminating process so that the germ is separated from the fiber, gluten, and starch. Finally, after a sieving process is carried out to separate the fiber, a primary separator splits by centrifugal force the stream in two fractions: gluten and slurry starch. The latter has a great interest as it has become a major industrial raw material.

The data available in this case are a compilation of eight raw material properties ($\mathbf{Z}$) of maize: promatest value, protein, acid value, specific weight, burnt grain, broken grain, starch and extractable lipids, and one response variable y (extraction yield of starch slurry). These variables are easily registered in order to assess the feasibility of a raw material batch. In total, 989 historical batches/observations were measured: $\mathbf{Z}$ (989×8) and $\boldsymbol{y}$ (989×1). Besides, a lower specification limit of 69% is considered for the response variable, hence this case refers to the third scenario.

Performance of the multivariate raw material specifications

Leave-one-out Cross-Validation (CV) was used for selecting the number of PLS components. Thus, two LVs were chosen to fit a PLS model ($R^2_{\mathbf{Z}_{cum}} = 37.6\%$, $R^2_{\mathbf{Y}_{cum}} = 26.73\%$ and $Q^2_{\mathbf{Y}_{cum}} = 25.63\%$) using the 989 calibration observations. The R^2 values (goodness of fit) give the percentage of the total sum of squares of $\boldsymbol{y}$ and $\mathbf{Z}$, respectively, that are explained by the fitted PLS model, while the $Q^2_{\mathbf{Y}_{cum}}$ (goodness of prediction) gives the percentage of the total sum of squares of the response that can be predicted with the PLS model by CV. It is also crucial to validate the model by monitoring charts for SPE and T^2 (shown in Figure 5.7), in order to determine whether historical/happenstance data are consistent with normal process conditions (i.e., common cause process variations).

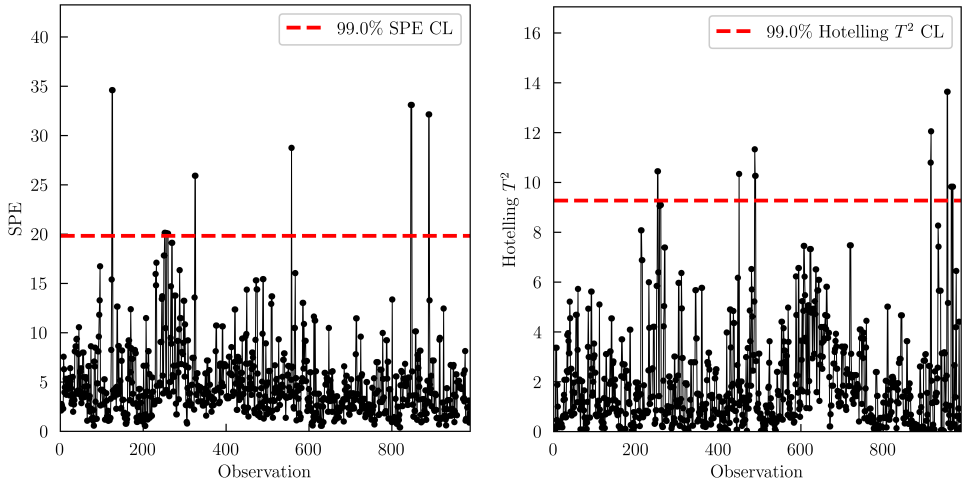


Figure 5.7: First Case Study: Monitoring charts for SPE (left) and T^2 (right).

Figure 5.7 shows that none of the historical batches exhibit any unusual behavior caused by special cause process variations. Notice that, although some of them slightly exceed the upper confidence limit, they correspond approximately to 1% false alarm rate (expected when using 99% confidence limits).

Figure 5.8 illustrates the 99% Hotelling T^2 confidence limit, the NS associated with the LSL, and its 90% confidence limit when considering the prediction uncertainty (i.e., the Low Specification Confidence Limit, LSCL). The intersection of all confidence regions, defined by their limits, yields the High-Confidence DS (i.e., the proposed multivariate raw material specifications in the latent space) within which there is assurance of obtaining superior or equal yields to 69% with at least 90% confidence level.

To evaluate the performance of the definition of the multivariate raw material specification region, a diagnostic test is carried out. In particular, type I risk, type II risk and the Negative Predictive Value (NPV) are calculated for the High-Confidence DS. The NPV is the proportion of batches that actually result in a good product out of all those within the High Confidence Design Space, and, hence, this metric is directly connected to the definition of the High-Confidence DS itself. The assessment of these metrics is carried out by leave-one-out CV.

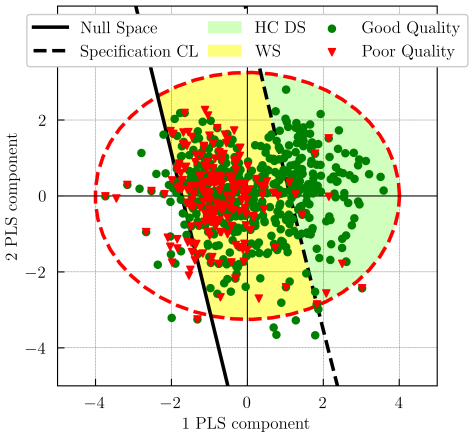


Figure 5.8: First Case Study: Graphical definition of the High Confidence Design Space, Warning Space and Low-Confidence Space by showing calibration data.

De Smet [79] and Duchesne and MacGregor [78] approaches would have ended up defining a straight line or an ellipsoid in a subjective way, that would best balance type I and type II risks in Figure 5.8. Besides, if the PLS model was of a higher dimension ($A \geq 3$), it would be difficult to decide the general shape and locus that best defines the separation between good and poor quality, unlike the proposed approach, which does not suffer from such handicap regardless of the dimensionality of the latent space. On the other hand, García-Muñoz, Dolph, and Ward [80] would have obtained a wider region, akin to the DS without considering the uncertainty (the joint of High-Confidence DS and Warning Space). However, because of the uncertainty, this approach would result in accepting almost every batch of raw materials (no matter if they are acceptable or unacceptable), leading to 6.10% type I risk, 88.89% type II risk, and 77.42% NPV. None of these approaches are probabilistic, and therefore they do not allow knowing the confidence level in meeting the final product quality specifications.

By contrast, our High-Confidence DS is defined with at least a 90% confidence level of obtaining superior or equal yields to 69%. Thus, one would expect that, of the batches lying within the High-Confidence DS, 90% or more would be acceptable batches (the NPV for the High-Confidence DS is 96.10%). On the other hand, the High-Confidence DS leads to 70.60% type I risk and 3.86% type II risk. This means that if only batches lying within the High-Confidence DS are accepted, 3.86% of unacceptable batches of raw materials will be accepted

at the expense of rejecting 70.60% of acceptable batches. These results are the consequence of the low PLS goodness of prediction ($Q^2_{\mathbf{Y}_{cum}} = 25.63\%$) in this case study, due to the fact that historical data presents a low signal to noise ratio. Alternatively, one could accept batches lying within the Warning Space knowingly that the NPV in such space would be 71.14% and, hence, likely end-product test should be required. Another option would be to balance the type I and type II risks by modifying the confidence level of the High-Confidence DS. Figure 5.9 shows the High-Confidence DS for different confidence levels (50, 70, 90 and 99%). The corresponding type I risk, type II risk and NPV for the High-Confidence DS, and NPV for the Warning Space are shown in Figure 5.10. Note that, in this case (i.e., scenario 3) the 50% confidence level case corresponds to the DS without considering the uncertainty.

Figure 5.9 shows that as confidence level increases, a tighter High-Confidence DS is spanned, thereby, the type II risk is reduced at the expense of increasing the type I risk, as is shown in Figure 5.10. Therefore, the confidence level of the High-Confidence DS must be chosen according to the users by balancing the consequences of having type I and type II errors in their processes and the total amount of such errors. Besides, for all cases, the NPV is equal or higher than its corresponding confidence level as expected.

Influence of the goodness of predictions

In order to investigate how PLS goodness of prediction $Q^2_{\mathbf{Y}_{cum}}$ affects the performance of the High-Confidence DS a simulation study is carried out. In these simulations, we assume that the true model relating $\mathbf{Z}$ and $\boldsymbol{y}$ is, indeed, the one calculated by the calibration set. Hence, individual values of $\boldsymbol{y}$, y^{obs}, are obtained using Eq. (21) given a batch of raw material $\boldsymbol{z}^{obs}$ and the weighting matrices $\boldsymbol{q}^2$ and $\mathbf{W}^*$:

$$y^{obs} = \boldsymbol{q}^{\mathrm{T}}\mathbf{W}^{*\mathrm{T}}\boldsymbol{z}^{obs} + e^{obs} \tag{5.3}$$

where e^{obs} is an independent random noise value from a normal distribution with zero mean and standard deviation σ. By modifying the value of such standard deviation, one can create simulated datasets yielding PLS models with different goodness of prediction. Figure 5.11 shows the High-Confidence DS with 90% confidence level of obtaining superior or equal yields to 69% for different datasets simulated from the exploiting dataset by using a standard

[2]Note that since there is only one CQA, $L = 1$ and $\mathbf{Q} = \boldsymbol{q}^{\mathrm{T}}$.

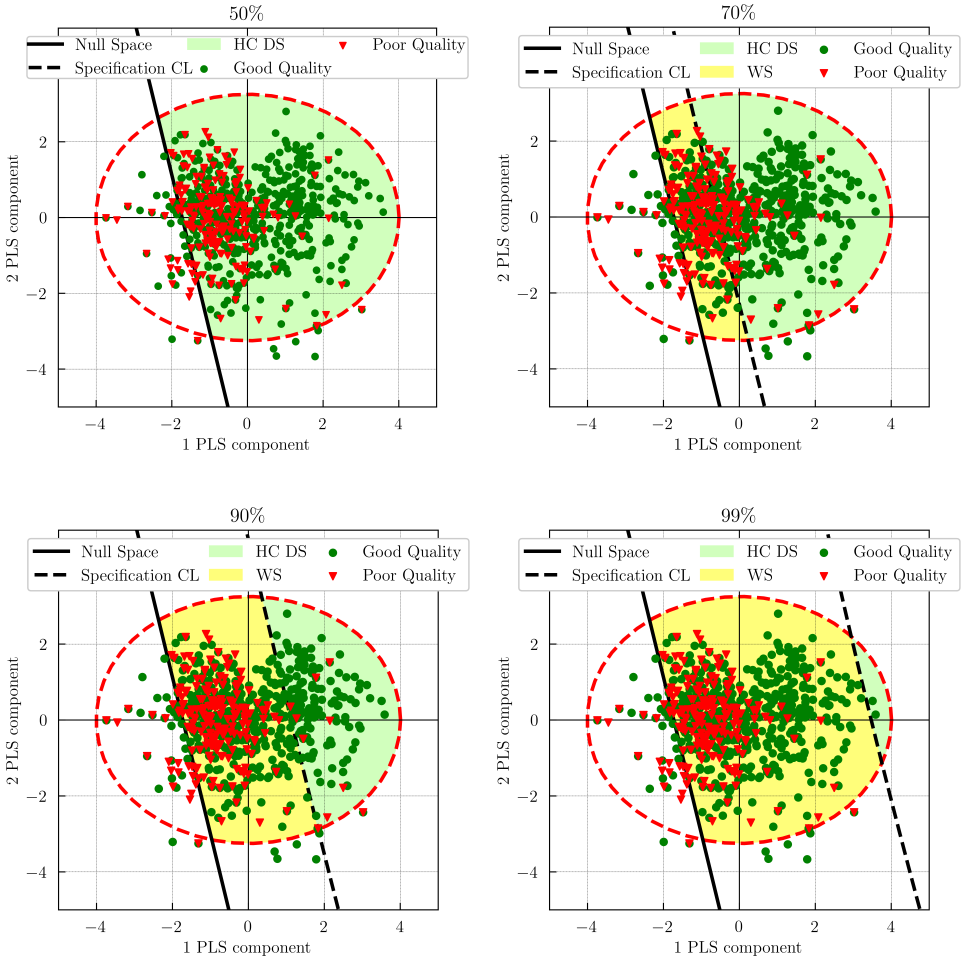


Figure 5.9: First Case Study: High-Confidence DS, Warning Space and Low-Confidence Space for several confidence levels.

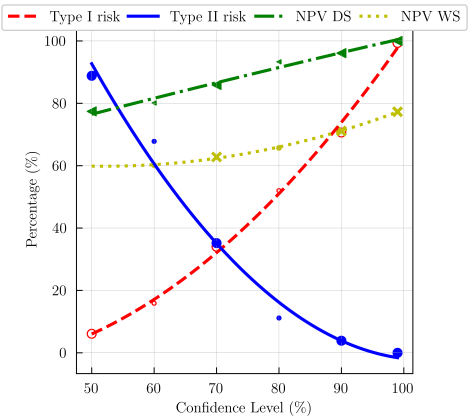


Figure 5.10: First Case Study: type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) vs confidence level: **50**%, 60%, **70**%, 80%, **90**% and **99**%. Bold values refer to those used in Figure 5.9 and are shown bigger.

deviation of 0.025, 0.1, 0.5 and 1 yielding $Q^2_{\mathbf{Y}_{cum}}$ of 91.76%, 73.92%, 41.06% and 20.19%, respectively. 989 batches have been simulated for each dataset to obtain more accurate results with respect to the original data.

Figure 5.11 shows that the lower the noise standard deviation, the higher the goodness of prediction and, consequently, the clearer the discrimination between acceptable and unacceptable raw materials. Besides, regardless the goodness of prediction, the proposed method defines the multivariate specification region given the same confidence level (90%). As can be seen, lower values for the goodness of prediction result in narrower multivariate specification region where more acceptable material is rejected to guarantee such confidence level. This will affect the type I and type II risks as shown in Figure 5.12.

Figure 5.12 shows that with moderate/high values of $Q^2_{\mathbf{Y}_{cum}}$ it is feasible to obtain DS with high confidence level and low type I and II risks, and high NPV. For example, given desired yields equal or superior to 69%, the DS with 90% confidence level and $\sigma = 0.025$ ($Q^2_{\mathbf{Y}_{cum}} = 91.76\%$) leads to 9.23% type I risk, 3.57% type II risk and 99.50% NPV. However, with low values of $Q^2_{\mathbf{Y}_{cum}}$ it is more critical to consider the prediction uncertainty for guarantying quality (i.e., high NPV in the High-Confidence DS) at the expense of increasing the type I risk.

Note that the apparently bad performance for low values of $Q^2_{\mathbf{Y}_{cum}}$ is solely due to the nature of the data and not the methodology, as noise refers to

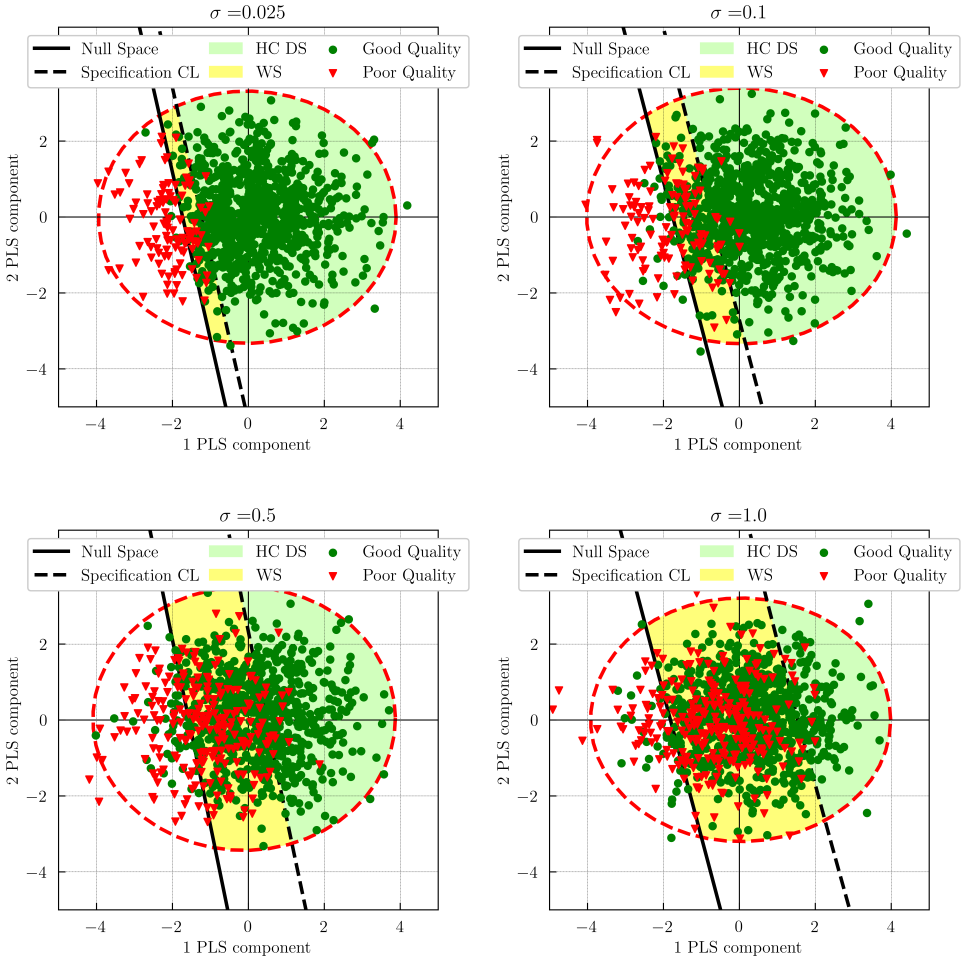


Figure 5.11: First Case Study: High-Confidence DS, Warning Space and Low-Confidence Space for simulated data with different noise variability σ.

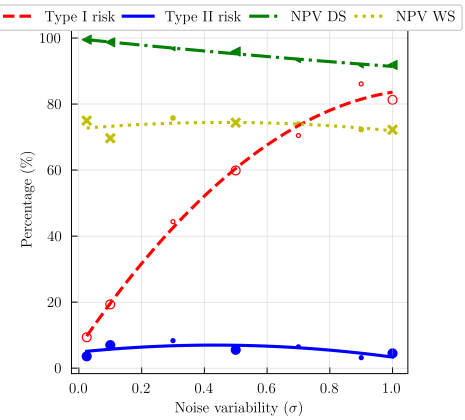


Figure 5.12: First Case Study: type I risk, type II risk and NPV for the High-Confidence DS, and NPV for the Warning Space (WS) with 90% confidence level vs noise variability: **0.025**, **0.1**, 0.3, **0.5**, 0.7, 0.9 and **1**. Bold values refer to those used in Figure 5.11 and are shown bigger.

random variation with no pattern, and therefore usually unavoidable and unpredictable.

In the case of desiring to increase the signal-to-noise ratio of the data sets, some process excitation is needed. Multivariate design of experiments can be used such that it provides the greatest amount of additional information with respect to the information available in the existing dataset [90] Considering these new observations from experimentation in addition to the historical/happenstance data will improve the estimation of the High-Confidence DS (i.e., wide multivariate specification region with high confidence level and low type I and type II risks will be obtained).

Sensitivity analysis of the number of PLS components

A sensitivity analysis was undertaken to assess the stability of the High-Confidence DS with respect to the number of PLS components. The number of components to be used is a very important property of a PLS model and their choice must be done according to the purpose of such model. In our case study, we have evaluated how changes in the number of components may affect the type I and type II risks of the High-Confidence DS with 90% confidence limit. Table 5.1 shows that no relevant differences in the performance of the diagnostic test are observed when adding PLS components. The reason for

this is the fact that the goodness of prediction ($Q^2_{\mathbf{Y}_{cum}}$) is quite similar among the models.

Table 5.1: First Case Study: Goodness of prediction ($Q^2_{\mathbf{Y}_{cum}}$), type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) as a function of the number of PLS components (High-Confidence DS for 90% confidence level).

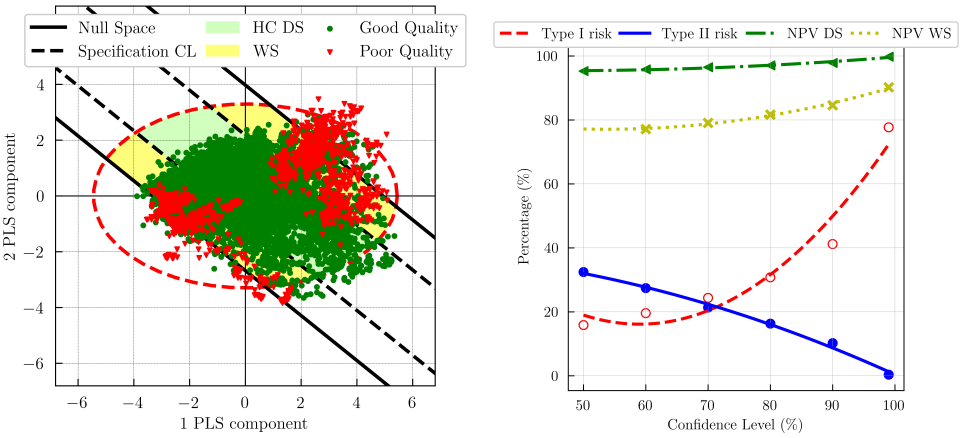
A	$Q^2_{\mathbf{Y}_{cum}}$ (%)	Type I (%)	Type II (%)	NPV HC DS (%)	NPV WS (%)
1	25.06	72.88	3.86	95.79	71.67
2	25.67	70.63	3.86	96.10	71.14
3	25.58	70.90	3.43	96.49	71.45
4	25.57	70.90	3.43	96.49	71.51
5	25.56	70.90	3.00	96.92	71.32
6	25.56	70.90	2.58	97.35	70.98
7	25.56	71.03	2.58	97.33	70.96
8	25.56	71.03	2.58	97.33	70.68

5.5.2 Second industrial case study: blown film process

Description of the dataset

This industrial case study refers to a catalytic afterburner used as control device for oxidation of undesirable combustible gases in a petrochemical process. The properties of the catalyst have an impact on the afterburn quality process and, hence, it is not only crucial to determine the raw material properties of the catalyst, but also to define its multivariate specifications for ensuring such quality.

The historical/happenstance data available are a compilation of nine properties of the afterburn catalyst ($\mathbf{Z}$) related to regenerated catalyst percentage, catalyst density, particle size distribution and chemical composition, and one response variable $\boldsymbol{y}$ (afterburn yield). In total, 9971 historical batches/observations were measured. Besides, both lower and upper specification limits are considered for the response variable, hence this case refers to the second scenario.



(a) Graphical definition of the High-Confidence DS, Warning Space (WS) and Low-Confidence by showing calibration data.

(b) Type I risk, type II risk and NPV for the High-Confidence DS, and NPV for the Warning Space vs confidence level

Figure 5.13: Second Case Study.

Performance of the multivariate raw material specifications

Leave-one-out CV [62, 91] was used for selecting the number of PLS components. Thus, two LVs were chosen to fit a PLS model ($R^2_{\mathbf{Z}_{cum}} = 56.44\%$, $R^2_{\mathbf{Y}_{cum}} = 73.32\%$ and $Q^2_{\mathbf{Y}_{cum}} = 73.29\%$) using calibration observations. This is a case study with a moderate goodness of prediction ($Q^2_{\mathbf{Y}_{cum}}$). None of the historical observations exhibit any unusual behavior caused by special cause process variations based on SPE and T^2 charts (charts not shown).

Figure 5.13a illustrates the High-Confidence DS with a 90% confidence level resulting in 41.20% type I risk, 10.23% type II risk and 97.82% NPV. However, if uncertainty had not been considered, 4.47% type I risk, 62.34% type II risk and 92.27% NPV would have been obtained. As expected, Figure 5.13b shows that as confidence level increases, the type II risk is reduced at the expense of increasing the type I risk. It should be noticed that the type I and II risks and NPV not only depend on the goodness of prediction but also on other factors such as the scenario, the value of the specification limits or the tested data. For that reason, different case studies with the same $Q^2_{\mathbf{Y}_{cum}}$ could result in slightly different type I and II risks for the same confidence level.

Sensitivity analysis of the number of PLS components

Since in the second case study there is a substantial variation in the goodness of prediction when adding the second PLS component, the sensibility analysis of the number of the PLS components is also undertaken (Table 5.2).

Table 5.2: Second Case Study: Goodness of prediction ($Q^2_{\mathbf{Y}_{cum}}$), type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) as a function of the number of PLS components (High-Confidence DS for 90% confidence level).

A	$Q^2_{\mathbf{Y}_{cum}}$ (%)	Type I (%)	Type II (%)	NPV HC DS (%)	NPV WS (%)
1	46.82	66.04	13.84	95.03	85.88
2	73.29	41.20	10.23	97.82	84.60
3	73.97	40.67	8.11	98.28	84.64
4	74.24	41.86	7.50	98.37	84.02
5	74.30	41.99	7.76	98.31	84.05
6	74.37	41.37	7.41	98.40	83.86
7	74.47	42.03	7.32	98.41	84.18
8	74.62	41.52	7.94	98.29	84.20

Unlike the first case study (Table 5.1), Table 5.2 shows relevant improvements in the reduction of type I and II risks when adding the second PLS component, but not after adding more components. For that reason, it is concluded that the CV criterion for the selection of two PLS components results in good performance indices.

5.6 Conclusion

In this chapter, we propose a novel approach to define an analytical expression for defining the multivariate raw material specification region in the latent space where there is assurance of quality with a certain confidence level for the CQAs of the final product (i.e., the so-called High-Confidence design space). Thus, it would allow evaluating the capability of the raw material batches of producing product with CQAs within specification limits, before producing a single unit of the product, and based on that information, making a decision about accepting or not the supplier raw material batch. This is totally different from existing approaches that evaluate (and also accept or reject) raw material batches based on their raw material properties but not on the desired final product properties.

This methodology is based on the inversion of the PLS model, and the most remarkable advantages are:

- It can be used with historical data (i.e., daily production data not coming from any experimental design but with varying raw material properties, typical from Industry 4.0 environment) since, when fitting PLS models, causality can be inferred in the latent space, which allows the meaningful inversion of the model.
- It considers a multivariate approach providing much insight into what constitutes acceptable raw material batches when their properties are correlated.
- The use of mathematical and statistical models as a way to define such raw material specifications by linking them with specification limits for CQAs of the final product.
- It allows a frequentist probabilistic interpretation. The multivariate raw material region is expected to produce product with CQAs within specification limits with a confidence level equal or higher than $(1 - \alpha) \times 100$.
- It provides the analytical definition of the limits of the multivariate raw material specifications.
- It provides a strategy where RTR (for batches in the multivariate raw material specification region or High-Confidence Design Space), or end-product testing (for batches in the Warning Space) can be used as needed.

Note that, while driven by the need to define meaningful specifications for raw materials, the developed methodology can be applied not only to defining High-Confidence DS for raw materials but also for other input variables, including process variables.

Appendices

5.A Specification confidence limits for the l-th critical quality attributes

Let $\boldsymbol{\tau}^{NS}$ be a vector of scores belonging to the NS associated to either the upper or lower specification limit for the l-th CQA (y_l^{SL}), and $\boldsymbol{\tau}^{SCL}$ the vector of scores belonging to such specification confidence limit. Thus, the vector defined by $(\boldsymbol{\tau}^{NS} - \boldsymbol{\tau}^{SCL})$ is orthogonal to the NS (i.e., as vector $\boldsymbol{v}_l$ defining the hyperplane of the NS (Equation 2.17)), and the direction depends on whether it refers to y_l^{LSL} ($\boldsymbol{\tau}^{LSCL}$) or y_l^{USL} ($\boldsymbol{\tau}^{USCL}$):

$$\boldsymbol{\tau}^{NS} - \boldsymbol{\tau}^{SCL} = \boldsymbol{v}_l \lambda \tag{5.A.1}$$

where λ is a scalar that can be negative or positive depending on it referring to the $\boldsymbol{\tau}^{LSCL}$ or $\boldsymbol{\tau}^{USCL}$, respectively. Besides, the lower (if y_l^{LSL} is considered) or upper (if y_l^{USL} is considered) endpoint of its prediction interval must match the specification limit.

$$y_l^{SL} = \boldsymbol{q}_l^{\mathrm{T}} \boldsymbol{\tau}^{NS} \tag{5.A.2}$$

$$y_l^{SL} = \boldsymbol{q}_l^{\mathrm{T}} - t_{N-df,\alpha/2} s_{e_l^{LSCL}} \tag{5.A.3}$$

$$y_l^{SL} = \boldsymbol{q}_l^{\mathrm{T}} + t_{N-df,\alpha/2} s_{e_l^{USCL}} \tag{5.A.4}$$

By substitution and reorganization of either Equations 5.A.1, 5.A.2 and 5.A.3, or Equations 5.A.1, 5.A.2 and 5.A.3 the same quadratic equation is defined (Equation 5.A.5).

$$s_{e_l^{SCL}}^2 t_{N-df,\alpha/2}^2 = \left(\boldsymbol{q}_l \boldsymbol{v}_l\right)^2 \lambda^2 \tag{5.A.5}$$

Notice that there will be a negative solution attributed to the y_l^{LSL} and a positive solution attributed to the y_l^{USL}. Furthermore, since $s_{e_l^{USCL}}^2$ depends on the leverage of the unknown $\boldsymbol{\tau}^{SCL}$ (either $\boldsymbol{\tau}^{LSCL}$ or $\boldsymbol{\tau}^{USCL}$) according to Equations 2.10 and 2.11, it must can be expressed as a function of $\boldsymbol{\tau}^{NS}$

by taking into account Equation 5.A.1 as follows:

$$s^2_{e_l^{SCL}} = SE_l^2\big(\boldsymbol{v}_l^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{v}_l\lambda^2 - 2\boldsymbol{v}_l^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{NS}\lambda + 1 + 1/N\boldsymbol{\tau}^{NS^{\mathrm{T}}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{NS}\big)$$

Substituting Equation 5.A.6 in Equation 5.A.5:

$$\left(\boldsymbol{q}_l\boldsymbol{v}_l\right)^2\lambda^2 = SE_l^2\big(\boldsymbol{v}_l^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{v}_l\lambda^2 - 2\boldsymbol{v}_l^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{NS}\lambda + 1 + 1/N\boldsymbol{\tau}^{NS^{\mathrm{T}}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{NS}\big)t^2_{N-df,\alpha/2}$$

and reorganizing terms:

$$a\lambda^2 + b\lambda + c = 0 \tag{5.A.6}$$

where:

$$\begin{aligned} a &= SE_l^2\boldsymbol{v}_l^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{v}_l t^2_{N-df,\alpha/2} - \left(\boldsymbol{q}_l\boldsymbol{v}_l\right)^2 \\ b &= -SE_l^2 2\boldsymbol{v}_l^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{NS}t^2_{N-df,\alpha/2} \\ c &= SE_l^2\big(1 + 1/N\boldsymbol{\tau}^{NS^{\mathrm{T}}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\boldsymbol{\tau}^{NS}\big)t^2_{N-df,\alpha/2} \end{aligned} \tag{5.A.7}$$

The values of λ that satisfy the Equation 5.A.6 are the solutions of a quadratic equation and, as commented above, there will be a positive and a negative one. Besides, it is known that c is positive given the terms that define it. For all this, it can be deduced that the quadratic function is concave down (i.e., the second derivative is negative) and, consequently, a must be negative. Because a is negative and c is positive, it is determined that the discriminant $(b^2 - 4ac)$ is positive and, therefore, there are two distinct roots as follows:

$$\begin{aligned} \lambda_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \\ \lambda_2 &= \frac{-b - \sqrt{b^2 - 4ac}}{2a} \end{aligned} \tag{5.A.8}$$

where both of them are, by definition, real numbers. Since the root of the discriminant is higher than b and a is negative, it is deduced that λ_1 is negative (it refers to y_l^{LSL}) and λ_2 is positive (it refers to y_l^{USL}). Thus, Equation 5.A.9

shows the analytical expression of the specification confidence limits when considering the prediction uncertainty.

$$\begin{aligned} \boldsymbol{\tau}^{LSCL} &= \boldsymbol{\tau}^{NS} - \boldsymbol{v}_l \lambda_1 \\ \boldsymbol{\tau}^{USCL} &= \boldsymbol{\tau}^{NS} - \boldsymbol{v}_l \lambda_2 \end{aligned} \tag{5.A.9}$$

Chapter 6

Defining multivariate raw material specifications via SMB-PLS

Part of the content of this chapter has been included in:

[17] J. Borràs-Ferrís, C. Duchesne, and A. Ferrer, "Defining Multivariate Raw Material Specifications via SMB-PLS," *Chemometrics and Intelligent Laboratory Systems*, vol. 240, 2023. DOI: 10.1016/j.chemolab.2023.104912

6.1 Introduction

In Chapter 5, we have discussed the advantages of being able to define precisely meaningful multivariate raw material specifications, i.e., a region that is expected to provide assurance of quality with a certain confidence level for the CQAs. To cope with that several authors suggest using multivariate approaches, such as Partial Least Squares (PLS) regression. Two approaches emerge from the literature when using PLS. The first is based on a direct mapping of good quality final product and associated batches of raw materials in the latent space, followed by a selection of boundaries that minimize or best balance type I and II errors. The second rather defines specification regions by inverting the PLS model for each point lying on final product acceptance limits [83]. Besides, assuming that both variations in raw materials properties and process operating conditions are responsible for CQAs variations, Azari et al. [42] proposed a Sequential Multi-block PLS (SMB-PLS) algorithm considering the direct mapping approach. The SMB-PLS imposes a sequential pathway between the regressor blocks according to the process flowsheet (e.g., raw material properties and process operating conditions), and then uses orthogonalization to separate correlated information between the blocks from orthogonal variations. Hence, the SMB-PLS captures the impact of variations in raw material properties on the process and on CQAs in the first block of latent variables. This allows identifying feedback/feedforward control actions made to compensate for variations in raw material properties. Then, the second block of latent variables captures process variations that are independent from raw material properties and also affect CQAs, e.g., certain (unplanned) excitations due to small changes in the process conditions during their daily operation. For that reason, the SMB-PLS is more efficient to establish the multivariate specifications when raw material properties and process conditions are correlated as it better sorts the contribution of both on the CQA variations.

However, since not only raw material properties influence the quality of the final product, but also process conditions, it is reasonable to consider also the possibility to modify process conditions to compensate for raw material properties variations. Thus, wider raw materials specifications could be used if an effective process control system attenuating most raw material variations is implemented. In this sense, García-Muñoz, Dolph, and Ward [80] already proposed a feed-forward controller based on the PLS model inversion. However, this approach requires solving an optimization problem by a non-linear programming method, where raw material properties are fixed to hard constraints reducing the degrees of freedom to only process conditions. Thus, once a new

raw material batch is received, the controller is executed in order to calculate the combination of the best process conditions, based on the desired CQAs, for such raw material batch. Note that, if too many constraints are specified for raw material properties, the model inversion solution may be forced to move away from the latent model [23]. Besides, this approach makes no attempt to differentiate between correlated and uncorrelated variations in process conditions with raw material properties and, hence, this proposed feed-forward controller does not identify properly the control actions from the past.

The purpose of this work is to develop a novel methodology taking advantage of the SMB-PLS model already discussed in the direct mapping approach but applied into the PLS model inversion approach. Thus, by means of the SMB-PLS model inversion, this methodology allows defining analytically such specifications by considering the possibility to modify process conditions prior to selecting a new raw material batch and, hence, it does not require solving an optimization problem each time a new raw material batch is received. In addition to that, unlike PLS, the SMB-PLS model does identify the variation in process conditions uncorrelated with both raw material properties and known disturbances, which is crucial as the modification of process conditions only must be inferred from such variations as it will be explained in Section 6.3.

6.2 Data requirements

The data required for developing raw materials multivariate specifications by considering the possibility to modify process conditions involves three blocks: $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{Z}$ $(N \times M)$ and $\mathbf{Y}$ $(N \times L)$ are defined as in Chapter 5, and $\mathbf{X}$ $(N \times K)$ is a matrix of inputs which includes a total of K process conditions used to process each one of the N batches of a particular raw material. In this chapter, it is assumed that process conditions refer to process manipulated variables. Finally, batches of raw materials are typically large, and it is assumed that the process will run for a long period at steady state on each batch. Thus, the three data blocks are collected in steady state.

6.3 The SMB-PLS model in the raw material paradigm

The SMB-PLS, presented in Section 2.3, is applied in the raw material paradigm to differentiate between process variations associated with raw material properties from other orthogonal sources of variations. Indeed, it imposes a hierarchical structure where the input blocks are ordered according to the process flowsheet with the first block $\mathbf{Z}$ containing incoming raw material properties, and process data in the second block $\mathbf{X}$. Besides, the SMB-PLS latent space can be expressed, similarly to PLS, but as two blocks of latent variables (Equation 6.1).

$$\mathbf{Y} = \left[\mathbf{T_T}\mathbf{T}^{orth}\right]\left[\mathbf{Q_T}\mathbf{Q}^{orth}\right]^{\mathrm{T}} + \mathbf{F}^{*} = \mathbf{TQ} + \mathbf{F}^{*} \tag{6.1}$$

where $\mathbf{F}^{*}$ are the residuals of $\mathbf{Y}$ after extracting the last SMB-PLS component. Thus, SMB-PLS captures the impact of variation in raw material properties on the process and on $\mathbf{Y}$ in the first modelling step represented by the first block of latent variables, $\mathbf{T_T}$, referring to $[\mathbf{Z}\ \mathbf{X}_{corr}]$. These latent variables allow identifying past operating procedures, and control actions from the past (i.e., feedback/feedforward control) implemented to compensate for raw material properties variations.

Note that as already commented, process data is collected in steady state, and hence, dynamics are not considered. Besides, if the controllers remove the disturbances completely (perfectly), no deviation in $\mathbf{Y}$ in steady state will be captured after a raw material disturbance occurred. In such a case, there will be a correlation between $\mathbf{Z}$ and the manipulated variable in the control loop $\mathbf{X}$, but that information should not be captured by any latent variable since there will be no correlation with $\mathbf{Y}$.

However, in the case of feedforward control on raw material properties, an ideal controller would compensate for any raw material disturbance completely only if it would know the "true" model, which is never the case. In the case of feedback control, if the controller transfer function includes an integrating element (e.g., the I mode in PID controller that seeks to eliminate the residual error according to the historic cumulative error), and if the manipulated variable does not reach an upper or lower bound (i.e., saturation), the impact of the disturbance on $\mathbf{Y}$ should not be captured if the data is collected truly in steady state (i.e., perfect controller). Note that, feedforward controllers are never ideal, nor feedback controllers are perfect. Therefore, these controllers do not compensate perfectly (i.e., there will be a residual effect on $\mathbf{Y}$). In addition to that, regarding the feedback control, the correlations between the manipulated and

controlled variable of the control loop are not causal but anti-causal, that is, these correlations capture the reciprocal of the control transfer functions, leading to the negative inverse of the controller gain for steady state data. Finally, note that control loops are known, and hence, when interpreting the SMB-PLS model, these correlations are not a surprise, but something expected.

In both feedback/feedforward control, the purpose is not to interpret these relationships as causal (which would be wrong), but to account for them in the first block of latent variables. Thus, in the second modelling step, the second block of latent variables, $\mathbf{T}_{orth}$ referring to $\mathbf{X}_{orth}$, is expected to capture only process variations that are independent of raw materials and also affect $\mathbf{Y}$ (e.g., certain (unplanned) excitations). The main aim of this study is to take advantage of the information captured in this second block to improve the control actions from the past in a feedforward control strategy.

6.4 Defining the design space in the latent space by means of SMB-PLS

In this section, a brief overview of defining the DS is shown based on Chapter 5, but by considering the process conditions by means of the SMB-PLS model instead of applying PLS. This is possible as the SMB-PLS latent space (Equation 6.1) is expressed similarly to PLS.

Defining the DS involves finding (predicting) a window of inputs (raw materials properties, process conditions, etc.) for a desired product quality by means of the model inversion. When considering the inversion of a SMB-PLS model, the set of input variables (column vector $\begin{bmatrix} \boldsymbol{z}^{new} \\ \boldsymbol{x}^{new} \end{bmatrix}$) that will yield the desired set of CQAs (column vector $\boldsymbol{y}^{des}$) are obtained by solving the following system of linear equations:

$$\boldsymbol{y}^{des} = \mathbf{Q} \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ \boldsymbol{\tau}_{ortho}^{new} \end{bmatrix} = \mathbf{Q}\boldsymbol{\tau}^{new} \tag{6.2}$$

where $\boldsymbol{\tau}^{new}$ is the vector of scores corresponding to the observation $\begin{bmatrix} \boldsymbol{z}^{new} \\ \boldsymbol{x}^{new} \end{bmatrix}$. The way to calculate $\boldsymbol{z}^{new}$ and $\boldsymbol{x}^{new}$ from $\boldsymbol{\tau}^{new}$ is explained in Section 6.5.

The SMB-PLS model inversion involves solving a system of linear equations represented in a matrix form (Equation 6.2), where there are as many linear

independent equations as the rank of $\mathbf{Y}$ (r_Y), and the number of unknown variables corresponds to the dimensionality of the latent space (A). Commonly, r_Y is lower than A and, hence, Equation 6.2 corresponds to an underdetermined system of linear equations. The multiple solutions $\boldsymbol{\tau}^{new}$ fall into a $(A - r_Y)$-dimensional hyper-plane of the A-dimensional space (i.e., Null Space (NS)), that theoretically yields the same desired set of CQAs. Finally, DS without uncertainty in the latent space is defined by the intersection of the scores fulfilling the specifications' NSs and the Hotelling T^2 confidence region.

In addition to that, when inverting the SMB-PLS model, the prediction uncertainty is accounted in the form of prediction intervals as in Section 5.3.2.2, with a certain confidence level, finding a window within which inputs variables are expected to produce product with CQAs within specification limits with at least the predefined confidence level. This window refers to the so-called High-Confidence Design Space (HC DS).

6.5 Multivariate raw material specification region

The HC DS, defined by the SMB-PLS model, simultaneously considers the raw material properties and process conditions. At this point, one could use such model to define the multivariate raw material specification region (i.e., the Raw Material HC DS) according to two strategies: without or under improved control.

6.5.1 Without improved control

In this section, it is assumed that process variations, correlated with raw material properties will remain in place in the future without any improvement. Thus, establishing specifications in raw material properties aims at penalizing those combinations that are not compensated for by the current control schemes.

A priori, in this strategy, there is no need to consider the orthogonal variations in process conditions and, hence, Raw Material HC DS refers to the HC DS of the SMB-PLS for $[\mathbf{ZX}_{corr}]$. Thus, given a new raw material batch, $\boldsymbol{z}^{new}$, its corresponding $\mathbf{Z}$ scores, $\boldsymbol{\tau}_{\mathbf{Z}}^{new}$, the expected process conditions according to the control actions from the past, $\boldsymbol{x}_{corr}^{new}$, and its corresponding $\mathbf{X}_{corr}$ scores, $\boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new}$, are calculated according to Equation 6.3.

$$\begin{aligned} \boldsymbol{\tau}_{\mathbf{Z}}^{new} &= \mathbf{W^*}_{\mathbf{Z}}^{\mathrm{T}} \boldsymbol{z}^{new} \\ \boldsymbol{x}_{corr}^{new} &= \mathbf{C}_{\mathbf{X}_{corr}} \boldsymbol{\tau}_{\mathbf{Z}}^{new} \\ \boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new} &= \mathbf{W^*}_{\mathbf{X}_{corr}}^{\mathrm{T}} \boldsymbol{x}_{corr}^{new} \end{aligned} \tag{6.3}$$

where $\mathbf{W^*}_{\mathbf{Z}}$ is the $\mathbf{Z}$ block weights transformed to be independent between components, $\mathbf{C}_{\mathbf{X}_{corr}}$ is the correlation coefficient matrix calculated in the first modelling step which directly relates $\boldsymbol{\tau}_{\mathbf{Z}}^{new}$ to $\boldsymbol{x}_{corr}^{new}$, and $\mathbf{W^*}_{\mathbf{X}_{corr}}$ is the $\mathbf{X}_{corr}$ block weights transformed to be independent between components. Both, $\mathbf{W^*}_{\mathbf{Z}}$ and $\mathbf{W^*}_{\mathbf{X}_{corr}}$, are calculated in the first modeling step as it is shown in Appendix 6.A. The corresponding projection into the first block of latent variables, $\boldsymbol{\tau}_{\mathbf{T}}^{new}$, are obtained in the super level score matrix Equation 6.4.

$$\boldsymbol{\tau}_{\mathbf{T}}^{new} = diag\left(\left[\boldsymbol{\tau}_{\mathbf{Z}}^{new} \boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new}\right] \mathbf{W}_{\mathbf{T}}\right) \tag{6.4}$$

where $[\boldsymbol{\tau}_{\mathbf{Z}}^{new} \boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new}]$ refers to the matrix of concatenated score vectors $(A \times 2)$, $\mathbf{W}_{\mathbf{T}}$ is the super weight matrix containing the super weight vectors organized by columns $(2 \times A)$, and *diag* is the matrix-to-vector diagonal operator. Then, if any point, $\boldsymbol{\tau}_{\mathbf{T}}^{new}$ is within the HC DS, one would expect good quality with a certain confidence level for such $\boldsymbol{z}^{new}$. Hence, the Raw Material HC DS (i.e., RM HC DS) can be defined as Equation 6.5).

$$RMHCDS := \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in HCDS\} \tag{6.5}$$

In the case of considering also $\mathbf{X}_{orth}$ in the second modeling step, the Raw Material HC DS would refer to the space defined in Equation 6.6.

$$\begin{aligned} Ort &:= \{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{orth}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathbf{T}}}, \boldsymbol{\tau}_{orth} = \boldsymbol{\tau}_{orth}^{new}\} \\ RMHCDS &:= \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in HCDS \cap Ort\} \end{aligned} \tag{6.6}$$

Note that, the Raw Material HC DS defined in Equation 6.6 a priori requires that the vector of scores referring to orthogonal variations in process conditions, $\boldsymbol{\tau}_{orth}^{new}$, is known beforehand. If this is not the case, it is assumed that $\boldsymbol{\tau}_{orth}^{new}$ will remain on average with respect to the past (i.e., $\boldsymbol{\tau}_{orth}^{new} = \mathbf{0}_{A_{orth}}$) where $\mathbf{0}_{A_{orth}}$ is a zero vector of size A_{orth}). However, as it is unknown, the confidence limits must be calculated disregarding the orthogonal latent space. In other words, the prediction uncertainty, back-propagated in the definition of the specification

confidence limits, must be estimated assuming that the $\boldsymbol{\tau}_{orth}^{new}$ remain on average with respect to the past.

6.5.2 Under improved control

Several works have already emphasized the control actions from the past could be improved in order to compensate for some of the raw materials variability [42, 80, 92]. Hence, wider raw materials specifications can be used if an effective process control system attenuating most raw material variations is implemented. In this sense, the SMB-PLS is particularly useful approaching this strategy as it models the orthogonal variations in process conditions in a second block of latent variables being orthogonal to the first one. Thus, one can infer causality interpretations in the reduced latent space of the second block. This information offers an effective way of manipulating the process conditions, with respect to the control actions from the past, for compensating raw material variations.

In this strategy, given a new raw material batch, $\boldsymbol{z}^{new}$, the expected process conditions according to the control actions from the past, $\boldsymbol{x}_{corr}^{new}$, and the first block of latent variables, $\boldsymbol{\tau}_{\mathbf{T}}^{new}$, are obtained as above (i.e., Section 6.5.1). Then, any raw material batch, resulting in $\boldsymbol{\tau}_{\mathbf{T}}^{new}$, is expected to have good quality with a certain confidence level by modifying process conditions (i.e., it belongs to the Raw material HC DS), if and only if there is any $\boldsymbol{\tau}_{orth} = \boldsymbol{\tau}_{orth}^{new}$ such that $\boldsymbol{\tau}^{new} = \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ \boldsymbol{\tau}_{orth}^{new} \end{bmatrix}$ belongs to the HC DS, where $\boldsymbol{\tau}_{orth}^{new}$ is the score values of the second block of latent variables. From $\boldsymbol{\tau}_{orth}^{new}$, one can figure out how to manipulate the process conditions to compensate for raw material variations according to Equation 6.7.

$$\boldsymbol{x}^{new} = \boldsymbol{x}_{corr}^{new} + \boldsymbol{x}_{orth}^{new} = \boldsymbol{x}_{corr}^{new} + \mathbf{P}_{orth}\boldsymbol{\tau}_{orth}^{new} \tag{6.7}$$

where $\mathbf{P}_{orth}$ is the loading matrix of the second latent block. Note that, $\boldsymbol{\tau}_{orth}^{new}$ represents the locus of the $\boldsymbol{x}_{orth}^{new}$ projections within the HC DS given a new raw material batch. Therefore, if it exists, the control actions could be improved in different ways without leaving the DS, which provides operational flexibility in process improvement.

Finally, we can define analytically the Raw Material HC DS, prior to selecting a new raw material, as the projection of the HC DS onto the space defined by the first block of latent variables as Equation 6.8.

$$RMHCDS := \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} = T_{\mathbf{P_T}}[\boldsymbol{\tau}], \forall \boldsymbol{\tau} \in HCDS\} \tag{6.8}$$

where $T_{\mathbf{P_T}}$ is the linear transformation that projects from $\mathbb{R}^{A_{\mathbf{T}}+A_{orth}}$ to $\mathbb{R}^{A_{\mathbf{T}}}$ defined by the matrix $\mathbf{P_T} = [\mathbf{I}_{A_{\mathbf{T}}}\mathbf{0}_{A_{\mathbf{T}},A_{orth}}]$, $\mathbf{I}_{A_{\mathbf{T}}}$ is the identity matrix of size $A_{\mathbf{T}}$, $\mathbf{0}_{A_{\mathbf{T}},A_{orth}}$ is a zero matrix of size $A_{\mathbf{T}} \times A_{orth}$, and $A_{\mathbf{T}}$ and A_{orth} are the latent dimensionality of the first and second block, respectively.

6.6 Presence of known disturbances affecting control actions

Until now, we have assumed that orthogonal process variations to raw material properties and related to CQAs are due to certain (unplanned) excitations. However, process conditions could present variations due to feedforward compensation for some known disturbances.

This issue needs special attention as if one decides to ignore the known disturbance for not being manipulatable, the SMB-PLS could model, in the orthogonal block, variations in process conditions that may be related to such disturbance. The fact that the correlation between process conditions and the known disturbance could still explain variations in CQAs is because the control adjustment may not be perfect (i.e., the effect of the known disturbances is not removed completely). This will yield misleading causality relations in the reduced latent space. Therefore, we suggest adding an intermediate block $\mathbf{D}$ ($N \times O$) being a matrix of inputs which includes a total of O known disturbances measured in each one of the N batches of a particular raw material. Thus, the SMB-PLS algorithm includes an intermediate modelling step that captures the impact of variation in disturbances orthogonal to $\mathbf{Z}$ (i.e., $\mathbf{D}_{orth}$) on the process and on $\mathbf{Y}$, represented by latent variables $\mathbf{T_D}$. This intermediate block of latent variables allows identifying control actions from the past implemented to compensate for disturbances not related to raw material properties. This ensures that the last modeling step only model certain (unplanned) excitations in process conditions, $\mathbf{X}_{orth}$, from which causality can be inferred.

In the same way as Section 6.5.1 but including the disturbance space, the Raw Material HC DS without improved control would refer to the space defined in Equation 6.9.

$$\begin{aligned} Dis &:= \{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}, \boldsymbol{\tau}_{orth}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathbf{T}}}, \boldsymbol{\tau}_{\mathbf{D}} = \boldsymbol{\tau}_{\mathbf{D}}^{new}, \boldsymbol{\tau}_{orth} \in \mathbb{R}^{A_{orth}}\} \\ Ort &:= \{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}, \boldsymbol{\tau}_{orth}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathbf{T}}}, \boldsymbol{\tau}_{\mathbf{D}} \in \mathbb{R}^{A_{\mathbf{D}}}, \boldsymbol{\tau}_{orth} = \boldsymbol{\tau}_{orth}^{new}\} \\ RMHCDS &:= \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in HCDS \cap Ort \cap Dis\} \end{aligned} \tag{6.9}$$

Equation 6.9 assumes that both the disturbance and the orthogonal space are not manipulatable and, hence, they must be defined as constraints, Dis and Ort respectively, that intersect with the HC DS. However, if control actions can be improved by means of the orthogonal space, such space must be projected onto the remaining space in the same way as Section 6.5.2. Thus, the Raw Material HC DS, by considering the possibility to modify process conditions prior to selecting a new raw material batch, can be defined analytically as the intersection between the projection of the HC DS onto the first and second block of latent variables (i.e., Pr), and the subspace defined by $\boldsymbol{\tau}_{\mathbf{D}}^{new}$ (i.e., Dis), as it is shown in Equation 6.10.

$$\begin{aligned} Dis &:= \{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathbf{T}}}, \boldsymbol{\tau}_{\mathbf{D}} = \boldsymbol{\tau}_{\mathbf{D}}^{new}\} \\ Pr &:= \{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}) : \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}} \\ \boldsymbol{\tau}_{\mathbf{D}} \end{bmatrix} = T_{\mathbf{P_{TD}}}[\boldsymbol{\tau}], \forall \boldsymbol{\tau} \in HCDS\} \\ RMHCDS &:= \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in Dis \cap Pr\} \end{aligned} \tag{6.10}$$

where $T_{\mathbf{P_{TD}}}$ is the linear transformation that projects from $\mathbb{R}^{A_{\mathbf{T}}+A_{\mathbf{D}}+A_{orth}}$ to $\mathbb{R}^{A_{\mathbf{T}}+A_{\mathbf{D}}}$ defined by the matrix $\mathbf{P_{TD}} = [\mathbf{I}_{A_{\mathbf{T}}+A_{\mathbf{D}}} \mathbf{0}_{A_{\mathbf{T}}+A_{\mathbf{D}},A_{orth}}]$, $\mathbf{I}_{A_{\mathbf{T}}+A_{\mathbf{D}}}$ is the identity matrix of size $A_{\mathbf{T}} + A_{\mathbf{D}}$, $\mathbf{0}_{A_{\mathbf{T}}+A_{\mathbf{D}},A_{orth}}$ is a zero matrix of size $A_{\mathbf{T}} + A_{\mathbf{D}} \times A_{orth}$, and $A_{\mathbf{D}}$ is the latent dimensionality of the disturbance block.

Note that, the Raw Material HC DS defined in Equation 6.9 and Equation 6.10 a priori requires that the vector of scores referring to orthogonal variations in disturbances, $\boldsymbol{\tau}_{\mathbf{D}}^{new}$, is known beforehand. If this is not the case, it is assumed that $\boldsymbol{\tau}_{\mathbf{D}}^{new}$ will remain on average with respect to the past (i.e., $\boldsymbol{\tau}_{\mathbf{D}}^{new} = \mathbf{0}_{A_{\mathbf{D}}}$ where $\mathbf{0}_{A_{\mathbf{D}}}$ is a zero vector of size $A_{\mathbf{D}}$), and the confidence limits must be calculated disregarding the disturbance latent space.

6.7 Industrial case study

Description of the dataset

A simulated polymer extrusion film blowing process was used to generate data in order to illustrate how to define multivariate specification regions for incoming raw materials [45, 78]. The dataset consists of two regressor blocks ($mathbfZ$ and $mathbfX$) and a response block ($mathbfY$). The raw material block ($mathbfZ$) contains the following polymer resin properties: ten temperature dependent viscosities (η), heat capacity (C_p), and density (ρ).

The second block ($mathbfX$) contains 3 process conditions, namely the air temperature (T_a), the polymer flow rate (Q) and the cooling air flow rate represented by the maximum local heat transfer coefficient along the film bubble (h_0). The response block ($\boldsymbol{y}$) is characterized by one quality attribute of the film, which is the full stress in the machine direction ($FMDS$), with a lower specification defined as its average.

The dataset was simulated in two steps. First, variability was introduced in raw material properties and process conditions in such a way that both regressor blocks affect $\boldsymbol{y}$, but variations in $\mathbf{Z}$ and $\mathbf{X}$ are uncorrelated to each other (initially blocks are orthogonal). This was achieved by introducing random variations in raw material properties ($\mathbf{Z}$) and processing conditions ($\mathbf{X}$) to simulate their effect on product quality. However, the variables within each block are collinear to a certain extent. Regarding $\mathbf{Z}$, correlation is due to viscosities measured at different temperatures. In a second step, similar uncorrelated variations were again implemented in both regressor blocks, but between block correlations were introduced by a feedforward controller, added to attenuate variations caused by raw material properties. This controller corrects for some of the variability in the polymer heat capacity C_p by adjusting the flow rate Q. The processing of 50 raw materials batches were simulated.

Building the SMB-PLS model

Three components were found sufficient to capture the impact of raw material properties (and correlated process variations) on $\boldsymbol{y}$ in the first modelling step. One additional component was also needed in the second modelling step to model the effect of orthogonal variations in process conditions on the remaining variations in $\boldsymbol{y}$. The goodness of fit, $R^2_{\mathbf{Y}_{cum}}$ (i.e., variability percentage explained by the model) for each one of the input blocks, $\mathbf{Z}$ and $\mathbf{X}$, and the output block, $\boldsymbol{y}$, and each component, is presented in Figure 6.1.

Figure 6.1 shows that the first three components of the first modelling stage explain 74.89% of the information in $\mathbf{Z}$ and 22.10% of the information in $\mathbf{X}$ that was correlated with $\mathbf{Z}$, to explain a great percentage of the response variability (86.22%). Component 4 (the unique component of the second modelling stage) shows that the 62.12% of the variation in $\mathbf{X}$, not related to $\mathbf{Z}$, is able to explain 8.65% of the response variability. Since the last two components explain the greatest variation in $\mathbf{X}$ (Figure 6.1), Figure 6.2 shows the bi-plot of the block weights and $\boldsymbol{y}$ loadings for these components to understand the behavior of process conditions. Figure 6.2 reveals that the explained variation in the polymer flow rate Q seems to be related to raw material properties according

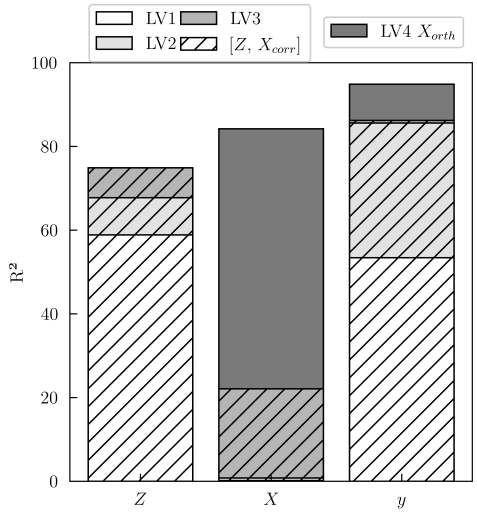


Figure 6.1: Explained **Z**, **X** and $\boldsymbol{y}$ variability for the SMB-PLS model depending on either the number of latent variables (LVS) or the two blocks of latent variables (LV1-LV3 explain the first block $[\mathbf{ZX}_{corr}]$, and LV4 explains the second block $\mathbf{X}_{orth}$).

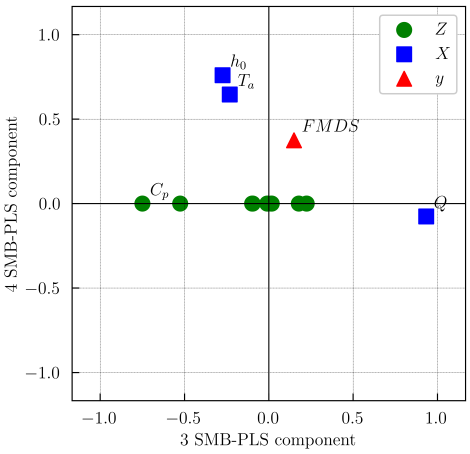


Figure 6.2: Bi-plot of the block weights and $\boldsymbol{y}$ loadings for last two components.

to the third component. In fact, Q is strongly negatively correlated with the heat capacity C_p because when C_p increases, Q is reduced (as a result of the feedforward controller) to mitigate its impact on quality product. However, component 4 shows that Q barely presents orthogonal variations to raw material properties related to y. By contrast, the air temperature T_a and the cooling air flow rate h_0 present orthogonal variations to raw material properties highly correlated with each other, from which one can infer causality in the reduced latent space. In other words, for any active change in the process conditions of T_a and h_0, being consistent with the correlation structure modeled by the latent orthogonal space, the SMB-PLS model will reliably predict the changes in $\boldsymbol{y}$.

Defining the high-confidence design space

The HC DS is defined with at least a 90% confidence level of obtaining superior or equal $FMDS$ values to the average of calibration data (lower specification limit). Figure 6.3 shows the HC DS by showing the calibration data for the first two components of the first modelling step [$\mathbf{ZX}_{corr}$] and the orthogonal one. The third [$\mathbf{ZX}_{corr}$] component from the first modelling step is omitted.

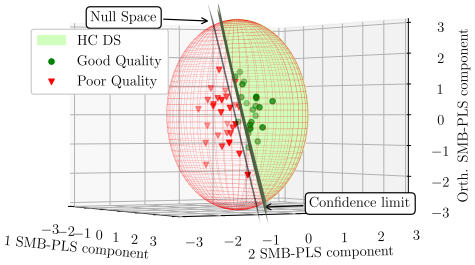


Figure 6.3: Graphical definition of the High-Confidence Design Space by showing calibration data.

One would expect that, of the batches lying within the HC DS, 90% or more would be acceptable batches. Indeed, the negative predictive value[1] is 95%. On the other hand, the HC DS leads to 3.57% type I risk and 13.64% type II risk. This means that if only batches lying within the HC DS are accepted, 13.64% of unacceptable batches of raw materials had been accepted at the expense of rejecting 3.57% of acceptable batches.

Multivariate raw material specification region without improved control

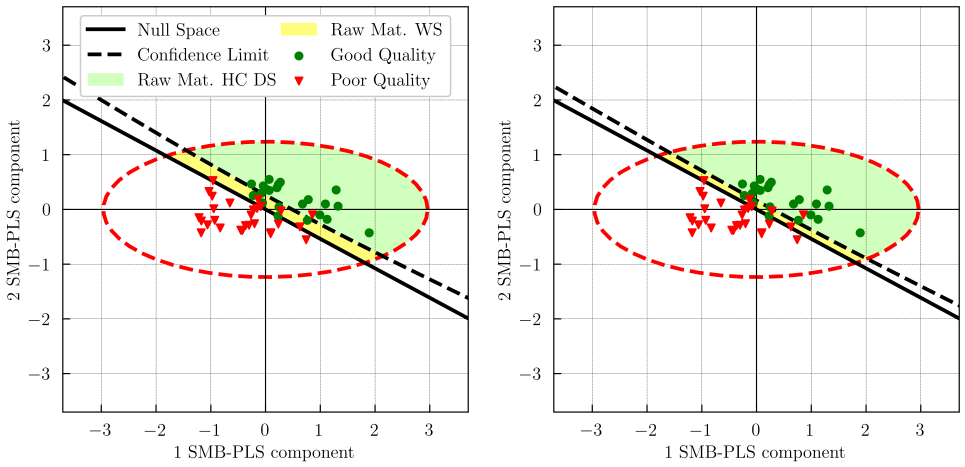
In this section, it is assumed that process variations, correlated with raw material properties due to control actions through manipulated variables, will remain in place in the future without any improvement. In such a case, a priori there is no need to consider process conditions to establish the specification regions associated with the raw material properties and, hence, one could define this region by the PLS model inversion by considering only raw material

[1]The negative predictive is the proportion of batches that actually result in a good product out of all those within the HC DS.

properties as in [14]. By contrast, without improved control, we propose to define the raw material HC DS as the HC DS of the first block of latent variables, referring to $[\mathbf{ZX}_{corr}]$ of the SMB-PLS. The amount of information/variability contained in the first input block depends on $\mathbf{Z}$ as $\mathbf{X}_{corr}$ does not provide a new source of variability. Therefore, the predictive power of both, PLS for $\mathbf{Z}$ with three components and SMB-PLS for only $[\mathbf{ZX}_{corr}]$ with three components, are the same. Consequently, a priori, the classification performance of new raw material batches is expected to be equivalent. However, incorporating process data by means of the SMB-PLS presents some advantages with respect to PLS as we will see below.

As Azari et al. [42] discussed, the SMB-PLS provides great insights in agreement with process knowledge for the effects of material variations and correlated process conditions (control schemes mainly). Firstly, since the SMB-PLS also can model the orthogonal variations in process conditions by the second block of latent variables, it provides a great capability for diagnosing assignable causes of such variations. In fact, by interrogating the underlying SMB-PLS model, one can extract diagnostic or contribution plots which reveal the group of process conditions making the greatest contributions to the deviations in the squared prediction errors, and the scores [28, 93]. In addition to that, the second block of latent variables provides a better understanding of the response variability with respect to both PLS and SMB-PLS for only $[\mathbf{ZX}_{corr}]$ (this increases the response variance percentage up to 95.87%). The latter results in less prediction uncertainty, and this affects the definition of the Raw Material HC DS. Figure 6.4 shows the graphical definition of such space for the SMB-PLS depending on whether the $\mathbf{X}_{orth}$ is considered or not.

Figure 6.4 shows graphically that, as expected, the Raw Material DS without uncertainty (i.e., the union of the Raw Material HC DS and the Raw Material WS) are equal regardless of whether $\mathbf{X}_{orth}$ is considered or not. In addition to that, the less uncertainty there is, the more similar the Raw Material HC DS and the Raw Material DS without uncertainty are. For that reason, the SMB-PLS Raw Material HC DS becomes wider when incorporating the $\mathbf{X}_{orth}$ block as can see in Figure 6.4. Therefore, it can be concluded that, for model building, the SMB-PLS provides useful information in order to achieve a higher level of process understanding when considering the $\mathbf{X}_{orth}$. However, it is crucial to bear in mind that for exploiting the model, Figure 6.4b requires that orthogonal variations are known beforehand. Indeed, the Raw Material HC DS shown in Figure 6.4b arises from the assumption that the orthogonal variation in process conditions will remain at the average value with respect to the past. If they are not known beforehand, it is assumed that $\boldsymbol{\tau}_{\mathbf{D}}^{new}$ will



(a) considering only $[\mathbf{ZX}_{corr}]$.

(b) considering $[\mathbf{ZX}_{corr}]$ and also $\mathbf{X}_{orth}$ assuming that orthogonal variations remain at the average value.

Figure 6.4: Graphical definition of the Raw Material High-confidence Design Space (Multivariate Raw Material Specification region) and Raw Material Warning Space when:

remain on average with respect to the past and, hence, the confidence limits must be calculated disregarding the orthogonal block yielding Figure 6.4a.

Multivariate raw material specification region under improved control

Let us consider the HC DS defined previously (see Figure 6.3). Then, a new raw material batch is considered prior to the manufacturing process (i.e., only raw material properties are known). Thus, the red triangle in Figure 6.5 refers to the projection onto the latent space assuming that the control actions of process conditions remain in place, and the orthogonal variation in process conditions remain at the average value with respect to the past (i.e., the orthogonal component is null): $\begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ 0 \end{bmatrix}$.

In such a case, as shown in Figure 6.5, this batch would be outside the specification region. However, if the orthogonal component is modified orthogonally, such batch can become part of the specification region (deep blue solid line). This is a batch that, a priori, would give place to a film with an unacceptable response value ($FMDS$), but that by improving the control actions it would

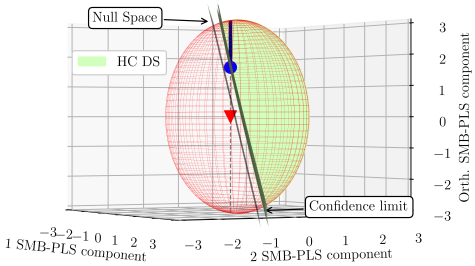


Figure 6.5: Graphical definition of the High-Confidence Design Space by showing the projection of the new raw material batch when: i) orthogonal variation in process conditions remain at the average value with respect to the past (red triangle), and ii) control actions are improved (blue circle).

yield a film with an acceptable response value ($FMDS$). As commented, since the control actions could be improved in different ways without leaving the HC DS, it provides operational flexibility in process improvement. As an example, the blue circle is selected among all process conditions yielding the score: $\begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ \boldsymbol{\tau}_{orth}^{new} \end{bmatrix}$. This solution belongs to the latent space and, therefore, it behaves in the same way as the ones used to create the model, in the sense that the correlation structure of the model is respected. A logical question then arises: how to manipulate the process conditions to get this solution? The answer is applying Equation 6.7. This is shown graphically in Figure 6.6.

Figure 6.6 shows the time series of manipulated variables with their historical limits. The red triangles refer to the expected process conditions due to the control actions from the past, $\boldsymbol{x}_{corr}^{new}$, while the blue circles show the final conditions after improving such control for compensating raw material variations. The latter arises from adding the orthogonal variation, $\boldsymbol{x}_{orth}^{new}$, which is obtained as $\mathbf{P}_{orth}\boldsymbol{\tau}_{orth}^{new}$. As expected, the flow rate Q is barely modified with respect to the expected control actions because, as it is shown in Figure 6.2, this process condition does not present a relevant amount of orthogonal variation related to y. By contrast, the air temperature T_a and the cooling air flow rate h_0 do and, hence, one can infer causality in the reduced latent space in order to attenuate most raw material variations. Note that, since causality is inferred in

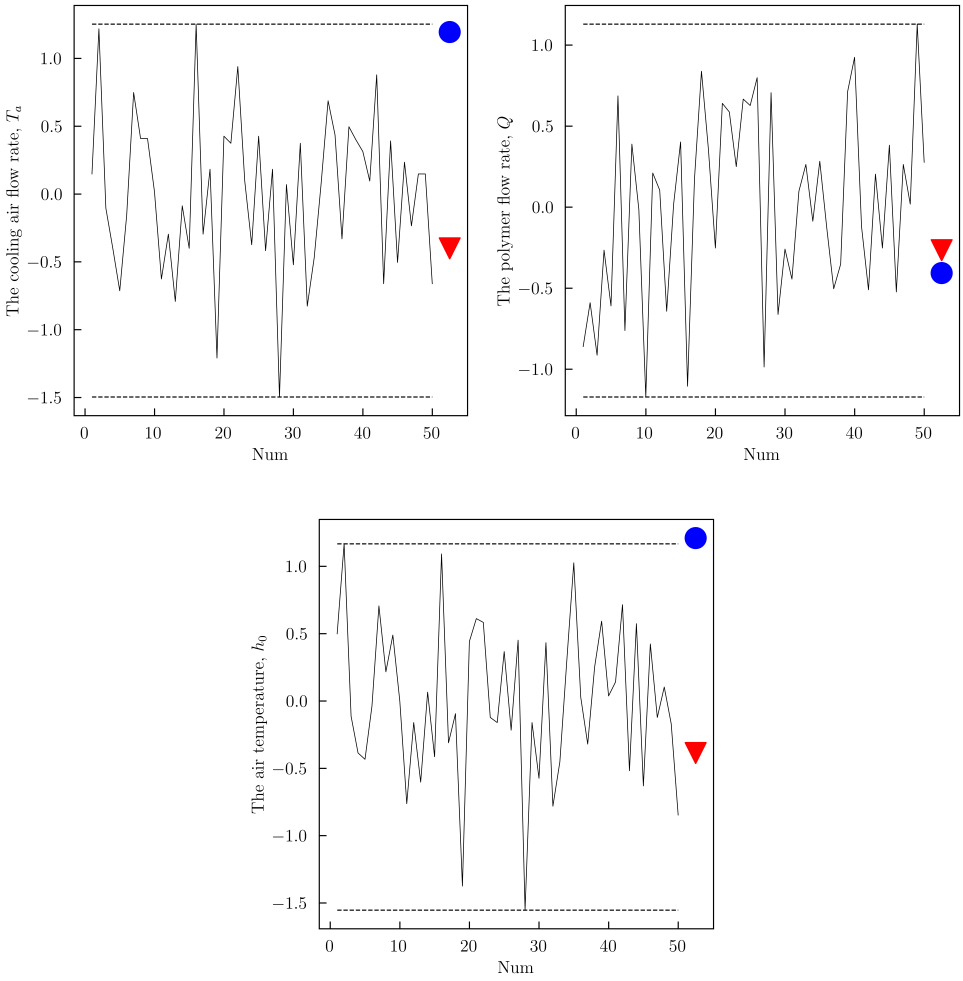
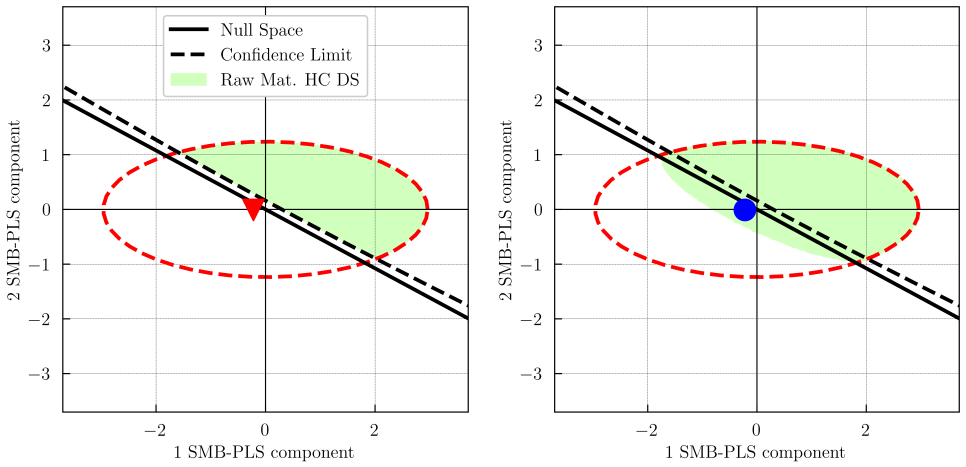


Figure 6.6: Time series of process conditions, T_a, Q and h_0, and new setpoints in two scenarios: no improved control (red triangle) and under improved control (blue circle).



(a) without improved control by showing the projection of the new raw material batch as a red triangle.

(b) under improved control by showing the projection of the new raw material batch as a blue circle .

Figure 6.7: Raw Material High-Confidence Design Space:

the reduced latent space, process conditions are manipulated to be consistent with the latent orthogonal space shown in Figure 6.2.

Finally, the Raw Material HC DS, by considering the possibility of modifying process conditions prior to selecting a new raw material batch, can be defined analytically as the projection of the HC DS onto the space defined by the first block of latent variables, according to Equation 6.8 (see Figure 6.7b).

Figure 6.7 shows that Raw Material HC DS is expanded when considering the possibility to modify process conditions for compensating raw material variations. Thus, one may be able to accept raw materials that will yield products with perfectly satisfactory quality properties as a consequence of the process conditions modification, as in the considered new raw material batch.

Presence of known disturbances affecting control actions

Process conditions could present variations due to feedforward compensation for some known disturbances. In fact, in the simulated polymer extrusion film blowing process, the air temperature T_a refers to the air ambient temperature. In such a case, this process condition cannot be manipulated but it is a major process known disturbance affecting cooling conditions and hence, quality properties. In addition to that, the cooling air flow rate h_0 is manipulated by a feedforward controller to compensate for some variations in the ambient air temperature T_a. To identify these variations as explained in Section 6.6, an intermediate block $\mathbf{D}$ must be added. In this case, since there is only one known disturbance, the intermediate block is defined as vector $\boldsymbol{d}$. The goodness of fit for the SMB-PLS model, R^2, for each one of the input blocks, $\mathbf{Z}$, $\boldsymbol{d}$ and $\mathbf{X}$, and the output block, $\boldsymbol{y}$, and each component, is presented in Figure 6.8.

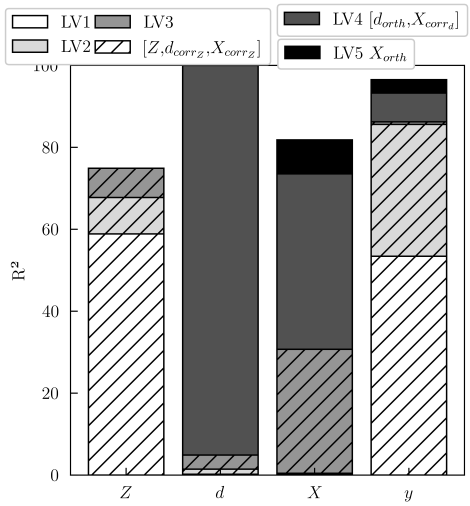
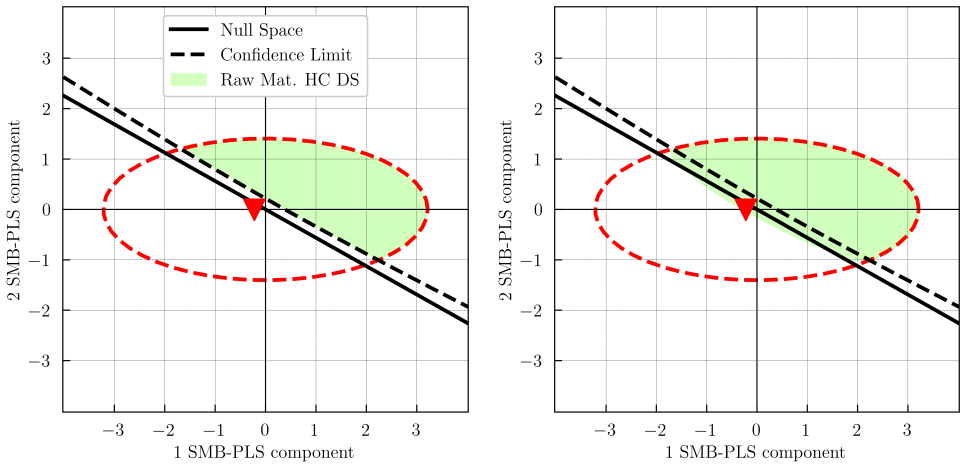


Figure 6.8: Explained $\mathbf{Z}$, $\boldsymbol{d}$, $\mathbf{X}$ and $\boldsymbol{y}$ variability for the SMB-PLS model depending on either the number of latent variables (LVs) or the three blocks of latent variables (LV1-LV3 explain the first block $[\mathbf{Z}\boldsymbol{d}_{corr_\mathbf{Z}}\mathbf{X}_{corr_\mathbf{Z}}]$, LV4 explains the second block $[\boldsymbol{d}_{orth}\mathbf{X}_{corr_\boldsymbol{d}}]$, and LV5 explains the last block $\mathbf{X}_{orth}$).

Figure 6.8 shows that three components were found sufficient to capture the impact of raw material properties (and correlated disturbances and process variations) on $\boldsymbol{y}$ in the first modelling step explaining 74.89% of the information in $\mathbf{Z}$, 4.89% of the information in $\boldsymbol{d}$ and 30.70% of the information in $\mathbf{X}$, to explain a high percentage of the response variability (86.22%). One additional



(a) without improved control. **(b)** under improved control.

Figure 6.9: Raw Material High-Confidence Design Space prior to knowing T_a by showing the projection of the new raw material batch as a red triangle:

component was also needed in the second modelling step to model the effect of orthogonal variations in disturbances (and correlated process variations) on the remaining variations in y. This component shows that the 95.11% of the variation in d, not related to $\mathbf{Z}$, is able to explain 42.83% of $\mathbf{X}$ and 7.02% of the response variability. The latter represents variations in $\boldsymbol{d}$ affecting y, but not compensated by the controller. Finally, one component was used to capture the orthogonal variations in process variations on the remaining variations in $\boldsymbol{y}$ showing that the 8.28% of variation in $\mathbf{X}$, not related to $\mathbf{Z}$ and $\boldsymbol{d}$, is able to explain 3.27% percentage of the response variability.

The most common case is that the ambient air temperature T_a is not known when receiving a raw material batch. Therefore, it is assumed that, for exploiting the model, this disturbance remains on average with respect to the past and, hence, the confidence limits are calculated disregarding the disturbance block. Thus, Figure 6.9 shows the Raw Material HC DS prior to knowing T_a without improved control (Figure 6.9a), and by considering the possibility to modify process conditions (Figure 6.9b), using Equations 6.9 and 6.10, respectively.

Figure 6.9 shows that the Raw Material HC DS is slightly expanded when considering the possibility to modify process conditions for compensating raw material variations. Indeed, the new raw material batch illustrated in the

Subsection 7.5 would be on the border of the Raw Material HC DS, since there is no control action that allows being within the HC DS. This happens because only 3.27% of the response variability can be inferred as the effect of 8.28% of the variation in $\mathbf{X}$ not related to $\mathbf{Z}$ and $\boldsymbol{d}$. The latter may not be sufficient to carry out effective improvement in the control action.

6.8 Conclusions

In this chapter, we propose a novel approach for defining analytically the multivariate raw material specification region by considering the possibility to modify process conditions to compensate for raw material properties variations. This methodology is based on the SMB-PLS model inversion where prediction uncertainty is back-propagated. The most remarkable advantages of the proposal approach are:

- It can be used with historical data (i.e., daily production data not coming from any experimental design but with varying raw material properties, typical from Industry 4.0 environment).
- It considers a multivariate approach providing much insight into the correlated nature of raw material properties and process conditions. Besides, the SMB-PLS does identify the variation in process conditions uncorrelated with raw material properties and known disturbances, which is crucial to implement an effective process control system attenuating most raw material variations.
- It allows expanding the multivariate raw material specification when considering the possibility to modify process conditions and, hence, one may potentially be able to accept lower cost raw materials that will yield products with perfectly satisfactory quality properties.

Definitely, this methodology takes advantage of the variation in process conditions uncorrelated with raw material properties and known disturbances to expand the raw material specification. However, this variation may result insufficient to carry out effective improvement. In such a case, process excitation would be needed by running design of experiments on process operating conditions.

Appendices

6.A SMB-PLS weights transformed to be independent between components

This appendix is applicable to blocks, $\mathbf{Z}$ and $\mathbf{X}_{corr}$ (hereinafter called $\mathbf{B}$). The weights matrix, $\mathbf{W_B}$, do not directly relate the matrix $\mathbf{B}$ to the score matrix $\mathbf{T_B}$, as $\mathbf{B}$ is deflated after each component by the loading matrix $\mathbf{P_B}$. However, the weights, $\mathbf{W_B}$, can be transformed to $\mathbf{W_B^*}$ by $\mathbf{M}$ (Equation 6.A.1) and, thus, $\mathbf{W_B^*}$ does directly relate $\mathbf{B}$ to $\mathbf{T_B}$, (Equation 6.A.2).

$$\mathbf{W_B^*} = \mathbf{W_B M} \tag{6.A.1}$$

$$\mathbf{T_B} = \mathbf{B W_B^*} = \mathbf{B W_B M} \tag{6.A.2}$$

If multiplying both sides of Equation 6.A.2 by the transpose of the super score matrix, $\mathbf{T_T}$, the $\mathbf{M}$ matrix can be expressed asEquation 6.A.3.

$$\mathbf{M} = \left(\mathbf{T_T^T B W_B}\right)^{-1} \left(\mathbf{T_T^T T_B}\right) \tag{6.A.3}$$

On the other hand, $\mathbf{T_T}$ are good "summaries" of $\mathbf{B}$ according to the loading matrix $\mathbf{P_B}$ (Equation 6.A.4).

$$\mathbf{B} = \mathbf{T_T P_B^T} + \mathbf{E_B} \tag{6.A.4}$$

where $\mathbf{E_B}$ is the residual matrix. Then, multiplying both sides of Equation 6.A.4 by the transpose of $\mathbf{T_T}$, the Equation 6.A.5 is obtained.

$$\mathbf{T_T^T B} = \mathbf{T_T^T T_T P_B^T} + \mathbf{T_T^T E_B} = \mathbf{T_T^T T_T P_B^T} \tag{6.A.5}$$

Note that, the super scores columns vectors of $\mathbf{T_T}$ are orthogonal to $\mathbf{E_B}$. Substituting Equations 6.A.3 and 6.A.5 in Equation 6.A.1, the relation between $\mathbf{W_B}$ and $\mathbf{W_B^*}$ is obtained according to Equation 6.A.6.

$$\mathbf{W_B^*} = \mathbf{W_B} \left(\mathbf{T_T^T T_T P_B^T W_B}\right)^{-1} \left(\mathbf{T_T^T T_B}\right) \tag{6.A.6}$$

Note that, regarding the block $\mathbf{X}_{corr}$, the matrix $A = \mathbf{T}_{\mathbf{T}}^{\mathrm{T}}\mathbf{T}_{\mathbf{T}}\mathbf{P}_{\mathbf{X}_{corr}}^{\mathrm{T}}\mathbf{W}_{\mathbf{X}_{corr}}$ may be rank-deficient as more latent variables could be extracted than the rank of $\mathbf{X}_{corr}$ and, hence, $\mathbf{A}$ would not be invertible. In such a case, $\mathbf{X}_{corr}$ and $\mathbf{T}_{\mathbf{X}_{corr}}$ cannot be directly related by $\mathbf{W}^{*}_{\mathbf{X}_{corr}}$.

Chapter 7

Latent space-based multivariate capability index

Part of the content of this chapter has been included in:

[19] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "A latent space-based Multivariate Capability Index: A new paradigm for raw material supplier selection in Industry 4.0," 2023, **SUBMITTED**

7.1 Introduction

Capability indexes (CIs) are used to estimate how likely is a given supplier of raw materials to meet consumer's requirements for these raw materials. It is therefore usually used as a criterion for selecting process's raw material suppliers [94]. When the process must be able to attain specification limits for multiple raw material properties, the use of independent univariate CIs for each one of them is often used with the implicit assumption that these properties are independent from one another. However, the use of this approach can lead to misinterpretation because the capability of each property is analyzed independently assuming that specifications are hyperrectangles, namely without considering its correlation to other properties of the raw material. Hence, different approaches were recently proposed to develop multivariate CIs which quantify with a single index the goodness of a raw material supplier by considering multiple properties simultaneously [95]. To the authors' knowledge, all of them assume that the specifications are hyperellipsoids defined in the original space of raw material properties without considering a precise relationship to the CQAs. Nevertheless, these specifications may result meaningless, i.e., based on these specifications, accepted raw material batches, once processed, the manufactured product is out of CQAs specifications. This may force the customer to be more restrictive (i.e., narrowing the raw material specification region) with the underlying assumption that accepting minimal variations in raw materials results in minimal fluctuations in CQAs. However, this would increase the costs in the acquisition of raw material batches with tighter variations in their properties [78]. But would it be feasible to define meaningful multivariate raw material specifications considering a precise relationship to the CQAs? This would allow increasing the number of potential suppliers, by allowing a wider range of raw material properties ($\mathbf{Z}$), without compromising the Critical Quality Attributes (CQAs) of the final product ($\mathbf{Y}$). In this sense, in Chapter 5 a novel approach has been proposed to define an analytical expression for defining the High-Confidence Design Space (HC DS) of raw materials, i.e., the High-Confidence Raw Material Specification Region (HC RMSR) where there is assurance of quality with a certain confidence level for the CQAs of the final product. The logical extension of defining meaningful specifications is to measure how far suppliers can consistently operate inside such latent space-based raw material specifications (i.e., HC-RMSR). This is the purpose of the novel Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) proposed in this chapter. The $LSb\text{-}MC_{pk}$ provides information on the ability of each supplier of a particular raw material to produce a certain percentage of the final product within its CQAs specifications. And this information can be obtained at the reception of the supplier´s raw material,

before producing a single unit of the product, and it can be used for ranking and selecting suppliers.

This chapter is organized as follows. Data requirements for defining the Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) are first discussed (Section 7.2). The proposed $LSb\text{-}MC_{pk}$ is then presented (Section 7.4), followed by the description of diagnosing assignable causes (Section 7.5). Finally, all methodology is summarized by a scheme (Section 7.6), and then illustrated by a case study Section 7.7.

7.2 Data requirements

The data required for developing the $LSb\text{-}MC_{pk}$ involves two blocks, $\mathbf{Z}$ $(N \times M)$ and $\mathbf{Y}$ $(N \times L)$ (as in Chapter 5) to define firstly the multivariate raw material specification region in the latent space. Then, only a set of raw materials from a given supplier is required to define its respective $LSb\text{-}MC_{pk}$.

7.3 Supplier's raw material operating space (RMOS)

The purpose of this section is to define the so-called supplier's Raw Material Operating Space (RMOS), that is, a region in the latent space connecting the raw material properties with the Critical Quality Attributes (CQAs) of the product manufactured that will contain the batches of a particular raw material supplier at a certain confidence level. Thus, the RMOS refers to the space where the supplier's raw material samples are expected to be located in the latent space connecting the properties of these raw material samples with the CQAs of the corresponding manufactured product.

To provide a reliable definition of the RMOS, a number of raw material samples for a particular supplier are required. As in any statistical estimation procedure, the larger the sample size, the better. Based on the authors' experience a minimum of 30 samples is recommended[1]. As the RMOS for any supplier is calculated from an empirical model, it is required to check if the supplier behaves in the same way as the historical raw material samples used to fit the PLS model. This is done by projecting the set of supplier´s raw material samples into the PLS latent space and calculating the SPE statistics (Equation 2.6). Then, if it is acceptable to consider that the samples from this

[1]Note that, given a supplier's raw material, the appropriate sample size will be conditioned by the degree of uncertainty of the estimated parameters used for defining the RMOS. If this uncertainty is higher than what is acceptable, the sample size should be increased.

supplier meet the correlation structure from the past ($SPE < SPE_{lim}$), the variability of the projected samples into the latent space (i.e., scores) can be modelled under the assumption that these scores follow a multivariate normal distribution (since they are linear combination of random variables [96]). In fact, this multivariate normal distribution in the latent space is characterised by the centroid vector ($\boldsymbol{\tau}_G$) and the covariance matrix ($\mathbf{S}$) of the scores obtained from the set of raw material samples.

Hence, given a certain confidence level, a region where we expect to operate the process according to the supplier is defined. This region is called Raw Material Operating Space (RMOS) and it is represented by the equation of an ellipsoid (Equation 7.1).

$$(\boldsymbol{\tau} - \boldsymbol{\tau}_G)^{\mathrm{T}} \mathbf{S}^{-1} (\boldsymbol{\tau} - \boldsymbol{\tau}_G) \leq c^2 \tag{7.1}$$

where τ is any score belonging to the RMOS, and c^2 represents the size of the elliptical region. The latter is the estimated squared Mahalanobis distance of any point belonging to the envelope of this ellipsoid to the centre, and can be considered distributed as [30]:

$$c^2 \sim \frac{A\,(S+1)\,(S-1)}{S\,(S-A)} F_{A,S-A} \tag{7.2}$$

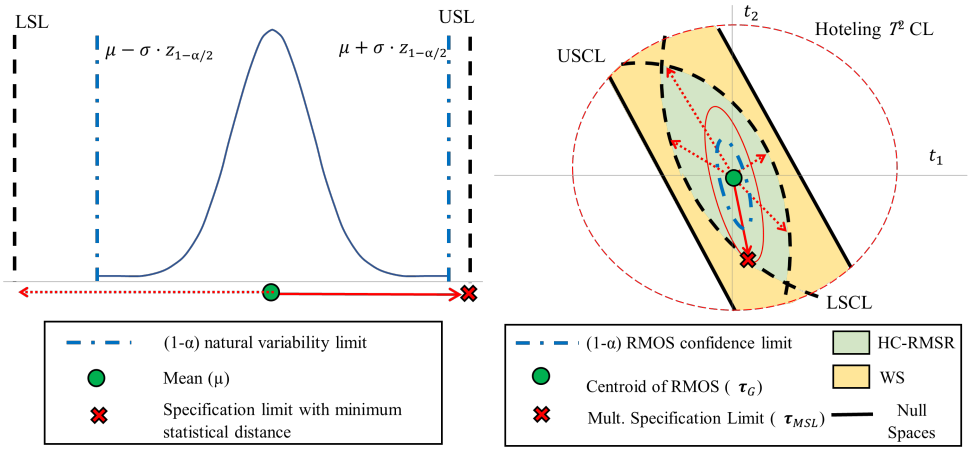
where S is the number of raw material samples. When considering a certain confidence level (α), the value of c^2 is calculated as follows:

$$c^2_{1-\alpha} = \frac{A\,(S+1)\,(S-1)}{S\,(S-A)} F_{(1-\alpha);A,S-A} \tag{7.3}$$

7.4 Latent space-based multivariate capability index

Once the HC-RMSR and the RMOS are defined, the proposed Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) can be calculated in an analogous way as it is done in the univariate case. This index quantifies the capacity of each supplier of the raw material of providing assurance of quality with a certain confidence level for the CQAs of the manufactured product.

Figure 7.1 shows graphically the similarity between the latent space-based multivariate capability index, $LSb\text{-}MC_{pk}$, assuming a two-component model,



(a) Univariate capability index, C_{pk}, in the original space.

(b) Latent space-based multivariate capability index, $LSb\text{-}MC_{pk}$ (assuming a two-components model). HC-RMSR (High-Confidence Raw Material Specification Region) and WS (Warning Space).

Figure 7.1: Graphical interpretation of:

and the classical univariate capability index in the original space, C_{pk}, for a particular supplier.

As commented above, the C_{pk} only focuses on one raw material property where its specifications are not model-linked to the CQAs (Figure 7.1a). This index measures how much "natural variation" a process experiences relative to its specification limits, and it is a ratio of the distance between the average of the raw material property (μ) and its closer specification limit (SL) and half the "natural variation" ($\sigma z_{1-\alpha/2}$) for a certain confidence level $1-\alpha$:

$$C_{pk} = \frac{min\{\mid SL - \mu \mid\}}{\sigma z_{1-\alpha/2}} \tag{7.4}$$

where σ is the standard deviation of the raw material property, and $z_{1-\alpha/2}$ is the percentile of a standard normal distribution corresponding to a given confidence level of $1-\alpha/2$. Per contra, $LSb\text{-}MC_{pk}$ focuses on a multivariate specification region in the latent space considering not just one but all raw material properties related to CQA (Figure 7.1b). Besides, analogously to C_{pk}, $LSb\text{-}MC_{pk}$ compares two terms:

- the minimum statistical distance between the multivariate specification limit ($\boldsymbol{\tau}_{MSL}$) and the RMOS centroid (i.e., $\boldsymbol{\tau}_G$), which is calculated taking into account the estimated covariance matrix ($\mathbf{S}$) of the scores of the raw material properties of the particular supplier,
- and the statistical distance of the RMOS envelope ($c_{1-\alpha}$) given a certain confidence level ($1-\alpha$).

by using the following expression:

$$LSb\text{-}MC_{pk} = \frac{min\sqrt{\left(\boldsymbol{\tau}_{MSL} - \boldsymbol{\tau}_G\right)^{\mathrm{T}} \mathbf{S}^{-1} \left(\boldsymbol{\tau}_{MSL} - \boldsymbol{\tau}_G\right)}}{c_{1-\alpha}} \tag{7.5}$$

Considering the minimum statistical distance between $\boldsymbol{\tau}_{MSL}$ and $\boldsymbol{\tau}_G$ equals to enlarge (if $LSb\text{-}MC_{pk} > 1$) or shrink (if if $LSb\text{-}MC_{pk} < 1$) the ellipsoid defined by Equation 7.1 until it intersects with the $\boldsymbol{\tau}_{MSL}$. $LSb\text{-}MC_{pk}$ lower than 1 is generally considered a poor capability index, as it suggests that the supplier cannot consistently operate within the HC-RMSR.

7.5 Diagnosing assignable causes

PLS models provide a great capability for diagnosing assignable causes [93]. By using contribution plots [97] the underlying PLS model can be interrogated to reveal the group of regressor variables making the greatest contributions to the deviations in the SPE and/or the scores. Although these plots will not unequivocally diagnose the root causes of the deviations, they will provide a great insight to find them.

For instance, a high value of the SPE ($SPE > SPE_{lim}$) for a particular sample could indicate that it is statistically different from the samples used to build the PLS model in the sense that it contains new sources of variability that have not been captured by the model (i.e., there is a breakage in the correlation structure) [82]. Therefore, the PLS model would be unsuitable for assessing this sample. In this case, the SPE contribution plot for each sample would show the contribution of each one of the raw material properties to the respective SPE value, giving insight into what is different with these raw material samples with respect to those used in the historical data base. These could be of great help to suppliers to try to achieve a profound understanding of these deviations.

On the other hand, if it is desired to diagnose an assignable cause for a poor $LSb\text{-}MC_{pk}$ ($LSb\text{-}MC_{pk} < 1$), we propose to use the score contribution plots. For that, we consider two scenarios: i) $LSb\text{-}MC_{pk} < 0$ and ii) $0 \leq LSb\text{-}MC_{pk} < 1$. The first scenario means that centroid of the RMOS (i.e., average operating point) is outside the HC-RMSR. In that case, it is of great interest to reveal the group of raw material properties making that deviation. This can be done by calculating the contribution from the centroid of the HC-RMSR (i.e., ideal point) to the centroid of the RMOS. In the second scenario the centroid of the RMOS is inside of the HC-RMSR and, hence, on average one would expect to yield a manufactured product meeting CQA specifications. In that case, it seems to be more valuable to reveal what variability direction inside the RMOS is most likely to yield samples outside the HC-RMSR. The latter can be done by calculating the contribution from the centroid of the RMOS to the minimum statistical distance between $\boldsymbol{\tau}_{MSL}$ and $\boldsymbol{\tau}_G$ (considering $\mathbf{S}$).

7.6 Proposed methodology

In this section we illustrate a schematic methodology, by a simple example (Figure 7.2), to rank and select suppliers for a particular raw material used in a manufacturing process. In this example, it is assumed that there are three raw material properties, the focus is on the l-th CQA, a PLS model is fitted using two components, and both lower and upper specification limits for the l-th CQA are considered.

The steps of the proposed methodology follow:

1. To build a PLS model from a rich database with historical information of the several properties measured for a particular raw material along with the CQAs of the corresponding manufactured product.

2. To define the High-Confidence Raw Material Specification Region (HC-RMSR) in the latent space connecting input and output spaces where the prediction uncertainty is considered, thus, this region is expected to provide assurance of quality with a certain confidence level for the CQAs.

3. To define the supplier's Raw Material Operating Space (RMOS), a region in the latent space where the supplier's raw material samples are expected to be located, at a certain confidence level, from a number of raw material samples that respect the correlation structure from the past ($SPE < SPE_{lim}$).

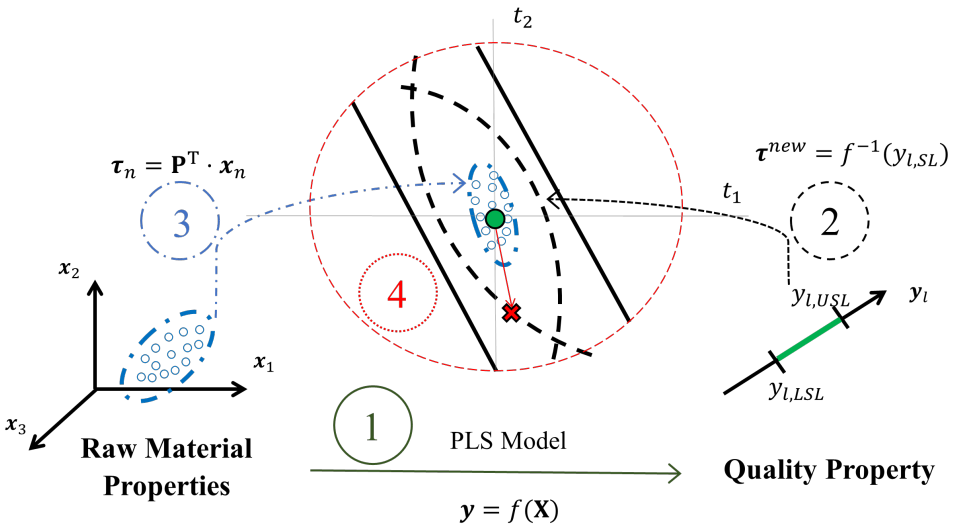


Figure 7.2: Schematic methodology of the proposal to define the $LSb\text{-}MC_{pk}$ and assess suppliers.

4. To calculate the latent space-based multivariate capability index, $LSb\text{-}MC_{pk}$, quantifying the capacity of each raw material supplier to produce a certain percentage of final product within its CQAs.

5. If it is required, to calculate the SPE and score contribution plots allowing the diagnosis of assignable causes.

7.7 Industrial case study

The present industrial case study refers to the maize cereal extraction process already presented in Section 5.5.1. To clearly illustrate this section, it is taken as a starting point the High-Confidence DS defined previously defined in Section 5.5.1 with at least 90% confidence level (Figure 5.8).

On the other hand, since historical data is mostly composed of three different suppliers, it is feasible to take advantage of the same data to assess them. First, we must check if these suppliers respect the correlation structure from the past by calculating the proportion of observations with SPE within SPE_{lim} (Table 7.1).

Table 7.1: Proportion of observation with SPE within the 99% SPE confidence limit for each supplier.

	Proportion of observation with SPE within the 99% SPE confidence limit
Supplier 1	0.93
Supplier 2	0.90
Supplier 3	0.99

From Table 7.1 is acceptable to consider that all of them respect the correlation structure from the past as most observations are within the SPE confidence limit. Then, the projections of these observations (i.e., the scores) are modelled assuming a multivariate normal distribution with a 99% confidence level as shown in Figure 7.3.

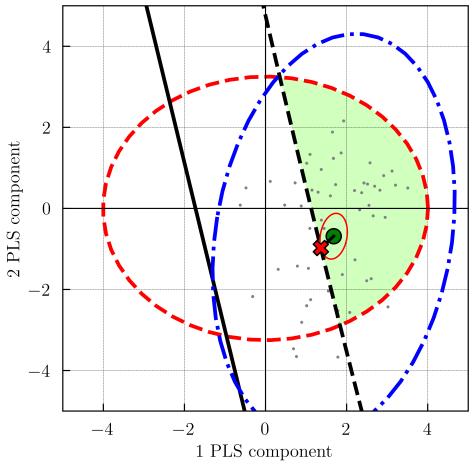
Figure 7.3c shows that the centroid of the supplier 3 is outside the HC RMSR. A priori, we should not be interested in using its raw materials. Indeed, a negative capacity index would be expected from this supplier. Conversely, the centroids of suppliers 1 and 2 are within the HC RMSR and, apparently, the Euclidean distances between the centroid of the RMOS for each supplier and the multivariate specification limit are similar. However, the RMOS is quite different when comparing suppliers 1 and 2 and, therefore, it is crucial to assess them by the $LSb\text{-}MC_{pk}$ in the latent space (Table 7.2).

Table 7.2: Multivariate capability index ($LSb\text{-}MC_{pk}$) for three suppliers.

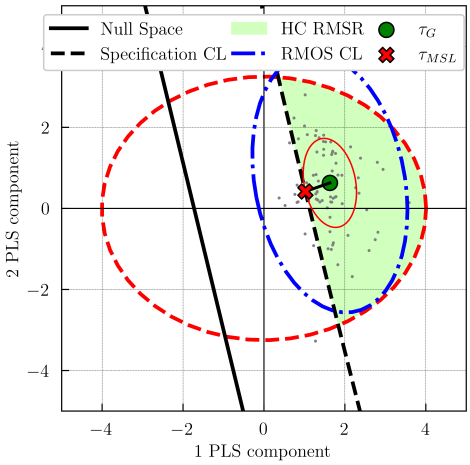
	$LSb\text{-}MC_{pk}$
Supplier 1	0.12
Supplier 2	0.35
Supplier 3	< 0

Table 7.2 shows that supplier 2 presents a higher $LSb\text{-}MC_{pk}$ than supplier 1. Thus, the expected ability to obtain a yield superior or equal to 69% is higher when accepting raw materials from the supplier 2. However, all suppliers still present a $LSb\text{-}MC_{pk}$ lower than 1. In all these cases, diagnosing assignable causes results very useful to isolate the deviating variables (Figure 7.4).

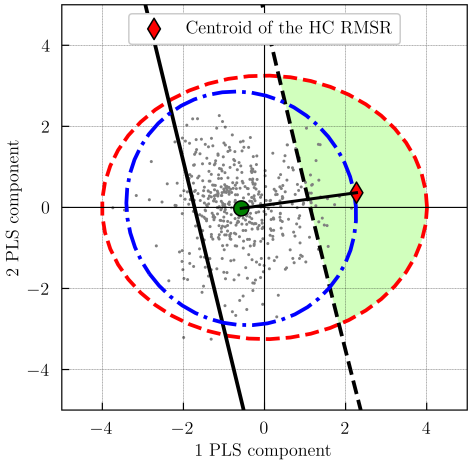
Since the Supplier 3 presents a $LSb\text{-}MC_{pk}$ lower than 0, Figure 7.4c shows the score contribution plot from the centroid of the HC RMSR (diamond in Figure 7.3c) to the centroid of the Supplier 3 RMOS (circle in Figure Fig-



(a) Supplier 1.

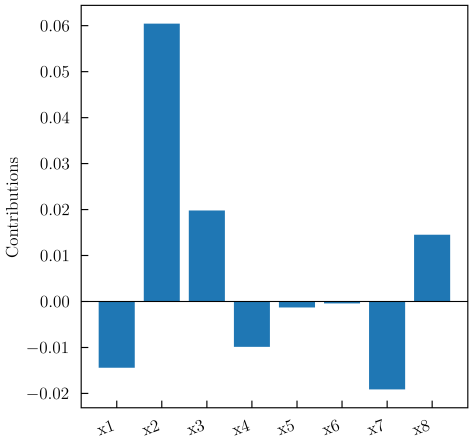


(b) Supplier 2.

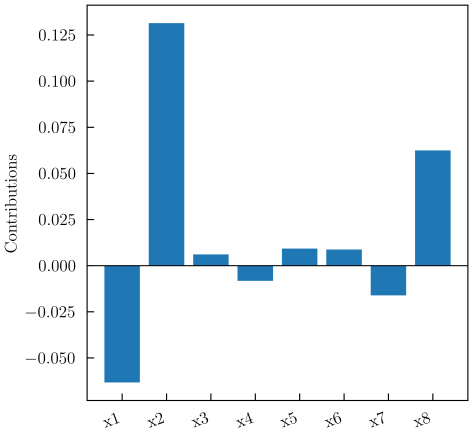


(c) Supplier 3.

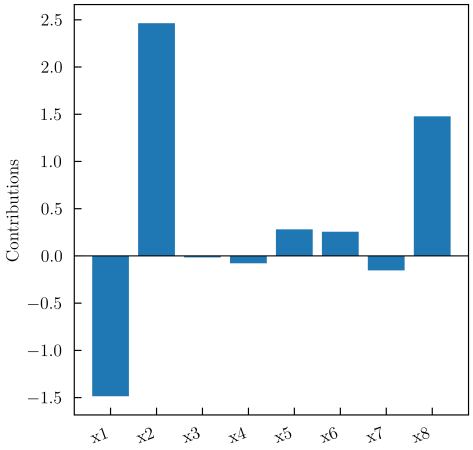
Figure 7.3: Graphical definition of the suppliers' assessing by means of the Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$).



(a) Supplier 1.



(b) Supplier 2.



(c) Supplier 3.

Figure 7.4: Score contribution plots.

ure 7.3c). Thus, it is concluded that mainly high values of the variable $\boldsymbol{x}_2$, but also slightly high values of $\boldsymbol{x}_8$ and slightly low values of $\boldsymbol{x}_1$, are responsible for the deviation from the ideal situation to the average operating space.

Conversely, Figures 7.4a and 7.4b show the score contribution plots from the centroid of the Supplier 1 and 2 RMOS (circle in Figure 6a and 6b) to the minimum statistical distance between $\boldsymbol{\tau}_{MSL}$ and $\boldsymbol{\tau}_G$ (x-cross marker in Figures 7.3a and 7.3b), respectively. In both cases, high values of the variable $\boldsymbol{x}_2$, but also slightly high values of $\boldsymbol{x}_8$ and slightly low values of $\boldsymbol{x}_1$, are responsible for the deviation in the direction most likely to yield samples outside the HC RMSR. However, for the Supplier 1 the direction is different as slightly high values of $\boldsymbol{x}_3$ and slightly low values of $\boldsymbol{x}_4$ and $\boldsymbol{x}_7$ also contribute to the deviation.

The Supplier 2 case is especially illustrative as a naive approach could have concluded that the direction of maximum variability in RMOS is the direction to consider when diagnosing assignable causes but, in fact, the orthogonal direction is responsible for the low $LSb\text{-}MC_{pk}$ despite the fact that it is the direction of less variability as can see in Figure 7.3b.

Even though we have used the same data in order to build the model and assess different suppliers, notice that it might be possible to assess new suppliers following a similar procedure.

7.8 Conclusions

In this chapter, we propose a novel Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) that arises from comparing the supplier's Raw Material Operating Space (RMOS) with the High-Confidence Raw Material Specification Region (HC-RMSR). RMOS is a region in the latent space linking the raw material properties (input space) with the Critical Quality Attributes (CQAs) of the product manufactured (output space), where the supplier's raw material samples are expected to be located at a certain confidence level. On the other hand, HC-RMSR is a region in the latent space connecting both input and output spaces associated with raw materials properties providing assurance of quality for the CQAs of the manufactured product with a certain confidence level.

All we need to calculate this novel multivariate capability index is a rich database with historical information of the several properties measured for a particular raw material along with the CQAs of the corresponding manufactured product, which is usually available in Industry 4.0.

The most remarkable advantages of the proposed $LSb\text{-}MC_{pk}$ are:

- It can be calculated with historical data (i.e., daily production data not coming from any experimental design, typical in Industry 4.0).
- It is a multivariate capability index, providing much insight into the correlated nature of raw material properties.
- It is not defined in the multivariate raw material space (as other multivariate capability indexes proposed in the literature) but in the latent space connecting the raw material properties with the Critical Quality Attributes (CQAs) of the product manufactured.
- It quantifies the ability of each supplier of a particular raw material to produce a certain percentage of final product within its CQAs specifications, and this information can be obtained at the reception of the supplier´s raw material, before producing a single unit of the product, and it can be used for ranking and selecting suppliers.
- Diagnosing assignable causes can be carried out when the samples of the supplier's raw material do not respect the correlation structure from the past (by using the SPE contribution plots), or when the supplier cannot consistently operate within the HC-RMSR (by using the score contribution plots).
- In case a supplier provides different raw materials, a $LSb\text{-}MC_{pk}$ can be calculated for each raw material of the supplier assuming a rich data base with historical information linking the several properties measured for each raw material with the CQAs of the corresponding manufactured product is available. In this case, another option would be to build a PLS model using as regressors all properties of all raw material types and calculating just one $LSb\text{-}MC_{pk}$ for each supplier. A comparison of the performance of both approaches deserves future research.

Part III

Novel applications

Chapter 8

Health application: COVID-19 Pandemic

Part of the content of this chapter has been included in:

[15] A. González-Cebrián, J. Borràs-Ferrís, J. P. Ordovás-Baines, M. Hermenegildo-Caudevilla, M. Climente-Marti, S. Tarazona, R. Vitale, D. Palací-López, J. F. Sierra-Sánchez, J. S. de la Fuente, and A. Ferrer, "Machine-learning-derived predictive score for early estimation of COVID-19 mortality risk in hospitalized patients," *PLoS ONE*, vol. 17, no. 9 September, pp. 1–17, 2022. DOI: 10.1371/journal.pone.0274171

8.1 Introduction

Since the end of 2019 to the present, one of the most contagious infections in history is going through; the pandemic produced by the SARS-CoV-2 influenza virus, named by the World Health Organization as COVID-19. This pandemic in 2020–2023 has caused to date (May 2023) more than 766 million infections and more than six million deaths worldwide, already ranking in many countries as one of the three main causes of death [98]. The world has paid a high toll in this pandemic in terms of human lives lost, economic repercussions, and increased poverty [99]. Therefore, faced with this emergency, the Ph.D. student was involved in a nationwide project sponsored by the Spanish Society of Hospital Pharmacy (SEFH). The data used in the SEFH project (SEFH data) were obtained from the RERFAR-COVID-19-SEFH Registry. It is a big repository of anonymized COVID-19 medical records admitted to Spanish hospitals. More details of the SEFH data are given in Section 8.4.2.

Within the context of SEFH project, the use of latent variable-based models was applied in both passive use and active use:

- **Passive use**: Development of statistical and machine-learning data-driven models that could be easily acquired at COVID-19 patients' admission to the hospital for the determination of their mortality risk. The findings of this work were included in Ref. [15] and are summarized in Section 8.2.
- **Active use**: Development of latent variable-based alternative to placebo-controlled clinical trials (see Section 8.4)

The proposed methods to address these novel applications with respect to their active use are presented in Section 8.3, as a reformulation of the process optimization problem in Section 2.2.3 is required.

8.2 Machine learning models for early estimation of COVID-19 mortality risk in hospitalized patients

The clinical course is extremely variable, most of the patients suffer minor symptoms, but around $10\% - 20\%$ of them require hospitalization due to the development of respiratory failure or pneumonia, requiring in some cases mechanical ventilation or admission to intensive care units, which increases the risk of death [100]. Progression to severe disease is linked to damage in the

respiratory tract and also to other organs because of the organic inflammatory syndrome related to massive cytokines release [101].

When a COVID-19 patient is admitted to the hospital, it is essential to predict the severity of the infection, both from the individual point of view and from potential health system collapses, whose prevention requires important decisions about patient management with appropriate triage criteria. The identification of involved mortality factors allows the application of targeted strategies to high-risk patients [102]. Most of the treatments that could improve the prognosis of the disease are usefully applied early, within the first days of symptoms, hence an early identification of the risk of death from COVID-19 can be critical.

In this sense, regarding the project sponsored by the SEFH, four supervised algorithmic techniques were used as classifiers that could accurately predict mortality (passive use): Logistic Regression [103], Partial Least Squares Discriminant Analysis [104], kernel-PLSDA [105], and Random Forest [106]. We used the information about these classifiers' performance metrics and about importance and coherence among the predictors to define a mortality score that can be easily calculated using a minimal number of mortality predictors and yielded accurate estimates of the patient severity status. In general, these algorithms had a similar performance, although there were differences in terms of the optimal number of variables. Indeed, in view of the findings, Random Forest was selected as the best classifier, showing slightly better results with the minimum number of predictors.

8.3 Methods: Reformulation of the optimization problem

8.3.1 PLS customized optimization problem formulation

The PLS customized optimization problem is proposed to find the customized combination of drugs (i.e., drug therapy) that maximizes the expected health of a new patient characterized by several features. Thus, the matrix of inputs, $\mathbf{X}$, refers to both patient features and drug therapy ($\mathbf{Z}$ and $\mathbf{X}$, respectively, in Section 8.3.3), and the output variable vector, $\mathbf{Y}$, refers to attributes related to patient health. This optimization problem is formulated as explained in Section 2.2.3, but modified to address the present problem. Indeed, patient features must be fixed to equality hard constraints reducing the degrees of freedom to only the set of drugs, as García-Muñoz, Dolph, and Ward [80] already proposed when defining a feed-forward controller. Thus, once a new

patient is admitted to the hospital, the optimization problem is executed in order to calculate the best drug therapy for this patient. Note that, if too many constraints are specified for patient features, the model inversion solution may be forced to move away from the latent model [23]. The optimization problem can be formulated in such a way as to take this occurrence into account, by including both soft and hard constraints for $SPE_{\boldsymbol{x}^{new}}$, namely, the mismatch of the model in representing $\boldsymbol{x}^{new}$. Differently from the previous scenario, the solution will lie outside the model space, although only slightly, as long as $SPE_{\boldsymbol{x}^{new}}$ is lower than a specified threshold (which can be represented by the historical confidence limit, SPE_{lim}). Thus, the optimization problem formulated as Equation 2.19 is modified as follows:

$$
\begin{aligned}
&\min_{\boldsymbol{x}^{new}} \left[g_0 \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau}\right)^{\mathrm{T}} \boldsymbol{\Gamma} \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau}\right) + g_1 \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} + g_2 SPE_{\boldsymbol{x}^{new}} \right] \\
&s.t. \\
&\boldsymbol{v}_0 = -\boldsymbol{y}^{des} \\
&\mathbf{V} = \mathbf{Q} \\
&\hat{\boldsymbol{y}}^{new} = \mathbf{Q}\boldsymbol{\tau} \\
&\hat{\boldsymbol{x}}^{new} = \mathbf{P}\boldsymbol{\tau} \\
&\boldsymbol{\tau} = \mathbf{W}^{*\mathrm{T}} \boldsymbol{x}^{new} \\
&SPE_{\boldsymbol{x}^{new}} = \left(\hat{\boldsymbol{x}}^{new} - \boldsymbol{x}^{new}\right)^{\mathrm{T}} \left(\hat{\boldsymbol{x}}^{new} - \boldsymbol{x}^{new}\right) \leq SPE_{lim} \\
&T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{\tau} \leq T_{lim}^2 \\
&\mathbf{A}_{\boldsymbol{\tau}} \boldsymbol{\tau} \leq \boldsymbol{d}_{\boldsymbol{\tau}} \\
&\mathbf{F}_{\boldsymbol{\tau}} \boldsymbol{\tau} = \boldsymbol{f}_{\boldsymbol{\tau}}
\end{aligned}
\tag{8.1}
$$

where g_2 is a parameter weighting the importance of the soft constraint for $SPE_{\boldsymbol{x}^{new}}$ in the objective function.

8.3.2 Nonlinear PLS customized optimization problem formulation

In many areas, such as the health area, a strong nonlinear relation between different sets of data may exist. While linear models, such as PLS, might be a good simple approximation to these problems, when nonlinearity is severe they often perform unacceptably. For that reason, Wold, Kettaneh-Wold and Skagerberg [107] introduced the concept of nonlinear PLS, where the authors

already distinguished and described two basic approaches for modeling curved relationships between sets of observed data. The first approach involves a nonlinear transformation to observed variables. Despite the ability to fit highly complex nonlinear data relationships, this approach usually has limited possibility to interpret the results with respect to the original data. On the contrary, the second approach involves a nonlinear inner relation between latent variables (LVs) which overcomes the problem of loss interpretability, but it is achieved at the expense of computational cost and optimization complexity [108]. The linear inner relation implicitly assumed between the scores vectors $\boldsymbol{t}$ and $\boldsymbol{u}$ in Equation 2.3 is replaced by a nonlinear inner relation (Equation 8.2). Finally, Equation 2.3 is replaced by Equation 8.3.

$$\boldsymbol{u}_a = g_a(\boldsymbol{t}_a) + \boldsymbol{h} = g(\mathbf{X}, \boldsymbol{w}) + \boldsymbol{h} \tag{8.2}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}^{*} \tag{8.3}$$

where the columns of the matrix $\mathbf{U}$ are the PLS output score vectors ($\boldsymbol{u}_a, a = 1, 2, 3, ..., A$), containing the first A latent variables (LVs) from PLS, g_a represents a continuous nonlinear function for the a-th LV, $\boldsymbol{h}$ denotes a vector of residuals, and $\mathbf{F}^{*}$ is the residual matrix being an indicator of how good the model is in predicting the $\mathbf{Y}$-space from $\mathbf{U}$. Note that, the use of a nonlinear model to relate the score vectors in the inner relation affects the computation of $\boldsymbol{w}$, and hence an update of the $\boldsymbol{w}$ needs to be considered. Wold, Kettaneh-Wold and Skagerberg [107] proposed to update $\boldsymbol{w}$ by means of a Newton-Rapshson-like linearization of g, and it was corrected later by Rosipal [108].

In this work, the second approach is considered in the customized optimization problem formulation for its suitable integration, as the assumption that the score vectors $\boldsymbol{t}$ and $\boldsymbol{u}$ are linear projections of the original variables is kept (in contrast to the first approach) [108]. The optimization problem is formulated as explained in Section 8.3.1, but modified to consider the nonlinear inner relation between scores.

$$
\begin{aligned}
&\min_{\boldsymbol{x}^{new}} \left[g_0 \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{v} \right)^{\mathrm{T}} \mathbf{\Gamma} \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{v} \right) + g_1 \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} + g_2 SPE_{\boldsymbol{x}^{new}} \right] \\
&s.t. \\
&\boldsymbol{v}_0 = -\boldsymbol{y}^{des} \\
&\mathbf{V} = \mathbf{Q} \\
&\hat{\boldsymbol{y}}^{new} = \mathbf{Q}\boldsymbol{\tau} \\
&\hat{\boldsymbol{x}}^{new} = \mathbf{P}\boldsymbol{\tau} \\
&\boldsymbol{\tau} = \mathbf{W}^{*\mathrm{T}} \boldsymbol{x}^{new} \\
&\upsilon_a = g_a(\tau_a) : a = 1, \cdots, A \\
&SPE_{\boldsymbol{x}^{new}} = \left(\hat{\boldsymbol{x}}^{new} - \boldsymbol{x}^{new} \right)^{\mathrm{T}} \left(\hat{\boldsymbol{x}}^{new} - \boldsymbol{x}^{new} \right) \leq SPE_{lim} \\
&T^2_{\boldsymbol{\tau}} = \boldsymbol{\tau}^{\mathrm{T}} \mathbf{\Lambda}^{-1} \boldsymbol{\tau} \leq T^2_{lim} \\
&\mathbf{A}_{\boldsymbol{\tau}} \boldsymbol{\tau} \leq \boldsymbol{d}_{\boldsymbol{\tau}} \\
&\mathbf{F}_{\boldsymbol{\tau}} \boldsymbol{\tau} = \boldsymbol{f}_{\boldsymbol{\tau}} \\
&\mathbf{A}_{\boldsymbol{\upsilon}} \boldsymbol{\upsilon} \leq \boldsymbol{d}_{\boldsymbol{\upsilon}} \\
&\mathbf{F}_{\boldsymbol{\upsilon}} \boldsymbol{\upsilon} = \boldsymbol{f}_{\boldsymbol{\upsilon}}
\end{aligned}
\tag{8.4}
$$

where $\boldsymbol{\tau}$ is the input score vector, composed by A elements, τ_a, and $\boldsymbol{\upsilon}$ is the output score vector of the solution, composed by A elements, υ_a. Besides, inequality (and equality) hard constraints of CPIs and CQAs are transferred to their respective LVs. This is feasible as the assumption that the scores are linear projections of the original variables is kept as commented. Therefore, $\mathbf{A}_{\boldsymbol{\tau}}$ and $\boldsymbol{d}_{\boldsymbol{\tau}}$ ($\mathbf{F}_{\boldsymbol{\tau}}$ and $\boldsymbol{f}_{\boldsymbol{\tau}}$) are a matrix and a vector used to define inequality (and equality) hard constraints of CPIs on the input LVs, and $\mathbf{A}_{\boldsymbol{\upsilon}}$ and $\boldsymbol{d}_{\boldsymbol{\upsilon}}$ ($\mathbf{F}_{\boldsymbol{\upsilon}}$ and $\boldsymbol{f}_{\boldsymbol{\upsilon}}$) are a matrix and a vector used to define inequality (and equality) hard constraints of CQAs on the output LVs.

Note that, now $\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{v} = \hat{\boldsymbol{y}}^{new} - \boldsymbol{y}^{des}$, and hence, this soft constraint keeps the same purpose as in Equations 2.19 and 8.1. In addition to that, it has been determined to include the soft and hard constraints on T^2 only for the input LVs, because it is desired to constrain the solution to be physically feasible and consistent with the correlation structure of inputs variables (i.e., patient features and the set of drugs) from the past.

8.3.3 SMB-PLS customized optimization problem formulation

It should be noted that the PLS customized optimization requires fixing as many equality hard constraints as patient features hindering the search for an optimal solution. Although it can be addressed by moving away from the latent model [23] as commented, it may face several difficulties:

- It requires moving from an optimization problem of the dimension of the number of LVs, A, to the dimension of the number of input variables, M. This may increase relevantly the computational cost of the optimization problem.
- Moving away from the latent model does not guarantee the existence of feasible solutions if too many constraints are considered.
- Although looking for a solution out of the model is admissible as long as $SPE_{\boldsymbol{x}^{new}}$ is lower than a specified threshold, it is preferable to search for a solution within the model latent space.

To overcome these difficulties, the SMB-PLS customized optimization problem formulation is proposed. As commented in Section 2.3, the SMB-PLS imposes a sequential pathway between the regressor blocks according to the process flowsheet, i.e., with the first block as the patient features, $\mathbf{Z}$, and the second block as the drug therapy, $\mathbf{X}$[1]. Therefore, the SMB-PLS captures the impact of variations in patient features on the drug therapy and on the patient's health in the first block of latent variables. This allows identifying the correlation structure between drug therapy and patient features from the past, which must be respected. Then, the second block of latent variables captures variations in drug therapy, that are independent from patient features and also affect health, e.g., certain excitations due to small changes in drug therapy, from which causality can be inferred. Thus, given a new patient whose features, $\boldsymbol{z}^{new}$ are known, one can estimate the expected drug therapy according to the past, $\hat{\boldsymbol{x}}^{new}_{corr}$, as in Equation 6.3, and the first block of latent variables, $\boldsymbol{\tau}_{\mathbf{T}}$, from Equations 6.3 and 6.4. Finally, the optimization problem can be formulated to search, in the second block of latent variables, for an optimal (and feasible) solution, $\boldsymbol{\tau}_{orth}$, that improves the expected patient's health with respect to the past. Thus, the SMB-PLS customized optimization problem is formulated as follows:

[1]Note that, the first and the second block refer to $\mathbf{Z}$ and $\mathbf{X}$, respectively, to keep the terminology used in Section 2.3

$$
\begin{aligned}
&\min_{\boldsymbol{\tau}_{orth}} \left[g_0 \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau} \right)^{\mathrm{T}} \boldsymbol{\Gamma} \left(\boldsymbol{v}_0 + \mathbf{V}\boldsymbol{\tau} \right) + g_1 \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} \right] \\
&s.t. \\
&\boldsymbol{v}_0 = -\boldsymbol{y}^{des} \\
&\mathbf{V} = \mathbf{Q} \\
&\hat{\boldsymbol{x}}^{new} = \hat{\boldsymbol{x}}_{corr}^{new} + \hat{\boldsymbol{x}}_{orth}^{new} = \hat{\boldsymbol{x}}_{corr}^{new} + \mathbf{P}_{orth}\boldsymbol{\tau}_{orth} \\
&\boldsymbol{\tau} = \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}} \\ \boldsymbol{\tau}_{orth} \end{bmatrix} \\
&\hat{\boldsymbol{y}}^{new} = \mathbf{Q}\boldsymbol{\tau} \\
&T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{\tau} \leq T_{lim}^2 \\
&\mathbf{A}_{\boldsymbol{\tau}}\boldsymbol{\tau} \leq \boldsymbol{d}_{\boldsymbol{\tau}} \\
&\mathbf{F}_{\boldsymbol{\tau}}\boldsymbol{\tau} = \boldsymbol{f}_{\boldsymbol{\tau}}
\end{aligned}
\tag{8.5}
$$

Note that, the optimization problem formulated in Equation 8.5 does not require fixing patient features to equality hard constraints, and hence, there is no need to move away from the latent model. The score vector, $\boldsymbol{\tau}$, will respect the observed patient features due to the fact the optimization problem reduces the degrees of freedom to only the second block of latent variables, where causality can be inferred. In addition to that, the SMB-PLS provides great insights in agreement with process knowledge for the effects of patient feature variations and the correlated drug therapy.

8.4 A Latent variable-based alternative to clinical trials upon new diseases

Many drug therapy developments consist in investigating through different clinical trials the effects of different specific drug therapies. A clinical trial refers to any form of a planned experiment that involves patients and is designed to elucidate the most appropriate treatment for future patients with a given medical condition. The essential characteristic of a clinical trial is that one uses results based on a limited sample of patients to make inferences about how the treatment should be conducted in the general population of patients who will require treatment in the future [109]. Usually, a clinical trial consists of a placebo-controlled trial: a test group receives a drug treatment and a control group receives a placebo. This is necessary to recognize the real drug effect while comparing it to a placebo.

Note that, the pursuit of an effective treatment requires an active use of a model, as it is intended to be used to actively alter/improve the patient's health. Indeed, it involves defining an optimization problem in order to obtain an effective treatment that maximizes a medical criterion. For active use, a causal model is required. When using conventional predictive methods that directly relate the registered input variables with the output variables, causality must be inferred from data obtained from clinical trials.

However, the appearance of a new dangerous and contagious disease, such as COVID-19, requires the development of a drug treatment faster than what is foreseen by usual mechanisms. In these cases, conventional clinical trials may not be feasible due to the time dilation in finding the best drug therapy [110]. On the contrary, hospitals tend to test different treatments or combinations of them for each of their patients based on their availability and potential effectiveness. This new context is yielding heterogeneous, collinear datasets with missing data, that are not the result of a conventional clinical trial. These data, as happenstance data from Industry 4.0, can not be used to infer causality in the original space.

Despite that, recently different approaches were tempted to use machine learning models to find an effective treatment using data from daily tested patients. For instance, Ezequiel et al. [110] use a complex neural network that learns from a diverse simulated dataset in which patients are tested in different drug therapies. This machine learning technique is used to find the best drug therapy for a new disease. The latter could give rise to two issues. First of all, if data comes from daily tested patients (and not from a clinical trial) surely there will be a correlation between patient features and drug therapy. However, looking for the best therapy involves suggesting a unique therapy for diverse patients. This unique therapy may not be suitable for several patients. Hence, it is crucial to design the best therapy customized for new patients considering their features. Secondly, when a machine learning technique is used to optimize or find the best drug therapy this usually involves a predictive (forward) approach. The latter requires the discretization of the input domain (i.e., set of drugs), and then obtaining the prediction for every discretization according to the predictive model with the purposed of finding the best one. Note that, these discretization points may not respect the correlation structure between the set of drugs already tested and, hence, any causal interpretation may be misleading.

On the contrary, as discussed in Section 1, latent variable-based models are of interest due to their capacity to infer causality in the reduced latent space no matter if the data come either from a clinical trial or daily tested patients.

Thus, by means of the latent variable-based optimization problem (backward approach) is feasible to find the best drug therapy in a customized way. Since the solution belongs to the latent space (defined by the latent variables), this solution is constrained to be physically feasible and consistent with the correlation structure of inputs variables (i.e., patient features and the set of drugs) from the past.

In this section, the latent variable-based optimization problem is applied to the dataset simulated according to Ref. [110] (see Section 8.4.1), and the data used in the SEFH project (see Section 8.4.2).

8.4.1 Simulated case study

8.4.1.1 Problem setup

The simulated case study allows showing, from a mathematical point of view, how latent variable-based models could be used to improve the efficacy in finding the best-customized drug therapy for a new unknown disease. The data is simulated according to Ezequiel et al. [110] where a simulated patient is composed of 5 features described with a multidimensional vector $\boldsymbol{z}$, and there are 10 drugs to be tested in any combination described with a multidimensional vector $\boldsymbol{x}$. These variables can be described with a number between 0 and 1. Moreover, it is assumed that the patient outcome can be described with a number between 0 and 1: 0 being dead and 1 being in excellent health conditions. This outcome is calculated by a health function $h(\boldsymbol{z}, \boldsymbol{x})$, which is a multi-variable non-linear unknown function that can take values in [0, 1]. In addition to that, there may be uncontrolled variables that also affect the outcome, such as a genomic factor. These uncontrolled variables could be included as a stochastic noise, S (in this work S is set to 40%).

$$\boldsymbol{y} = H(\boldsymbol{z}, \boldsymbol{x}) = h(\boldsymbol{z}, \boldsymbol{x})S \tag{8.6}$$

where $H(\boldsymbol{z}, \boldsymbol{x})$ represents the health outcome of a patient with features $\boldsymbol{z}$ when receiving a treatment consisting of drug therapy $\boldsymbol{x}$. This modeling does not rely on any physiological, biological, or molecular behavior or interaction, but instead just on the outcome of the patient as a number in [0, 1] related to the patient's health at the end of treatment. More details about how the outcome is obtained can be found in Appendix 8.A. The data available in this case are 2000 historical patients and 1000 new patients, both simulated according to 8.6 for uniformly and independently distributed $\boldsymbol{z}$ and $\boldsymbol{x}$.

This case study aimed to build a latent variable-based model that, provided the features for new patients, could be used to find the customized drug therapy that maximizes the distribution $H(\boldsymbol{z}, \boldsymbol{x})$ over them.

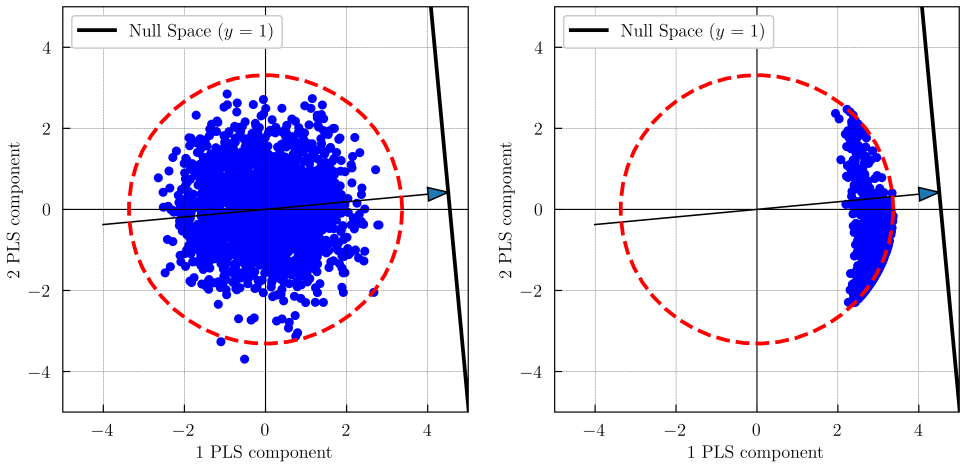
8.4.1.2 Results

PLS customized

First of all, the PLS customized optimization problem formulated in Section 8.3.1 is proposed. For that, we train a PLS model with the historical patients. Then, we apply the PLS optimization problem independently for each new patient yielding the distribution of $H(\boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{z}$. For that, y^{des} is set to 1, and inequality hard constraints are defined as the result of transferring the historical restrictions on the drugs (i.e., [0, 1]) to the latent space. Besides, given a new patient, its features are fixed by equality hard constraints. g_0 is set to 1, g_1 is set to 0, and g_2 is set to 0.15. Note that, in this problem, g_1 is neglected in order to ease the solution to reach the hard constraints in an attempt to comprehend the performance of the models in this simulated case study. Moreover, g_2 has a low weight to bring the solution closer to the model, but not so much as g_0. Anyway, in both cases the confidence limits, T^2_{lim} and SPE_{lim}, are also accounted as hard constraints.

Leave-one-out Cross-Validation (CV) was used for selecting the number of PLS components. Thus, three LVs were chosen to fit a PLS model ($Q^2_{\mathbf{Y}_{cum}} = 76.38\%$ (goodness of prediction)) for the 2000 historical patients. None of the historical patients exhibit any unusual behavior caused by patient features and drugs based on SPE and T^2 charts (charts not shown). Figure 8.1 shows the score plot by showing the calibration data (Figure 8.1a), and the optimal solution of new patients calculated by means of the PLS customized optimization approach (Figure 8.1b). In both cases, it is shown the null space (NS) with respect to $y = 1$, and an arrow stating the maximum variability direction of the health response.

What stands out in Figure 8.1b with respect to Figure 8.1a is that the optimal solutions for the new patients have moved in order to try to reach the NS of $y = 1$. However, these solutions are not able to attain such NS as they encounter the T^2_{lim} hard constraint, which is illustrated as a Hotelling T^2 confidence hyperellipsoid. In some cases, they do not reach this confidence limit because several equality and inequality hard constraints prevent them to move "freely" along the latent space.



(a) calibration data. **(b)** the optimal solution of new patients.

Figure 8.1: PLS customized optimization approach: Score plot by showing

To assess the performance of the PLS customized optimization approach, Figure 8.2 shows the health distribution of new patients, according to the true model (Equation 8.6), using the proposed PLS optimal solutions (H PLS). Besides, it is compared to:

- Prediction of the health distribution of new patients according to the PLS model using the proposed PLS optimal solutions (PLS prediction)
- Neural Network Drug Theraphy technique (hereinafter NN) [110]: Train a Neural Network whit the 2000 historical patients to make it learn $H(\boldsymbol{z}, \boldsymbol{x})$. Then, a large set of pseudo-patients and drug therapies from the discretization of the input domain is simulated. The therapy that yields the maximum average for the trained Neural Network output on $(\boldsymbol{z}, \boldsymbol{x})$, in a predictive (forward) approach, is considered the best therapy. This drug therapy is defined as $\boldsymbol{x}_B$, and we estimate the distribution of $H(\boldsymbol{z}, \boldsymbol{x}_B)$ over $\boldsymbol{z}$ from the 1000 new patients (H NN).
- Placebo treatment (hereinafter Placebo): We estimate the distribution of $H(\boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{x}$ from the 1000 new patients assuming $\boldsymbol{x} = \boldsymbol{0}$ (H Placebo).

Figure 8.2 shows that both approaches PLS customized optimization and NN yield similar health distributions of new patients, H-PLS and H-NN, respectively, being substantially superior that the Placebo performance (H Placebo).

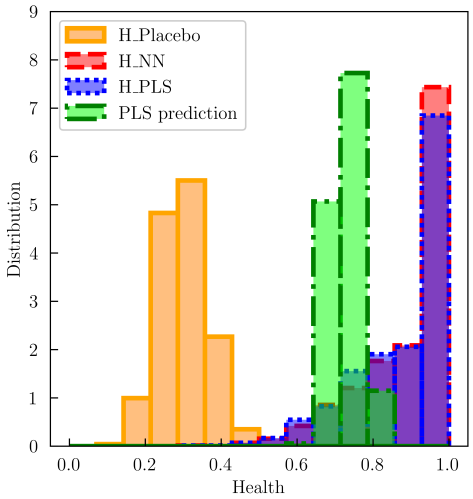
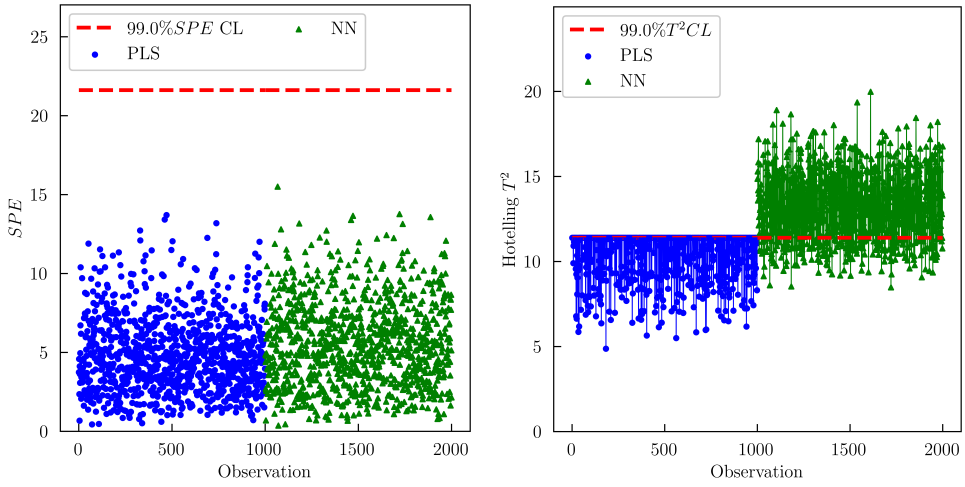


Figure 8.2: PLS customized optimization approach: Health distribution of new patients.

To comprehend the optimal solutions given by both approaches, Figure 8.3 shows the monitoring charts, SPE and T^2 of the new patients, using the PLS model for the optimal solutions of both approaches.

Figure 8.3a shows that in this simulated case, both approaches provide similar SPE values under the SPE_{lim} (feasible solutions). Note that, for the PLS customized optimization approach, this is expected because the SPE_{lim} hard constraint is taken into account in the optimization formulation. By contrast, the NN approach provides feasible solutions due to the fact that patient features and drug therapy were simulated uniformly and independently and, hence any drug therapy for any new patient is feasible. In a more realistic case, it may not happen and, consequently, the NN optimal solution would not guarantee its feasibility. In these cases, as commented, it is crucial to design the best therapy customized for new patients considering their features, as there could be a correlation between patient features and drug therapy. In addition to that, the integration of latent variable models is strongly recommended. Thus, the input space would be projected into a low-dimensional space, that defines the feasible space for simulating pseudo-drug therapies for a given patient. Then, the predictive (forward) approach would be applied to find the optimal solution for such patient.

In Figure 8.3b there is a clear difference between the PLS customized optimization and the NN approaches. Indeed, the PLS approach provides optimal



(a) SPE. **(b)** T^2.

Figure 8.3: Monitoring charts of new patients.

solutions under the T^2_{lim} as expected, however, the NN approach apparently considers an extreme drug therapy that, combined with each new patient, usually gives T^2 values higher than T^2_{lim}. Note that, when discretizing the input space in the NN approach, any combination is possible, but in the PLS approach, extreme combinations are not viable due to the T^2_{lim} hard constraint. This would explain why the NN approach yields slightly better results than the PLS approach (Figure 8.2).

A closer inspection of Figure 8.2 shows that the PLS prediction is not accurate. Indeed, Figure 8.1b already showed that none of the optimal solutions reaches the NS of $y = 1$, but they yield a healthy distribution, according to the true model, tilted towards 1 (Figure 8.2). To perceive this discrepancy, the inner relation in the first latent variable is quite revealing (Figure 8.4).

From Figure 8.4, it can be seen that the inner relation in the first latent variable involves a strong nonlinear relationship, but a linear fit is assumed in the PLS model building. However, since this inner relation is a monotonic increasing relationship, the PLS model is able to figure out the maximum variability direction with respect to health response, resulting in effective drug therapies, but it does not result in good predictions. Therefore, the PLS customized optimization problem is formulated as Section 8.3.2, assuming an exponential inner relation in the latent variables.

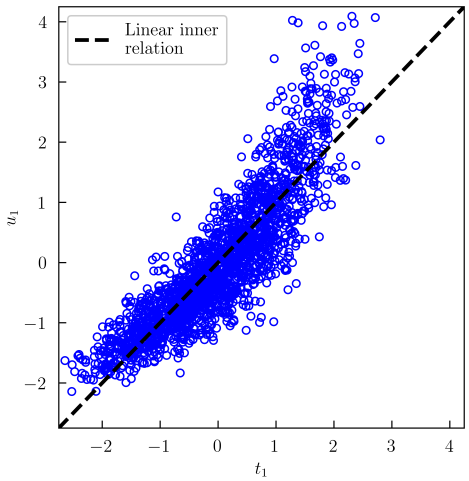
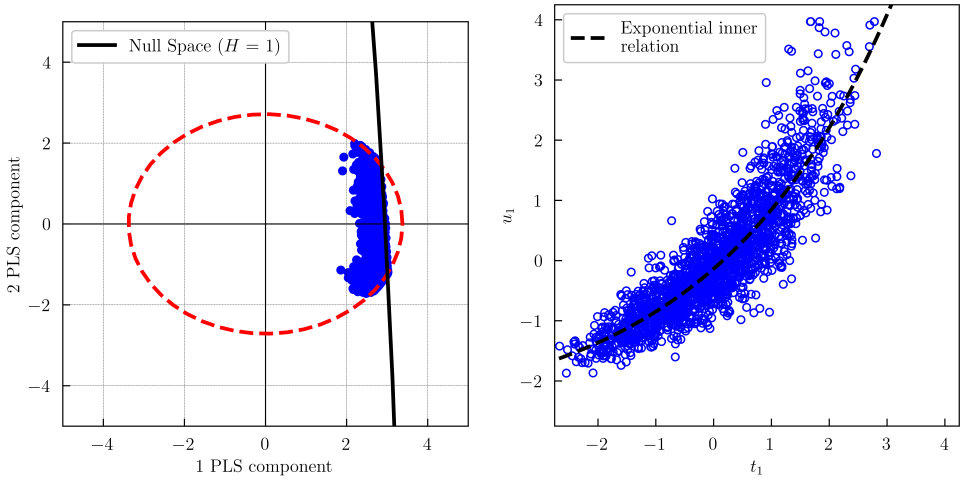


Figure 8.4: PLS customized optimization approach: Inner relation in the first latent variable assuming a linear fit by showing calibration data.

Thus, three LVs were chosen to fit a nonlinear PLS model ($Q^2_{\mathbf{Y}_{cum}} = 79.47\%$) using the 2000 historical patients, resulting in slightly better goodness of prediction that the liner PLS ($Q^2_{\mathbf{Y}_{cum}} = 76.38\%$). Besides, none of the historical patients exhibit any unusual behavior caused by patient features and drugs based on SPE and T^2 charts (charts not shown). Then, the PLS optimization problem is applied independently for each new patient considering the same optimization problem parameters as PLS yielding the distribution of $H(\boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{z}$. Figure 8.5a shows the score plot by showing the optimal solution of new patients by means of the nonlinear PLS customize optimization approach, and Figure 8.5b shows the inner relation in the first latent variable.

Figure 8.5b shows that the exponential inner relation is more suitable in the attempt to model the nonlinear relationship. Besides, the NS of $y = 1$ could be projected as a linear NS of the output latent variables, but it is represented as a nonlinear space in the score plot of inputs variables as can see in Figure 8.5a. In addition to that, Figure 8.5a shows that the NS itself is feasible because part of this space is within the Hotelling T^2 confidence hyperellipsoid, and hence, the optimal solutions reach this NS.

To assess the performance of the nonlinear PLS customized optimization approach, Figure 8.6 shows the health distribution of new patients, according to the true model, using the proposed nonlinear PLS optimal solutions (H NL



(a) Score plot by showing the optimal solution of new patients

(b) Inner relation between the first latent variable assuming an exponential fit by showing calibration data.

Figure 8.5: Nonlinear PLS customized optimization approach:

PLS). Besides, it is compared to Placebo, the nonlinear model prediction (NL PLS prediction), and H-PLS.

The most interesting aspect of Figure 8.6 is that the prediction accuracy is markedly better than the PLS approach in Figure 8.2, but it is not yet ideal. This may be improved by using other non-linear inner relations.

On the other hand, the nonlinear PLS approach yields slightly worse results than the linear PLS approach. This appeared to be due to the slight lack of accuracy in predicting the output first latent variable from high values of the first input latent variable as can see in Figure 8.5b. Notice that, high values of the first input latent variable are precisely those associated with the optimal values (Figure 8.5a). In addition to that, unlike the linear PLS approach, the optimal solutions in the nonlinear PLS approach do reach the NS as commented, and hence, they do not keep moving along the maximum variability direction of the health until encountering the T^2_{lim} hard constraint.

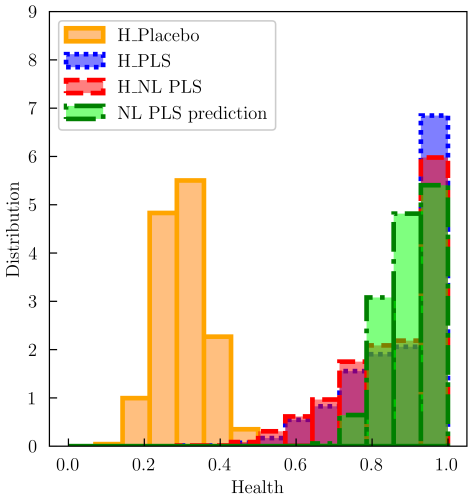
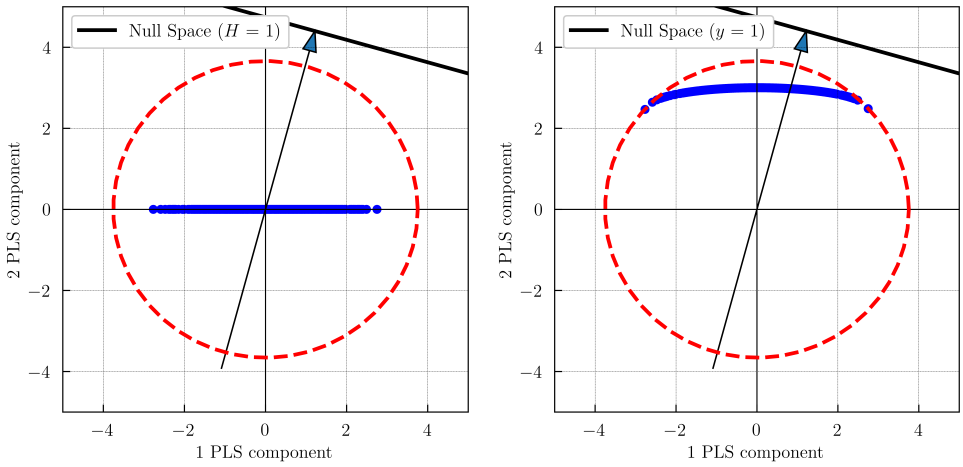


Figure 8.6: Nonlinear PLS customized optimization approach: Health distribution of new patients.

SMB-PLS customized

The SMB-PLS customized optimization problem formulated in Section 8.3.3 instead of PLS is illustrated for comparison. First, the SMB-PLS model is built with historical patients where the first block corresponds to patient features, and the second one to drug therapy. Then, the PLS optimization problem is applied independently for each new patient considering the same optimization problem parameters as PLS yielding the distribution of $H(\boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{z}$.

Thus, one component was found sufficient to capture the impact of patient features (and correlated drug therapy) on $\boldsymbol{y}$ in the first modelling step. This component explains 5.69% of the information in $\mathbf{Z}$ and 0.05% of the information in $\mathbf{X}$ that was correlated with $\mathbf{Z}$, to explain a small percentage of the response variability (5.69%). It is deduced that there is hardly variation in $\mathbf{X}$, correlated with $\mathbf{Z}$, explaining the response variability, as $\mathbf{Z}$ and $\mathbf{X}$ were simulated independently. Three additional components were also needed in the second modelling step to model the effect of orthogonal variations in drug therapy on the remaining variations in $\boldsymbol{y}$. These components show that the 10% of the variation in $\mathbf{X}$, not related to $\mathbf{Z}$, are able to explain 70.96% of the response variability. This block enables one to figure out in advance if there is any chance of searching for optimal drug therapy. In fact, if there is no variation in $\mathbf{X}$, not related to $\mathbf{Z}$, able to explain the response variability,



(a) the expected solution of new patients (no optimization).

(b) the optimal solution of new patients.

Figure 8.7: SMB-PLS customized optimization approach: Score plot by showing

it implies that this dataset does not present certain excitations due to small changes in drug therapy, from which causality can be inferred. To sum up, four LVs were chosen to fit the SMB-PLS model ($Q^2_{\mathbf{Y}_{cum}} = 76.29\%$) using the 2000 historical patients. This goodness of prediction is similar to the linear PLS model ($Q^2_{\mathbf{Y}_{cum}} = 76.38\%$), but the SMB-PLS latent variables better sort the contribution of both patient features and drug therapy on the health variations. None of the historical patients exhibit any unusual behavior caused by patient features and therapies based on SPE and T^2 charts (charts not shown).

Figure 8.7a shows the score plot by showing the expected solution of new patients, namely, it refers to the first latent variable composed of the patient features and their expected drug therapy due to the correlation structure from the past. Then, one could move the next three latent variables (i.e., orthogonal block) in order to improve the drug therapy with respect to the expected one. Thus, Figure 8.7b shows the optimal solution of new patients by means of the SMB-PLS customized optimization approach.

Figure 8.7b shows that, with respect to Figure 8.7a, the optimal solutions move along the second latent variable (also third and four but not shown) in order to reach the NS of $y = 1$, but they are restricted by the Hotelling T^2 confidence hyperellipsoid. Unlike Figure 8.1b, the optimal solutions shown in

Figure 8.7b locate on the Hotelling T^2 confidence limit[2], because there are no equality hard constraints that prevent the optimal solution to move "freely" along the orthogonal space. Only there are inequality hard constraints for historical restrictions on the drugs, that apparently do not intervene.

To assess the performance of the SMB-PLS customized optimization approach, Figure 8.8 shows the health distribution of new patients, according to the true model, using both the proposed SMB-PLS optimal solutions (H-SMB) and the expected solution of new patients with no optimization (H-SMB (no opt.)). Besides, it is compared to Placebo, the nonlinear model prediction (SMB-PLS prediction), and H-PLS.

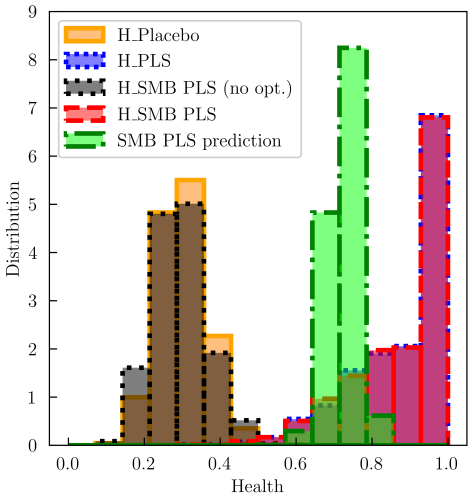


Figure 8.8: SMB-PLS customized optimization approach: Health distribution of new patients.

As shown in Figure 8.8, H-SMB (no opt.) results in practically the same health distribution as the Placebo. As commented, there is hardly variation in drug therapy, correlated with patient features, explaining the response variability. Therefore, the expected drug therapy for a new patient is, indeed, the Placebo. Then, since a linear inner relation is assumed in the SMB-PLS model, the SMB-PLS prediction is not accurate. Hence, as in PLS, a different inner relation could be used with the intention of improving the prediction.

[2]Some of the optimal solutions apparently do not locate on the T^2 confidence limit, when showing the first and second latent variables, due to the fact that the third and fourth latent variables are not shown. Indeed, all these optimal solutions have a T^2 equal to T^2_{lim} (chart not shown)

On the other hand, Figure 8.8 shows that, for this simulated case, there are no health distribution differences between both the H-SMB and H-PLS. Regarding the computational time and SPE values of the proposed drug therapy of new patients, Figure 8.9 shows multiple boxplots comparing these metrics by means of both approaches.

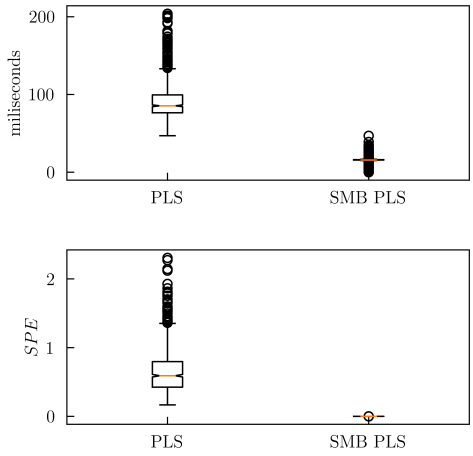


Figure 8.9: Multiple boxplots of computational time and SPE values of the proposed drug therapy by means of PLS and SMB-PLS customized optimization approaches.

What is interesting about the comparison in Figure 8.9 is that SMB-PLS approach requires much less computation time to find the optimal solution than the PLS approach. This finding can be attributed to the dimension and the number of equality hard constraints of the optimization problem. In fact, the PLS approach searches for a solution in a 15-dimensional space with 5 equality hard constraints, by contrast, the SMB-PLS approach searches for a solution in a 3-dimensional space without equality hard constraints. In a more realistic case, the number of patient features could be much higher resulting in unfeasible computational times and it may hinder the convergence of the optimization problem. Regarding the SPE values, note that the contribution of this value comes from two parts: patient features and drug therapy. The first part is given by the patient, hence it is common in both approaches. However, the second part is only present in the PLS approach as it allows the solution to move away from the model, by contrast, the SMB-PLS approach searches for the optimal solution in the second latent block itself. The contribution of this second part is shown in Figure 8.9.

8.4.2 Spanish society of hospital pharmacy case study

8.4.2.1 Problem setup

The data used in this case study were obtained from the RERFAR-COVID 19 SEFH Registry. It s a big repository of anonymized COVID-19 medical records of 15.628 patients admitted to Spanish hospitals from March 20th to July 15th, 2020. All registered patients were diagnosed using SARS-CoV-2 testing at the time of admission. The primary outcome was all-cause mortality, codified as the binary variable $\boldsymbol{y}$ with levels "alive" (numerically as zero) or "deceased" (numerically as one"). The follow-up censoring date was July 15th, 2020, and hence, if a patient had not reached the outcome (death) by the time the data were obtained, their outcome was considered null. The input data can be structured in two blocks, Z and X. The first block involves 22 variables that correspond to information available at admission (i.e., patient features), such as clinical conditions, medical records, demographic variables, chronic medications, etc. In addition to that, the second block involves 16 variables that correspond to the therapies used during admission, mainly, pharmacotherapy. The second block variables are codified as binary variables with the levels "not used therapy" (numerically as zero) or "used therapy" (numerically as one). The initial dataset ($n = 15628$ with 2846 deceased individuals) was preprocessed to obtain a clean dataset ($n = 12427$ with 2019 deceased individuals). This process eliminated observations with excessive missing values or errors in the data. The data were divided randomly into two datasets: the calibration dataset (80%) and the validation dataset (20%).

This case study intended to build a latent variable-based model that, provided the information available at the hospital admission, could predict the mortality risk of a patient with COVID-19 assuming that the therapies used during admission will respect the correlation structure from the past. Then, the same model could be used to provide an optimal therapy by means of the latent variable-based customized optimization approach.

Given the nature of the dataset, the PLS customized optimization may result in convergence problems and elevated computational times due to the fact that both the 22 information variables yield 22 equality hard constraints in the optimization problem formulation, and the optimization problem is framed in the 38-dimensional original space. By contrast, as commented, the SMB-PLS does not require fixing patient features to equality hard constraints, and the dimensionality of the optimization problem dimensional will depend on dimensional orthogonal latent space. In addition to that, contrary to the simulated case

study in Section 8.4.1, this case study involves a binary outcome. Hence, it is required a discriminant analysis method that classifies groups of individuals based on their input variables [111]. For all this, Sequential Multi-block Partial Least Squares Discriminant Analysis (SMB-PLSDA) is suggested in this case study.

8.4.2.2 Results

First of all, the SMB-PLSDA model is built for the calibration data. 12-groups Cross-Validation (CV) was used for selecting the number of PLS components. Since the goal of this discriminant model is to classify an individual between "alive" and "deceased", the goodness of prediction, Q^2, may be not the best criterion. In this case, the number of misclassifications (MC) is preferable as a criterion of selection of the latent variables number. Thus, four components were found sufficient to capture the impact of patient features (and therapies) on the classification of $\boldsymbol{y}$ in the first modelling step ($MC = 21.11\%$). Two additional components were also needed in the second modelling step to model the effect of orthogonal variations in therapies on the remaining variations in $\boldsymbol{y}$ ($MC = 17.53\%$). None of the historical patients exhibit any unusual behavior caused by patient features and therapies based on SPE and T^2 charts (charts not shown).

Figure 8.10 presents the goodness of fit, $R^2_{\mathbf{Y}_{cum}}$ (i.e., variability percentage explained by the model) for each one of the input blocks, $\mathbf{Z}$ and $\mathbf{X}$, and the output block, $\boldsymbol{y}$, and each component.

Figure 8.10 shows that the first four components of the first modelling stage explain 30.16% of the information in $\mathbf{Z}$ and 3.29% of the information in $\mathbf{X}$ that was correlated with $\mathbf{Z}$, to explain a relevant percentage of the response variability (26.63%). Given the features of a new patient, the information captured in this block will allow predicting the mortality risk of a patient with COVID-19 assuming that the therapies used during admission will respect the correlation structure from the past. In this case study, only a little amount of therapy variation is related to the patient features, however, it is crucial to capture it in this block to avoid misinterpretation of causality when searching for optimal therapies for a new patient in the second block. Then, the next two components of the second modelling step show that the 24.08% of the variation in $\mathbf{X}$, not related to $\mathbf{Z}$, is able to explain 5.81% of the response variability. The potential of this block is to take advantage of the information captured, from which causality can be inferred, to provide the optimal therapies for a new

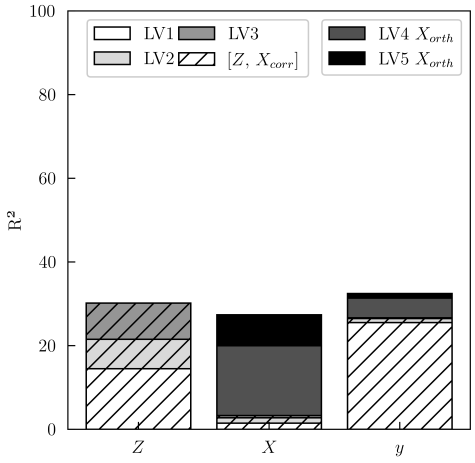


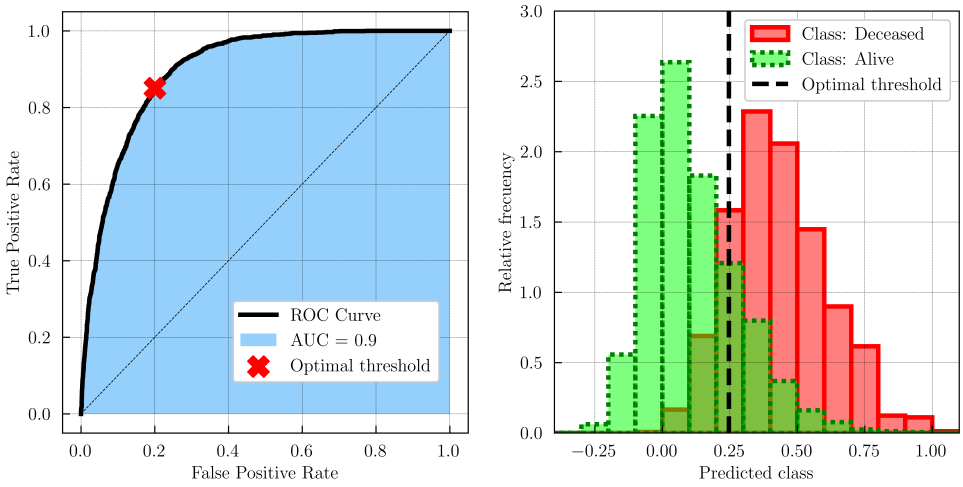
Figure 8.10: Explained **Z**, **X** and $\boldsymbol{y}$ variability for the SMB-PLS model depending on either the number of latent variables (LVS) or the two blocks of latent variables (LV1-LV3 explain the first block $[\mathbf{ZX}_{corr}]$, and LV4-LV5 explains the second block $\mathbf{X}_{orth}$).

patient. Although it only explains a 5.81% of the response variability, in some cases, it could be vital.

To assess the discriminant model, Figure 8.11 shows the Operating Characteristic Curve (ROC curve) (Figure 8.11a), and the prediction risk distribution of the calibration dataset divided into: "alive" and "deceased" (Figure 8.11b).

The ROC curve is devised as a graphical means to explore the trade-offs between two metrics at various decision thresholds when a particular quantitative variable, $\boldsymbol{y}$, is used to guide the decision [112]. In discriminant analysis, the True Positive Rate (TPR) and the False Positive Rate (FPR) are usually used to make the decision. An ideal discriminant model would have a ROC curve that passed through TPR = 1 and FPR = 0, by contrast, a useless discriminant model would correspond to the dash-line drawn through the diagonal of the ROC axes as can see in Figure 8.11a. To assess the discriminant power of the model the Area Under the ROC curve (AUC) is frequently used. In this case study, the SMB-PLSDA has an AUC of 0.90 (see Figure 8.11a).

The expected cost of operating at the various possible decision thresholds is used as a criterion to select the optimal threshold. In this case study, a default optimal threshold is calculated that balances the cost of having FPR and TPR alike. This criterion is used for being objective, even though it may not be



(a) Operating Characteristic Curve (ROC curve) of the calibration set. AUC: Area Under Curve. **(b)** Prediction risk distribution of the calibration set depending on classes.

Figure 8.11

realistic in the current context. Thus, the optimal default is shown in Figure 8.11a. In fact, Figure 8.11b shows how at the optimal threshold more of a balance is struck, as both positive and negative events are missed.

Regarding the validation step, the built SMB-PLSDA model can be used to classify the validation dataset using only the information available at the hospital admission, i.e., patient features. This is feasible as the SMB-PLSDA considers the first block of latent variables to estimate the expected therapy treatment from the patient features, and then, to predict the risk value. Thus, the risk prediction curve (Figure 8.12) for the validation dataset can be obtained.

The validation risk prediction curve, shown in Figure 8.12, is created by binning predicted probabilities, then plotting the rounded predicted probability in each bin against the observed frequency (observed mortality %). An ideal calibration curve would be located in the same diagonal. By contrast, curves located in regions aside from the diagonal would indicate an underestimation of the mortality risk (leading to under-treatment) or an overestimation (leading to over-treatment). Figure 8.12 shows that the validation risk prediction curve appears to be close to the diagonal and also well-balanced regarding possible issues of underestimation and overestimation.

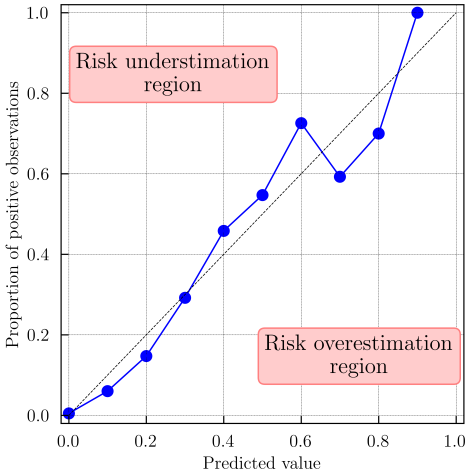


Figure 8.12: Validation risk prediction curves. Predicted risk values were rounded to the first decimal digit, e.g., predicted value 0.1 refers to predictions between 0.05 and 0.15.

Finally, as commented, the second block of latent variables can be used to search optimal therapies with respect to the past. For that, the proposed SMB-PLS customized optimization problem formulated in Section 8.3.3 is used, where y^{des} is set to 0, and inequality hard constraints are defined as the result of transferring the historical restrictions on the therapies (i.e., [0, 1]) to the latent space. g_0 is set to 1, g_1 is set to 0.25. The confidence limit, T^2_{lim}, is also accounted as a hard constraint. In addition to that, the optimization problem is applied only if the expected risk value according to the patient features is higher than 0, otherwise, it would end up searching for worse therapies. Figure 8.13 compares the prediction risk distribution for the SMB-PLSDA expected outcome only using the patient features (expected), and the SMB-PLSDA customized optimization approach (optimal).

It can be seen from the histograms in Figure 8.13 that the right tail of the optimal histogram shows a slight shift toward values closer to 0, with respect to the expected histogram. These results suggest that patients, with high prediction risk values, could slightly improve their prediction risk with the optimal drug therapy proposed by the SMB-PLSDA customized optimization approach. Only slight improvements are achievable since only 5.81% of the response variability can be inferred as the effect of 24.08% of the variation in $\mathbf{X}$ not related to $\mathbf{Z}$.

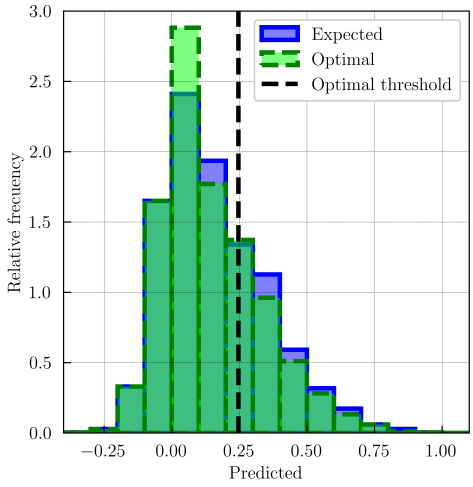


Figure 8.13: Prediction risk distribution of the validation set for the SMB-PLSDA expected outcome only using the patient features (expected), and the SMB-PLSDA customized optimization approach (optimal).

Note that, the optimal therapies are constrained to be between their historical restrictions (i.e., [0, 1]), but until now, any therapy could present values between 0 and 1, despite the fact that they are binary variables. In this case, a basic approach is proposed to address this particular problem. This involves rounding the therapies variables, after the optimization problem, to 0 or 1 based on, for instance, a naive threshold of 0.5. Then, the therapy will be considered feasible only if the new SPE value is under the SPE_{lim}. An alternative approach would be to formulate the optimization problem as an integer programming (or mixed-integer programming problem if only some therapies are not continuous), instead of continuous optimization. Then, the so-called exhaustive integer optimization could be applied, which involves evaluating all possible combinations of the binary values of the therapies within the specified domain. Although exhaustive integer optimization is guaranteed to find the optimal solution, it can be computationally expensive and practically unfeasible in problems with a large number of drugs. This option may require moving away from the latent model for both PLS and SMB-PLS optimization problems. The benefits and drawbacks of these approaches deserve further research.

Additionally, it is important to bear in mind that the optimal solutions proposed in this case study are not validated by empirical verification since this

would require using the proposed therapy in a real patient and measuring how well the proposed therapy performs. This is an important issue for future research.

8.5 Conclusion

The most clinically relevant finding is that, despite having data from daily tested patients (instead of classical clinical trials), one could use this available information to infer causality in the latent space. This may allow providing improvements over the past with respect to the use of therapies. In this sense, the latent variable-based models are of interest, particularly the SMB-PLS model, as it identifies the variation in therapies uncorrelated with patient features, which is convenient to formulate an effective optimization problem.

Finally, the proposed customized method is a valuable tool for the evaluation and optimization of the patient's therapy plan, but it should not be understood as a mandatory protocol, but rather as a support tool for the medical team in the patient care process.

Appendices

8.A Function: H

$H(\boldsymbol{z}, \boldsymbol{x})$ is defined as H_2 in Ref. [110]. This function has a compact form as follows:

$$H = \frac{1}{15} \mid z_1 + (x_1 + 3x_2 - x_3)(x_5 - x_3) + \sinh(x_7 - x_6) - 5e^{x_9 - x_3} \mid S \quad (8.A.1)$$

where H is a multi-variable (highly) non-linear function, it takes values in $[0, 1]$, and it has a stochastic component, S. $H(\boldsymbol{x}, \boldsymbol{0})$ represent the no-drugs expected outcome.

Chapter 9

Industrial application: Multivariate Six Sigma

Part of the content of this chapter has been included in:

[13] D. Palací-López, J. Borràs-Ferrís, and L. Thaise da Silva de Oliveria, "Multivariate Six Sigma: A Case Study in Industry 4.0," *Processes*, vol. 8, pp. 1–20, 2020. DOI: doi:10.3390/pr8091119

9.1 Introduction

Six Sigma is a strategy for process improvement widely used in various sectors such as manufacturing, finance, healthcare, and so on. It is defined by Linderman et al. [113] as "an organized and systematic method for strategic process improvement and new product and service development that relies on statistical methods and the scientific method to make dramatic reductions in customer defined defect rates". Moreover, Six Sigma, as a quality tool, has fostered a never-ending improvement culture based on a strong and professionalized organization for improvement, a clear and well thought methodology (DMAIC), and also powerful tools and statistical techniques to carry out the improvement projects within the DMAIC framework that has proved highly effective in a large variety of situations.

The DMAIC methodology in Six Sigma is a five-step improvement cycle, i.e., Define, Measure, Analyze, Improve, and Control. Reliable data and objective measurements are critical at each step of the method and, hence the statistical techniques are incorporated into the structured method as needed [113]. Traditionally, classical statistical techniques (e.g., Multiple Linear Regression (MLR)) have been used within the DMAIC framework in a data-scarce context mainly from experimental designs. However, with the emergence of Industry 4.0 and the Big Data movement gaining momentum, data abounds now more than ever, and the speed at which they accumulate is accelerating [3]. For all this, the Six Sigma statistical toolkit traditionally focused on classical statistical techniques must incorporate new approaches to be able to handle complex data characteristics from this current Industry 4.0 context. In such context, latent variable-based multivariate statistical techniques are widely recommended as commented in Section 1.1. In the literature, there are some examples of this integration of multivariate statistical tools into the Six Sigma toolkit [13, 114, 115].

This chapter reinforces conclusions from previous works in the literature on how Six Sigma's DMAIC methodology can be used to achieve competitive advantages, efficient decision-making, and problem-solving capabilities within the Industry 4.0 context, by incorporating latent variable-based techniques such as Partial Least Squares (PLS) into the statistical toolkit leading to the so-called Multivariate Six Sigma [2].

9.2 Methods

9.2.1 Six Sigma's DMAIC methodology

The DMAIC (Define-Measure-Analyze-Improve-Control) method in Six Sigma is often described as an approach for problem-solving. In this section, a rational reconstruction of the DMAIC methodology is shown [116].

- **Define**: problem selection and benefit analysis.
- **Measure**: translation of the problem into a measurable form, and measurement of the current situation.
- **Analyze**: identification of influence factors and causes that determine the Critical Quality Attributes (CQAs) behavior.
- **Improve**: design and implementation of adjustments to the process to improve the performance of the CQAs.
- **Control**: empirical verification of the project's results and adjustment of the process management and control system in order that improvements are sustainable.

9.2.2 Optimization of the latent space-based multivariate capability index

The purpose of this section is to provide a Sequential Multiblock (SMB) PLS optimization problem to figure out how to manipulate the process in order to maximize the Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$), namely, the ability of a particular supplier of a particular raw material to produce a certain percentage of final product within its CQAs specifications. For that, it is required to merge the novel applications proposed in Section 6 and Section 7.

In Section 6.5.1, the High-Confidence Multivariate Raw Material Specification Region (HC RMSR) was defined depending on the vector of scores referring to orthogonal variations in process conditions, $\boldsymbol{\tau}_{orth}^{new}$. Moreover, the supplier's Raw Material Operating Space (RMOS) was defined in Section 7.3 as the space where the supplier's raw material samples are expected to be located in the latent space of the PLS model. In the SMB-PLS model, the RMOS can be defined analogously from the first block of latent variables, and in function of the orthogonal variations in process conditions, $\boldsymbol{\tau}_{orth}^{new}$. Therefore, by comparing

the RMOS and HC RMSR, the $LSb\text{-}MC_{pk}$ could be defined as in Section 7.4, but in function of $\boldsymbol{\tau}_{orth}^{new}$ (i.e., $LSb\text{-}MC_{pk} = f(\boldsymbol{\tau}_{orth}^{new})$. The latter provides an opportunity to optimize the $LSb\text{-}MC_{pk}$ from $\boldsymbol{\tau}_{orth}^{new}$, i.e., For that, the following optimization problem is formulated (Equation 9.1).

$$\begin{aligned} &\max_{\boldsymbol{\tau}_{orth}} LSb\text{-}MC_{pk} \\ &s.t. \\ &LSb\text{-}MC_{pk} = f(\boldsymbol{\tau}_{orth}) \\ &\hat{\boldsymbol{x}}_{orth}^{new} = \mathbf{P}_{orth}\boldsymbol{\tau}_{orth} \end{aligned} \tag{9.1}$$

where $\hat{\boldsymbol{x}}_{orth}^{new}$ states how to manipulate the process conditions to maximize the $LSb\text{-}MC_{pk}$ for a particular supplier.

9.3 Results

In this section, the results from applying each of the DMAIC steps (Define, Measure, Analyze, Improve, and Control) are shown. Note that, latent variable-based models such as the Sequential Multiblock (SMB) PLS, are used, instead of more classical ones such as MLR. Hence, the tools implemented in some of the steps of the DMAIC cycle differ from more traditional approaches, but the original purpose of each stage remains.

9.3.1 Define

The purpose of this stage is to identify opportunities for improvement that lead to e.g., an increase in benefits, reduced costs or losses, a mitigation of the environmental impact, etc. This requires pinpointing observed problems, framing them within the context of the corresponding processes, evaluating the costs and benefits of addressing them, and locating the most appropriate people to do it given the existing constraints on time and resources.

In this Six Sigma project, the focus was set on the ratio between the cheese produced and the corresponding quantity of milk used in a cheese production process (i.e., CQA). This came as a result of an observed seasonal variability of this CQA with time yielding in some situations, ratios lower than a lower specification limit. These situations involve a decrease in the average value of the cheese produced.

9.3.2 Measure

9.3.2.1 Available Data

Altering or interrupting the production of this cheese production process was not allowed to any extent, and hence experimenting on the plant itself was not an option either. Due to this, only historical data from past production could be used. Thus, data from a total of 1917 curd vats (i.e., batches) were available from January 1, 2021, to August 30, 2022, containing information about:

- Four milk properties ($\boldsymbol{z}_1$ to $\boldsymbol{z}_4$).
- One manipulated variable referring to the cheese ingredient ($\boldsymbol{x}_{17}$).
- 18 manipulated variable referring to the cheese curd maker process ($\boldsymbol{x}_1$ to $\boldsymbol{x}_{16}$, and $\boldsymbol{x}_{18}$ to $\boldsymbol{x}_{19}$).
- Information on one CQA ($\boldsymbol{y}$), measured at the end of the process. This refers to the ratio between the cheese produced and the corresponding quantity of milk used.

Thus, the historical dataset can be structured in three blocks, milk properties $\mathbf{Z}$, process manipulated variables $\mathbf{X}$, and CQA, $\boldsymbol{y}$. Finally, note that the detection of outliers was carried out by a PCA model, yielding a validated dataset of 1917 observations.

9.3.2.2 Define initial situation

The purpose of this part is to establish process performance baselines. In this sense, Figure 9.1 is presented to show the evolution of $\boldsymbol{y}$ with time.

Most striking in Figure 9.1 is the noticeable fluctuation in seasons over a period of time. For that reason, it was decided to structure the dataset in two campaigns. The first campaign (C1) involves data from October to May (i.e., winter campaign), and the second campaign (C2) involves data from June to September (i.e., summer campaign). Figure 9.1 shows that for both 2021 and 2022 the C1 yield values of $\boldsymbol{y}$ substantial lower than the C2. In fact, the C1 has values out of specifications.

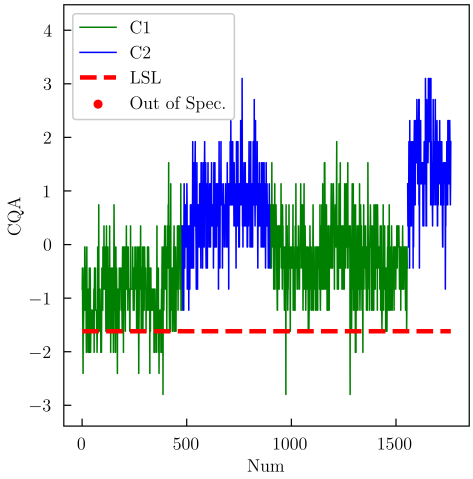


Figure 9.1: Evolution of $\boldsymbol{y}$ with time for the first campaign (C1) and second (C2). LSL (Lower Specification Limit).

9.3.3 Analyze

The main goal of this stage was to identify which input variables are related to the $\boldsymbol{y}$, taking into account the process flowsheet with the first block $\mathbf{Z}$ containing incoming milk properties, and process data in the second block $\mathbf{X}$. As discussed in Chapter 6, the SMB PLS model is of particular interest in these scenarios.

Thus, one component was found sufficient to capture the impact of milk properties (and correlated process variations) on $\boldsymbol{y}$ in the first modelling step. One additional component was also needed in the second modelling step to model the effect of orthogonal variations in process conditions on the remaining variations in $\boldsymbol{y}$. The goodness of fit, $R^2_{\mathbf{Y}_{cum}}$ (i.e., variability percentage explained by the model) for each one of the input blocks, $\mathbf{Z}$ and $\mathbf{X}$, and the output block, $\boldsymbol{y}$, and each component, is presented in Figure 9.2.

Figure 9.2 shows that the first component explains 74.14% of the information in $\mathbf{Z}$ and 18.26% of the information in $\mathbf{X}$ that was correlated with $\mathbf{Z}$, to explain a relevant percentage of CQA (54.35%). Then, the orthogonal component of the second modelling step shows that the 15.72% of the variation in $\mathbf{X}$, not related to $\mathbf{Z}$, is able to explain 9.37% of the CQA. To analyze and comprehend the behavior of the process, Figure 9.3a shows the score plot, and Figure 9.3a

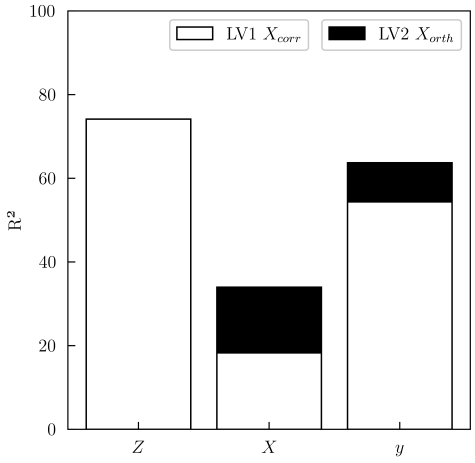
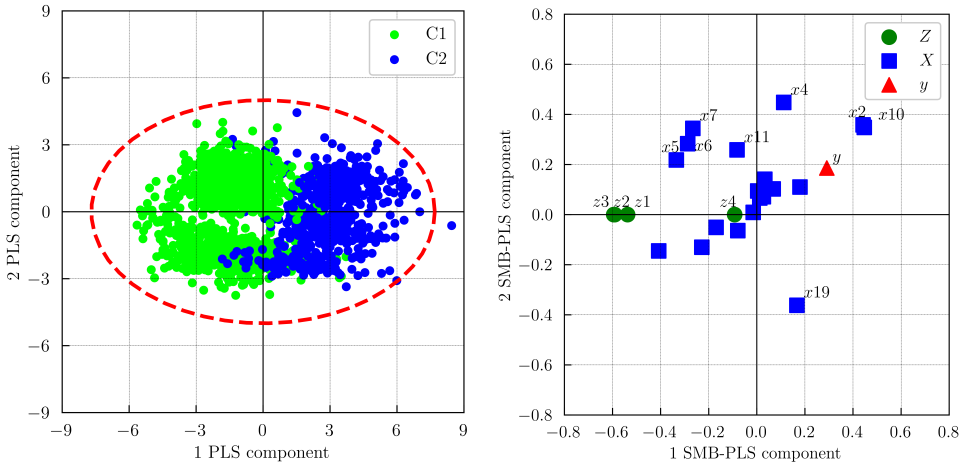


Figure 9.2: Explained **Z**, **X** and $\boldsymbol{y}$ variability for the SMB-PLS model depending on either the number of latent variables (LVS) or the two blocks of latent variables (LV1 explain the first block [$\mathbf{ZX}_{corr}$], and LV2 explains the second block $\mathbf{X}_{orth}$).

shows the bi-plot of the block weights and $\boldsymbol{y}$ loadings for these components to understand the behavior of process conditions.

The most obvious finding to emerge from Figure 9.3a is that the first latent variable is clearly associated with behavioral variations between C1 and C2. This implies that the milk properties (and the corresponding process correlation) have a significant impact on the different campaigns. Indeed, this latent variable shows that C1 tends to have negative scores. This means that this campaign has relatively higher values of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$ with respect to C2 (see Figure 9.3b). Besides, this tendency seems to be negatively related to some process variables, such as $\boldsymbol{x}_2$ and $\boldsymbol{x}_{10}$, as they present positive values in the first latent variable. These two variables also present high values in the second latent variable. Hence, it is concluded that some of their variability is related to the $\boldsymbol{y}$ from variation in milk properties, while there is also additional orthogonal variability that is related as well. Figure 9.3b also shows that $\boldsymbol{x}_4$ and $\boldsymbol{x}_{19}$ are negatively related according to the second component. This correlation barely is associated with the first component and, hence, a relation with milk properties is hardly expected. Finally, Figure 9.3b shows that $\boldsymbol{y}$ is located in the first quadrant where both the first and the second latent variables are positive. This implies that the differences between campaigns according to the first latent variable are also associated with the variations in $\boldsymbol{y}$. This point seems to explain the differences in $\boldsymbol{y}$ between campaigns found in Figure 9.1.



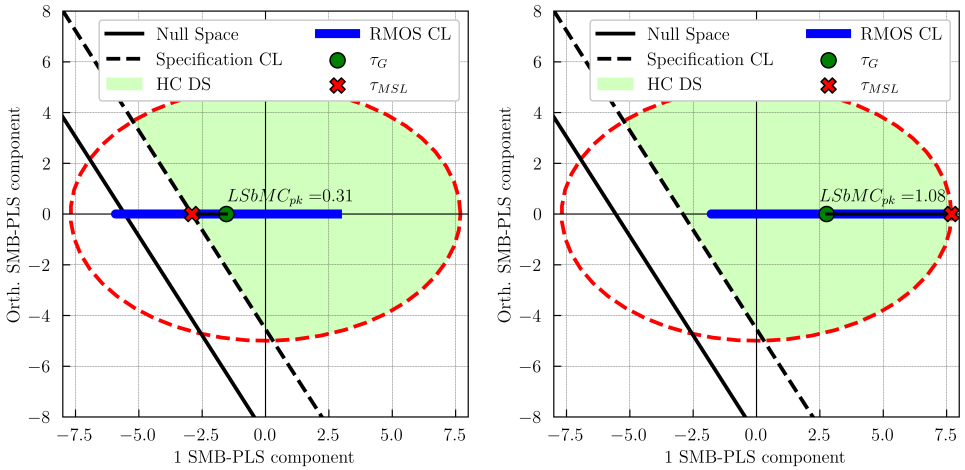
(a) Score plot by showing calibration data for the first campaign (C1) and second (C2). **(b)** Bi-plot of the block weights and $\boldsymbol{y}$ loading.

Figure 9.3: SMB PLS model.

In addition to that, 9.37% of the variation in $\boldsymbol{y}$, which is associated with the second component, can be interpreted as the impact of $\mathbf{X}$ not related to $\mathbf{Z}$. This inference supports the idea that there is potential for improvements or optimizations by manipulating the process.

On the other hand, even though this case study does not present information about suppliers, there are two campaigns that could be assessed by the Latent Variable-space Multivariate Capability Index ($LSb\text{-}MC_{pk}$) analogously to Section 7. Thus, this index would allow quantifying the capacity of each campaign's milk of providing assurance of quality with a certain confidence level for the CQA of the cheese. Figure 9.4 shows the $LSb\text{-}MC_{pk}$ for the first campaign (Figure 9.4a) and the second (Figure 9.4b) assuming that orthogonal variations remain at the average value.

Figure 9.4 shows that C2 presents a higher $LSb\text{-}MC_{pk}$ than C1 (1.08 vs 0.31, respectively). Thus, the expected ability to obtain a $\boldsymbol{y}$ superior or equal to 7.2 is higher for C2. This finding is consistent with the already discussed in Figure 9.1.



(a) First campaign. **(b)** Second campaign.

Figure 9.4: Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) for:

9.3.4 Improve

As a result of the analyses performed and summarized in the previous section, it is concluded that the expected ability to obtain a $\boldsymbol{y}$ superior or equal to 7.2 for C1 is poor yielding in some cases values outside of specifications. Besides, the SMB PLS model exhibited that 9.37% of the variation in $\boldsymbol{y}$ can be inferred as the effect of variation in $\mathbf{X}$ not related to $\mathbf{Z}$. Therefore, the proposed improvement involves optimizing the $LSb\text{-}MC_{pk}$ for C1 by manipulating the process as proposed in Section 9.2.2 (Figure 9.5).

Figure 9.5 shows that expected (Exp.) RMOS control (i.e., $\boldsymbol{\tau}_{orth}^{new} = \mathbf{0}$) is moved to search for an optimal $LSb\text{-}MC_{pk}$. The latter arises from adding the orthogonal variation, and thus, the $LSb\text{-}MC_{pk}$ improves from 0.31 to 0.83. Finally, Figure 9.6 shows how to manipulate the process conditions, with respect to the expected values according to the first block of latent variables, to achieve this optimal $LSb\text{-}MC_{pk}$ for the C1.

Most relevant contributions for the optimal $LSb\text{-}MC_{pk}$ are highlighted in Figure 9.6. Note that, since causality is inferred in the reduced orthogonal latent space, process conditions are manipulated to be consistent with the latent orthogonal structure (i.e., the second component) shown in Figure 6.2.

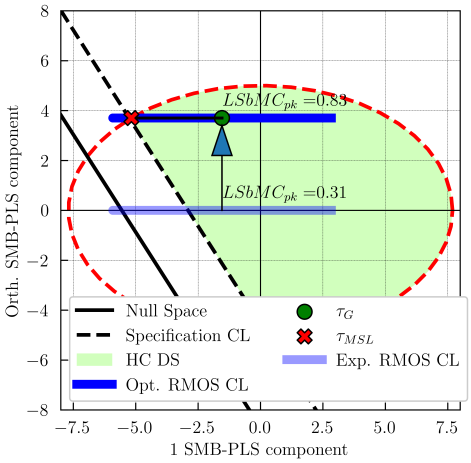


Figure 9.5: Graphical interpretation of the optimization of the Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) for the first campaign.

9.3.5 Control

Once the potential improvement is defined, the purpose of the control step is to embed the changes and ensure sustainability. In the current case study, the changes have not been applied yet, and hence, empirical verification is not available. However, the SMB PLS model can be used to predict $\boldsymbol{y}$ in the case of using historical milk properties data, jointly with the proposed improvement only for C1 (i.e., optimal prediction). Thus, one can compare the evolution of $\boldsymbol{y}$ according to the optimal prediction, and the historical CQA data that are available (Figure 9.7).

Figure 9.7 reveals that, for C1, the optimal scenario yields higher values of $\boldsymbol{y}$ than historical values due to the proposed improvement. Moreover, since the C2 was not optimized, the predicted values seem to correspond to the historical evolution of $\boldsymbol{y}$ as expected. Finally, note that, despite the improvement, C1 still has slightly lower $\boldsymbol{y}$ values than C2. This finding is in agreement with the $LSb\text{-}MC_{pk}$ obtained, as the optimal value of $LSb\text{-}MC_{pk}$ for C1 (i.e., 0.83) was lower than C2 (1.08).

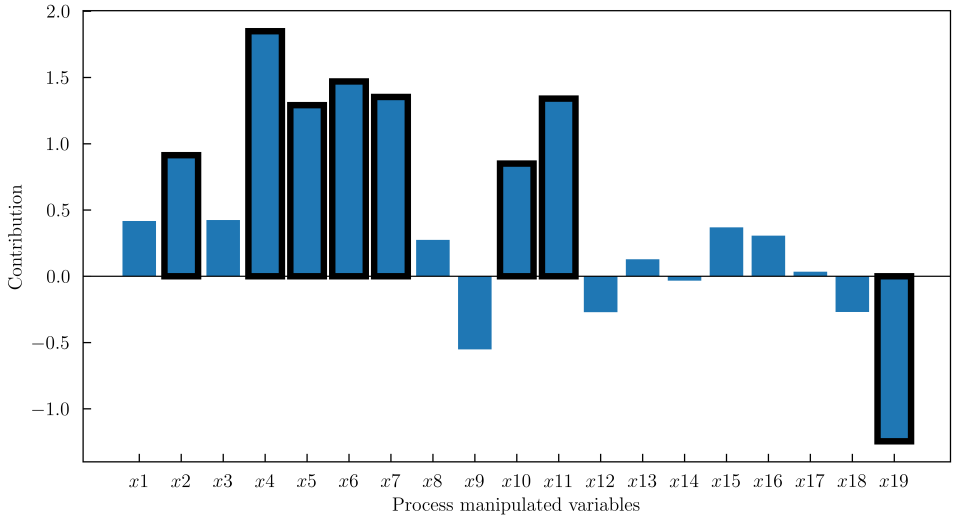


Figure 9.6: Contribution of the manipulated process variables for the optimal $LSb\text{-}MC_{pk}$.

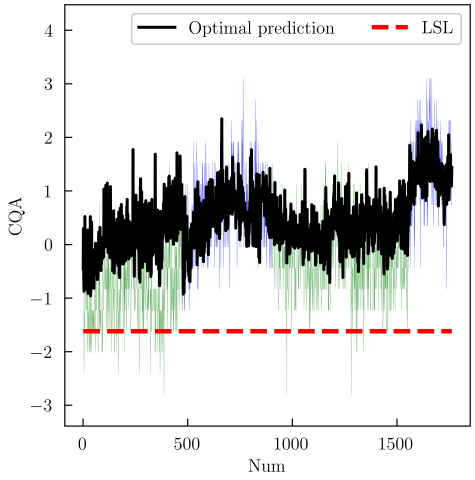


Figure 9.7: Evolution of the $\boldsymbol{y}$ according to the proposed improvement. The historical evolution of $\boldsymbol{y}$ with time for the first campaign (green) and second (blue) is also shown in the background for comparison.

9.4 Conclusion

Traditional Six Sigma statistical toolkit, mainly focused on classical statistical techniques (such as scatterplots, correlation coefficients, and linear regression models from experimental designs), is seriously handicapped for problem-solving using process data coming from Industry 4.0. In this context, abundant historical process data involving hundreds/thousands of variables highly correlated with missing values are registered from daily production.

Latent variable-based regression models can be used in this context providing unique and causal models in the latent space. In particular, the SMB PLS model is used for optimizing the Latent Space-based Multivariate Capability Index ($LSb\text{-}MC_{pk}$) for a particular campaign by manipulating process variables. Thus, the ability to obtain a $\boldsymbol{y}$ superior or equal to the corresponding specification limit is enhanced for this campaign.

Part IV

Graphical user interface

Chapter 10

Dragonet: a software for data analysis and process optimization

10.1 Introduction

The problem of data analysis and process optimization from historical datasets by means of latent variable-based models arises in several research areas. To assist scientists across various research areas, we introduce here a GUI in Python, called Dragonet, devoted to introducing not only the conventional latent variable-based uses but also the novel methodological contributions presented in this thesis.

The aim of this chapter is to show the main uses by means of an illustrative tutorial. This tutorial involves five steps as follows:

1. Importing data (see Section 10.2).
2. Building a model (see Section 10.3).
3. Data analysis (see Section 10.4).
4. Process optimization (see Section 10.5).
5. Defining the High-Confidence Design Space (HC DS) (see Section 10.6).

10.2 Importing data

To start a project, start up Dragonet resulting in the main menu of the software that is shown in Figure 10.1.

Subsequently, to import data, click `File | New` from the menu bar, or directly click `New` from the toolbar in (Figure 10.1). The data manager window will appear (Figure 10.2). From here one can search for the data file by clicking `Browser`. Note that, the use of Excel spreadsheets is required, where each worksheet refers to a different block. One can import as many blocks as desired. These blocks will appear in the display called `available sheets` (Figure 10.2), in this case, X and Y blocks. Data must be organized simply, with each row representing an observation and each column representing a variable. The primary ID is a mandatory column for all blocks, and the secondary ID columns are optional. Every block must have the same number of rows referring to the corresponding primary ID.

When clicking a block, the corresponding worksheet will be displayed in the display called `Data` (Figure 10.2), in this case, the X block. Then, one can choose the type of data for each column by selecting `ID` (for primary ID), `Sec.`

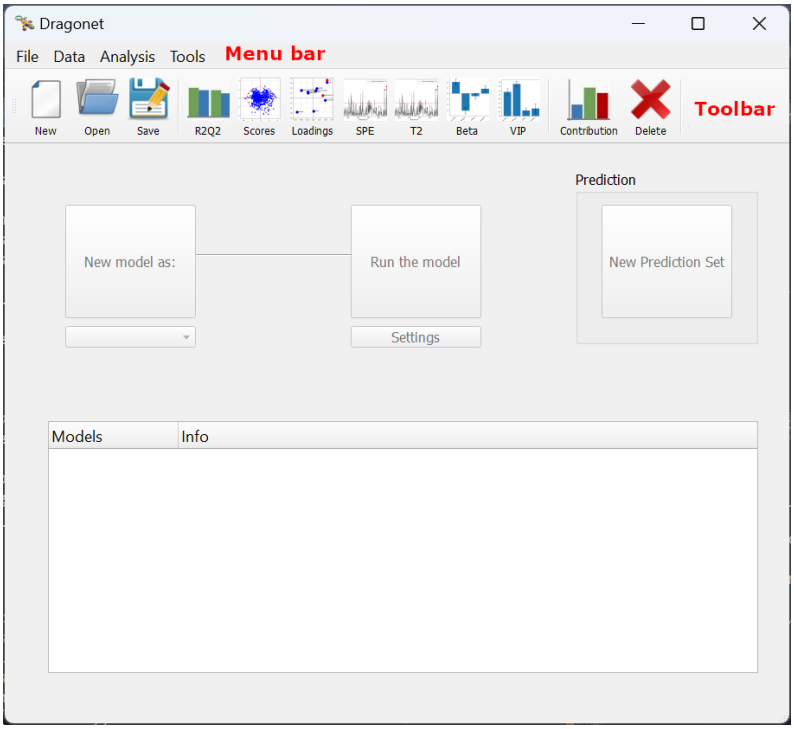


Figure 10.1: Dragonet main menu.

`ID` (for secondary ID), `Numerical` (for continuous variable), and `Categorical` (for categorical variables). Finally, to add the block, select `Add block`. Once all desired blocks are imported, click `Ok` to finish importing data, and the data manager window will close.

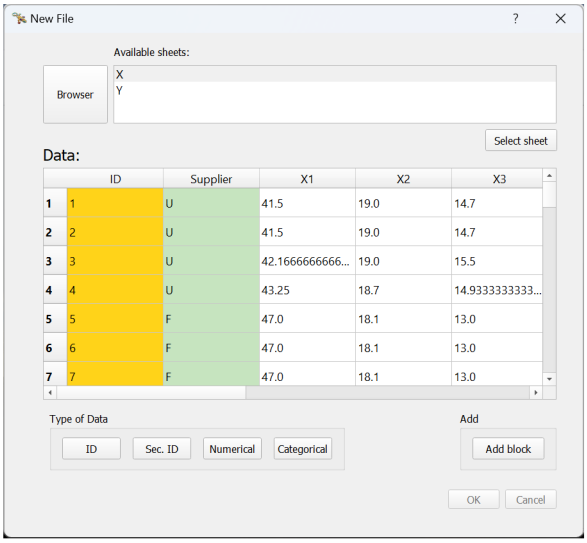
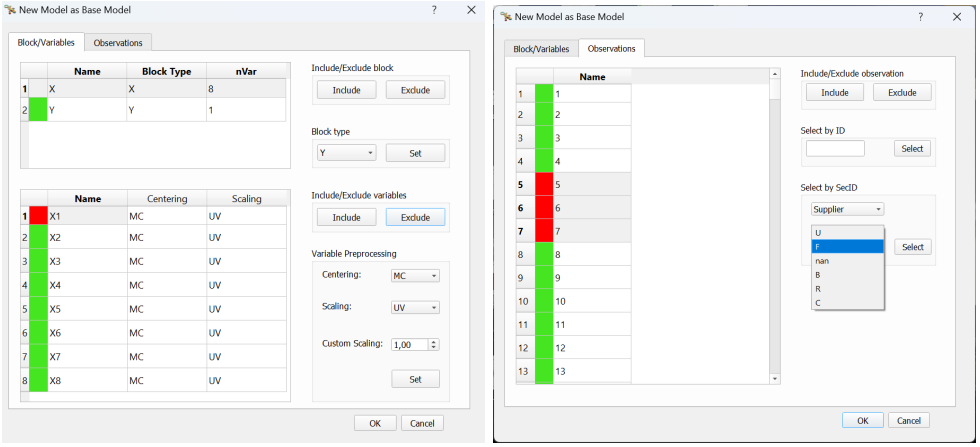


Figure 10.2: Data manager window.

10.3 Building a model

To build a new model, select the reference model in the drop-down below the button called `New Model as:` in the main menu (Figure 10.1). If the first model is built, the reference model will correspond to the `Base model`, namely, the model including all observations and variables. Then click the button called `New model as:` in the main menu (Figure 10.1), and the new model window will appear (Figure 10.3) in order to define the model specifications. The default model specifications are those of the reference model.

The new model menu has two tabs, the `Block/Variables` tab (Figure 10.3a), and the `Observation` tab (Figure 10.3b). The `Block/Variables` tab presents two displays. The top display shows the available blocks, and by clicking one of them the corresponding variables will be shown in the bottom display. This tab allows including and excluding the blocks as a whole (`Include/Exclude blocks`), or some of their variables (`Include/Exclude variables`). In this case, the variable X1 is excluded. Moreover, the block type, X (for input variables) or Y (for output variables), can be set (`Block type`). If only X blocks are set, the model will refer to a Principal Component Analysis (PCA) model. By contrast, if there are both X and Y blocks, the model will refer to a PLS or SMB PLS model. In addition to that, the preprocessing of the variables can be also set (`Variable Preprocessing`).



(a) Variable tab. **(b)** Observation tab.

Figure 10.3: New model window

On the other hand, the `Observation` tab presents only one display showing all observations. From this display, one can include or exclude some observations manually or can filter according to the level of a Secondary ID. For instance, level F of the Secondary ID Supplier is selected and then excluded. Finally, click `Ok` to finish the model specifications, and the new model window menu will close.

Before running the model the settings of the model can be defined. For that, click the button called `Settings` in the main menu (Figure 10.1), and the model settings window will appear (Figure 10.4).

At the top, this window shows the type of model (in this case PLS) and the name of the model (in this case Model 1). First, one can set the number of components, the number of groups or folds in the Cross-Validation process, and the confidence limits. In addition to that, if having a PLS model, one may select the check box `SMB PLS` to build the SMB PLS model and define the number of blocks and their characteristics (it is not the case). Finally, click `Ok` to finish the model settings, and the model settings window will close.

Once the specifications and the settings of the model are defined, click `Run the model` in the main menu (Figure 10.1) to build the corresponding model. Each time a model is built, one can find the main information for the corresponding model in the model list located at the bottom of the main menu. For instance, Figure 10.5 shows that the current project has four different models. Regarding

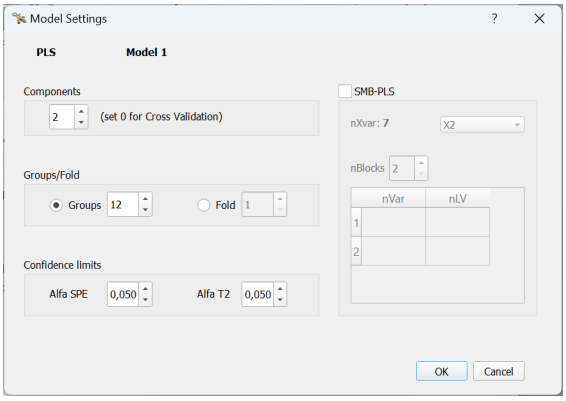


Figure 10.4: Model settings window.

the first model, it corresponds to a PLS model with two latent variables, 969 observations, and 7 regressor variables.

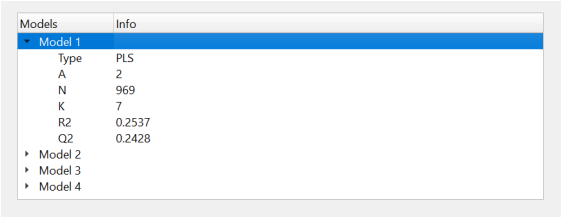


Figure 10.5: Dragonet main menu - Model list

10.4 Data analysis

To analyze the data it is necessary to select the corresponding model in the `model list`. Thus, one can show any of the model analysis plots of the software for the selected model. The following plots for model analysis are available in `Analysis` from the menu bar in the main menu, or by directly clicking the desired plot from the toolbar:

- `R2Q2` to show the cumulative R^2 and Q^2 metrics for each model component.
- `Scores` to create a score plot.
- `Loading` to create a loading plot (for PCA) or weighting plot (for PLS).

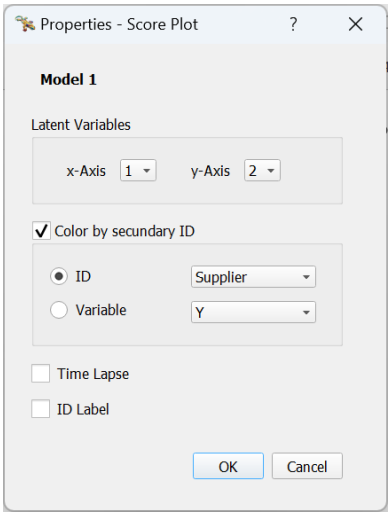
- `SPE` to create a Squared Prediction Error (SPE) plot.
- `T2` to create a Hotelling's T^2 plot,.
- `Beta` to create a Coefficients plot. This is available for PLS models only.
- `VIP` to create a Variable Importance to the Projection (VIP) plot.

To display the properties window for each of these plots, click right on the corresponding plot. In addition to that, these plots are interactive. Indeed, Beta and VIP plots allow the variable selection to exclude these variables by clicking `Delete` from the toolbar, and the score plot allows the observation selection to both exclude these observations by clicking `Delete` from the toolbar, and calculate the contribution plots by clicking `Contribution` from the toolbar. Note that, `Delete` and `Contribution` are not dependent on the selected model in the model list, but they are dependent on the model corresponding to the selected window.

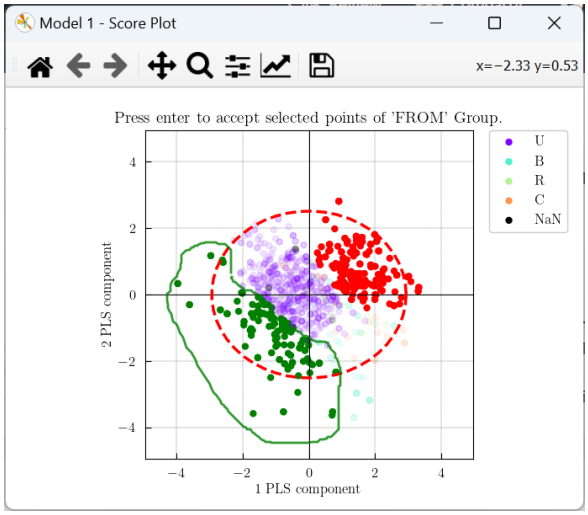
The score plot is used as an example to illustrate how to work with plots. Figure 10.6a allows setting the properties of the score plot, and Figure 10.6b shows the plot score after setting from the properties window that the observations will be colored by the Supplier secondary ID. Notice that, by selecting `Time Lapse` from Figure 10.6a, this properties window also allows showing the plot score as a Graphics Interchange Format (GIF) where observations are plotted chronologically according to the import order. This may be useful to comprehend the evolution process through the latent space.

Additionally, Figure 10.6b shows an example of observation selection. First, a red group is selected and, then, a green group is being selected. Thus, to calculate the contribution plots between two groups (from green to red group), click `Contribution` from the toolbar. This action will display the contribution plot (Figure 10.7a). Besides, one can click on any variable of the contribution plot, and press enter to show its corresponding time series. On the contrary, if two variables are clicked before pressing enter, the corresponding scatter plot will be displayed. In this case, only the X2 variable is selected from Figure 10.7a in order to show its corresponding time series in Figure 10.7b.

Note that, the observation selection in Figure 10.6b is kept in Figure 10.7b. This is feasible since all observation plots (i.e., score plots, SPE plots, T^2 plots, time series plots, and scatter plots) for a model are in sync. Namely, any selection change in a particular observation plot will be instantly carried out in all observation plots of the same model, and it will keep in upcoming observation plots.

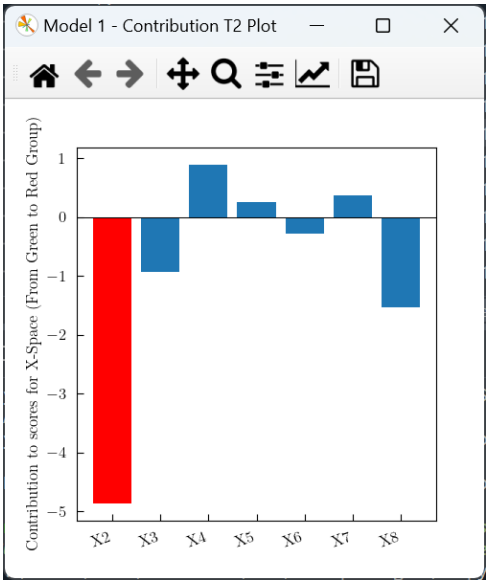


(a) Score plot properties window.

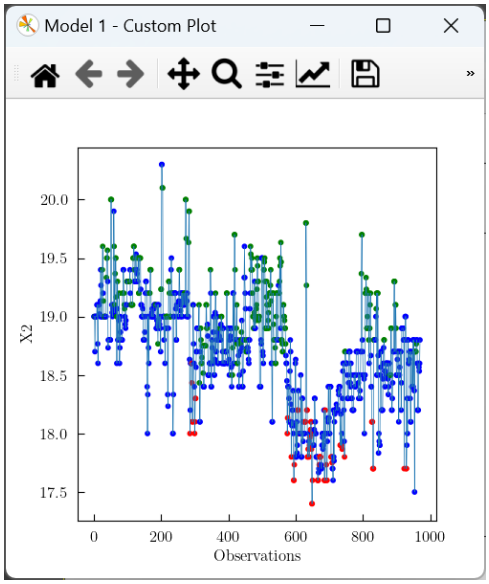


(b) Score plot colored by colored secondary ID.

Figure 10.6



(a) Contribution plot.



(b) Time series plot.

Figure 10.7

10.5 Process optimization

Once a PLS model has been built in Dragonet, the optimization menu can be displayed from `Tools | Optimization` from the menu bar in the main menu showing Figure 10.8. This tool can be enabled to optimize a process using PLS models built entirely on historical data.

The top of this menu is structured into three tabs:

- `Settings` to set configuration settings such as the target and penalty weight of the outputs (see at the top of Figure 10.8).
- `X Constraints` to define the X constraints. In this case, historical constraints are set for all input variables except X3, and an equality hard constraint is set to the X3 variable (see Figure 10.9).
- `T2 and SPE Constraints and Weights` to configure T^2 and SPE constraints and overall weights of the optimization problem (see Figure 10.10).

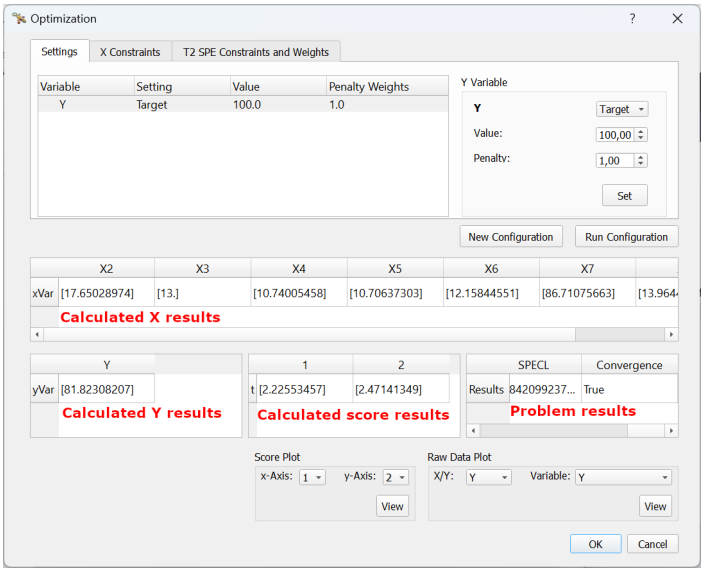


Figure 10.8: Optimization menu.

Once the configuration of the optimization problem is defined, click `Run Configuration` from the optimization menu, and the results will be displayed in the four displays (Figure 10.8). Notice that, the optimization problem converges

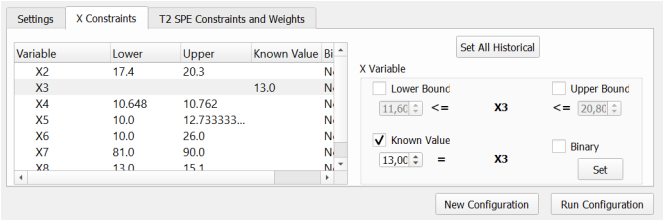


Figure 10.9: Optimization menu - **X** Constraints tab.

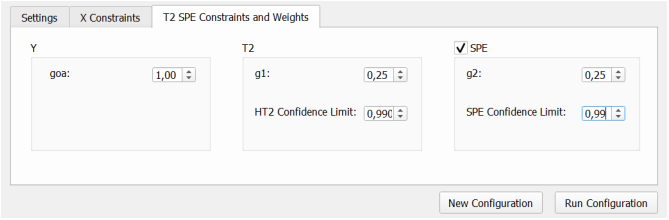
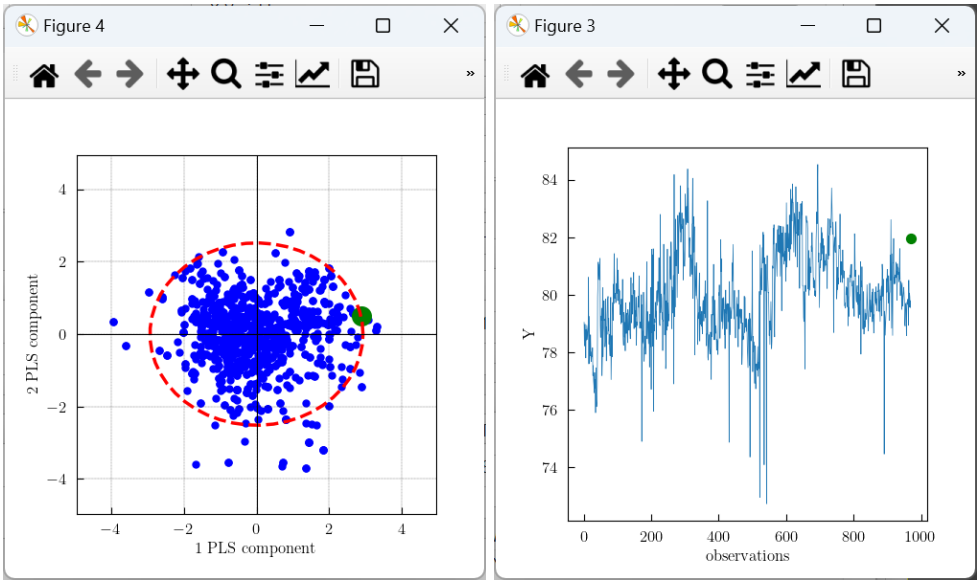


Figure 10.10: Optimization menu - T^2 and SPE constraints, and weights tab.

in an optimal solution (see Problem results), and the hard constraints are respected (see Calculated X results). Finally, one can also create new plots by showing the optimal solution, jointly with the historical data. For instance, Figure 10.11a shows the score plot by highlighting the optimal score, and Figure 10.11b highlights the expected output for this optimal score.



(a) Score plot by showing the optimal solution (green point).

(b) Output time series by showing the optimal solution (green point).

Figure 10.11

10.6 Defining the high-confidence design space

Once a PLS model has been built in Dragonet, the High-Confidence Design Space (HC DS) can be defined. For that, click `Tools | HC DS`, and the High-Confidence Design Space (HC DS) menu will be displayed (Figure 10.12).

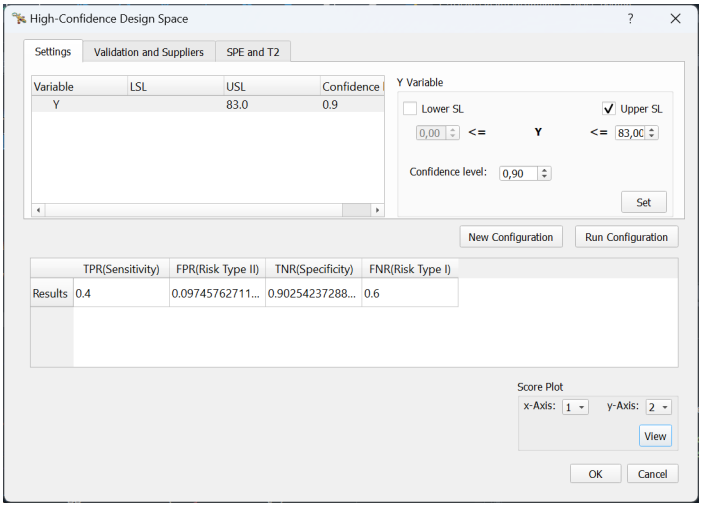


Figure 10.12: High-Confidence Design Space (HC DS) menu.

In the settings tab, one can specify the specification limits and the confidence level for the output. Then, click `Run Configuration` to get the results of Sensitivity, Risk Type II, Specificity, and Risk Type I. Additionally, one can visualize the HC DS from the `Score Plot` container (Figure 10.13).

The results presented in the display of Figure 10.12 are calculated from the validation set selected in the `Validation and Suppliers tab` (see Figure 10.14). In this case, the calibration set of the current model is selected as validation. The same tab allows selecting a particular level of a secondary ID to calculate the validation results only for observations belonging to this level. Note that, the available validation sets must be defined previously from the `New Prection Set` button in the main menu.

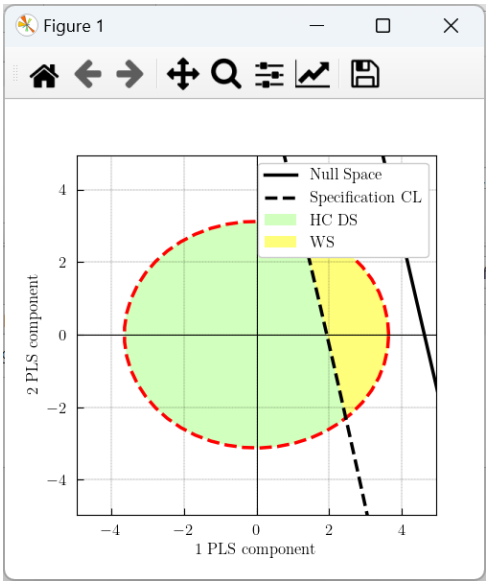


Figure 10.13: High-Confidence Design Space (HC DS) plot

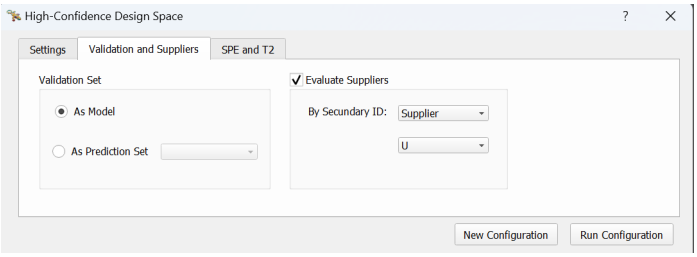


Figure 10.14: High-Confidence Design Space (HC DS) menu - Validation and Suppliers tab.

10.7 Conclusions

In this chapter, a new Python GUI is presented for analyzing historical datasets by means of latent variable-based models: Dragonet. This software integrates the developed methods in the thesis with the aim of being self-explanatory and user-friendly. The main uses of Dragonet are illustrated by a tutorial. Thereby, this tutorial guides the user step by step through the process of importing data, building a model, process optimization, and definition of the High-Confidence Design Space (HC DS).

Although it has not been explicitly shown in the tutorial, Dragonet also allows the user to diagnose problems with respect to past operations by contribution plots and establish multivariate control charts for monitoring the SPE and T^2 statistics using PCA and PLS models.

Finally, it is concluded that Dragonet integrates a digital representation of an intended process (a physical twin) that serves as the effectively indistinguishable digital counterpart, commonly referred to as the Digital Twin.

Part V

Epilogue

Chapter 11

Conclusions

11.1 Meeting the objectives

This thesis is devoted to developing causal latent variable-based models for scientific learning in Industry 4.0. The main conclusions of the thesis are summarized, and organized according to the objectives presented in Section 1.2.

Objective I: To study the properties of Partial Least Squares (PLS) regression to analyze data from Design of Experiments (DOE).

A novel methodology to analyze two-level full and fractional factorial designs, with or without missing runs, with one single technique, PLS, was proposed in Chapter 4. This property is very attractive for practitioners since, to the best of our knowledge, no other statistical tool has comparable versatility. In the case of a full and fractional factorial design, the one-PLS component model yields the same analytical solution as Multiple Linear Regression (MLR), not only in the estimation of the effects, but also in their statistical significance. When having missing runs in the factorial design, PLS is of particular interest as it is a powerful tool when dealing with complex correlation structures, as opposed to MLR. Thus, we challenge the widely held view that PLS is useful only when dealing with non-experimental design (i.e., correlated observational data). The methodology was synthesized by an easy-to-follow route map useful for practitioners.

Objective II: To define the raw material design space via latent variable-based models.

The first contribution devoted to accomplishing objective II is presented in Chapter 5. For that, it was proposed a novel methodology, making use of the PLS model inversion, to define multivariate raw material specification region in the latent space where there is assurance of quality with a certain confidence level for the Critical Quality Attributes (CQAs) of the final product (i.e., the so-called high-confidence raw material design space). Thus, it allows the evaluation of the capability of the raw material batches of producing products with CQAs within specification limits, before producing a single unit of the product, and based on that information, making a decision about accepting or not the supplier raw material batch. This is totally different from existing approaches that evaluate (and also accept or reject) raw material batches based on their raw material properties but not on the desired final product properties.

Since not only raw material properties influence the quality of the final product, but also the process conditions, Chapter 6 considered the possibility to modify process conditions to compensate for raw material properties variations

by means of a novel methodology based on the SMB-PLS model inversion. The model enables the identification of variation in process conditions uncorrelated with raw material properties and known disturbances, which is crucial to implement an effective process control system attenuating most raw material variations. The latter allows expanding the specification region and, hence, one may potentially be able to accept lower-cost raw materials that will yield products with perfectly satisfactory quality properties.

Both novel methodologies are based on the latent variable-based model inversion, and the most remarkable advantages are:

- They can be used with historical data (i.e., daily production data not coming from any experimental design but with varying raw material properties, typical from Industry 4.0 environment) since, when fitting latent variable-based models, causality can be inferred in the latent space, which allows the meaningful inversion of the model.
- They consider a multivariate approach providing much insight into what constitutes acceptable raw material batches when their properties are correlated.
- They use mathematical and statistical models as a way to define such raw material specifications by linking them with specification limits for CQAs of the final product. It enables a frequentist probabilistic interpretation, namely, the multivariate raw material region is expected to produce products with CQAs within specification limits with a confidence level equal or higher than $(1 - \alpha) \times 100$.
- They provide the analytical definition of the limits of the multivariate raw material specifications.

Objective III: To develop a latent space-based multivariate capability index.

Chapter 7 presented a Latent-Space based Multivariate Capability Index (LSb-MC_{pk}) that allows ranking and selecting suppliers for a particular raw material used in a manufacturing process. This index arises from comparing the supplier's Raw Material Operating Space (RMOS) with the High-Confidence Raw Material Specification Region (HC-RMSR). RMOS is a region in the latent space linking the raw material properties (input space) with the CQAs of the product manufactured (output space), where the supplier's raw material samples are expected to be located at a certain confidence level. On the other hand, HC-RMSR corresponds to the high-confidence raw material design space

defined in Objective II. This is a region in the latent space connecting both input and output spaces associated with raw materials properties providing assurance of quality for the CQAs of the manufactured product with a certain confidence level.

This index quantifies the ability of each supplier of a particular raw material to produce a certain percentage of final product within its CQAs specifications, and this information can be obtained at the reception of the supplier´s raw material, before producing a single unit of the product. Finally, diagnosing assignable causes is carried out when the samples of the supplier's raw material does not respect the correlation structure from the past (by using the SPE contribution plots), or when the supplier is not able to consistently operate within the HC-RMSR (by using the score contribution plots).

Objective IV: To illustrate the use of PLS for process optimization using happenstance data.

Part III (Chapters 8 and 9) illustrated how causal latent variable-based models, such as PLS, can be used for process optimization using happenstance data with respect to two novel applications. Indeed, by employing causal latent variable-based models, it was possible to identify the underlying causal relationships between variables and leverage them to optimize processes. These applications provided a deeper understanding and allowed for better decision-making and optimization strategies.

Chapter 8 involved a health application related to the COVID-19 Pandemic. Within the context of the Spanish Society of Hospital Pharmacy (SEFH) project, latent variable-based models were applied to develop an alternative to placebo-controlled clinical trials. In fact, latent variable-based models used data from daily tested patients (instead of classical clinical trials) in order to infer causality in the latent space. This could allow for improvements over the past with respect to the use of therapies.

Chapter 9 presented an industrial case study related to a cheese production process where the Six Sigma DMAIC (Define-Measure-Analyze-Improve-Control) method was used as an approach for problem-solving. For that, the use of latent variable-based models was integrated into the Six Sigma statistical toolkit yielding the Multivariate Six Sigma: a powerful process improvement methodology for Industry 4.0.

Objective V: To integrate the developed methods by means of a Graphical User Interface (GUI).

In Chapter 8, a new Python Graphical User Interface (GUI) was presented for analyzing historical datasets by means of latent variable-based models: Dragonet. Dragonet integrates the developed methods in the thesis with the aim of being self-explanatory and user-friendly. The main uses of Dragonet were illustrated by a tutorial. Thereby, this tutorial is able to guide the user step by step through the process of importing data, building a model, process optimization, and definition of the High-Confidence Design Space (HC DS).

11.2 Future research lines and transfer activities

This Ph.D. manuscript opens some future lines:

- Extend the method of estimating the $LSb\text{-}MC_{pk}$ in the case a supplier provides different types of raw materials.
- Validate, by empirical verification, the optimal solution proposed according to the latent variable-based customized optimization problems.
- Apply the idea behind the latent variable-based customized optimization problems, arising from health applications, to industrial case studies.
- Adapt the latent variable-based optimization problems to integer programming for those problems subjected to binary variables.
- Transfer the work done in this thesis to educational systems relating to both academia and industry.

Bibliography

[1] G. E. Box, "Science and Statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[2] A. Ferrer, "Multivariate six sigma: A key improvement strategy in industry 4.0," *Quality Engineering*, 2021, ISSN: 15324222. DOI: 10.1080/08982112.2021.1957481.

[3] M. S. Reis and G. Gins, "Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis," *Processes*, vol. 5, no. 3, 2017, ISSN: 22279717. DOI: 10.3390/pr5030035.

[4] J. F. MacGregor, "Empirical Models for Analyzing "BIG" Data – What's the Difference?" *Spring AIChE Conf.*, 2018.

[5] G. E. P. Box, J. S. Hunter, and W. G. Hunter, *Statistics for Experimenters: Design, Innovation and Discovery*. John Willey & Sons, 2005.

[6] L. H. Chiang and R. D. Braatz, "Process monitoring using causal map and multivariate statistics: Fault detection and identification," *Chemometrics and Intelligent Laboratory Systems*, vol. 65, no. 2, pp. 159–178, 2003, ISSN: 01697439. DOI: 10.1016/S0169-7439(02)00140-5.

[7] Y. Shu and J. Zhao, "Data-driven causal inference based on a modified transfer entropy," *Computers and Chemical Engineering*, vol. 57,

pp. 173–180, 2013, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2013.05.011.

[8] R. Paredes, T. J. Rato, and M. S. Reis, "Causal network inference and functional decomposition for decentralized statistical process monitoring: Detection and diagnosis," *Chemical Engineering Science*, vol. 267, p. 118 338, 2023, ISSN: 00092509. DOI: 10.1016/j.ces.2022.118338.

[9] C. M. Jaeckle and J. F. MacGregor, "Product Design through Multivariate Statistical Analysis of Process Data," *AIChE Journal*, vol. 44, no. 5, pp. 1105–1118, 1998. DOI: 10.1016/0098-1354(96)00182-2.

[10] ICH Harmonised Tripartite, *Guidance for Industry Q8(R2) Pharmaceutical Development*, Rockville, USA, 2009.

[11] R. P. Cogdill and J. K. Drennen, "Risk-based quality by design (QbD): A Taguchi perspective on the assessment of product quality, and the quantitative linkage of drug product parameters and clinical performance," *Journal of Pharmaceutical Innovation*, vol. 3, no. 1, pp. 23–29, 2008, ISSN: 18725120. DOI: 10.1007/s12247-008-9025-3.

[12] T. Davis, "Science, Engineering, and Statistics," *Applied Stochastic Models in Business and Industry*, vol. 22, pp. 401–430, 2006, ISSN: 1524-1904. DOI: 10.1002/asmb.

[13] D. Palací-López, J. Borràs-Ferrís, and L. Thaise da Silva de Oliveria, "Multivariate Six Sigma: A Case Study in Industry 4.0," *Processes*, vol. 8, pp. 1–20, 2020. DOI: doi:10.3390/pr8091119.

[14] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "Defining multivariate raw material specifications in industry 4.0," *Chemometrics and Intelligent Laboratory Systems*, vol. 225, 2022, ISSN: 18733239. DOI: 10.1016/j.chemolab.2022.104563.

[15] A. González-Cebrián *et al.*, "Machine-learning-derived predictive score for early estimation of COVID-19 mortality risk in hospitalized patients," *PLoS ONE*, vol. 17, no. 9 September, pp. 1–17, 2022. DOI: 10.1371/journal.pone.0274171.

[16] A. González-Cebrián, J. Borràs-Ferrís, Y. Boada, A. Vignoni, A. Ferrer, and J. Picó, "PLATERO: A calibration protocol for plate reader green fluorescence measurements," *Frontiers in Bioengineering and Biotechnology*, vol. 11, pp. 1–19, 2023, ISSN: 22964185. DOI: 10.3389/fbioe.2023.1104445.

[17] J. Borràs-Ferrís, C. Duchesne, and A. Ferrer, "Defining Multivariate Raw Material Specifications via SMB-PLS," *Chemometrics and Intelligent Laboratory Systems*, vol. 240, 2023. DOI: 10.1016/j.chemolab.2023.104912.

[18] J. Borràs-Ferrís, A. Folch-Fortuny, and A. Ferrer, "On the properties of PLS for analyzing Design of Experiments," 2023, **SUBMITTED**.

[19] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, "A latent space-based Multivariate Capability Index: A new paradigm for raw material supplier selection in Industry 4.0," 2023, **SUBMITTED**.

[20] A. Ferrer, D. Palací-López, J. Borràs-Ferrís, M. Barolo, and P. Facco, "Inverse design via PLS model inversion," in *The Digital Transformation of Product Formulation Concepts, Challenges, and Applications for Accelerated Innovation*, Due to 2023, Taylor & Francis, ch. 10.

[21] A. Ferrer, J. Borràs-Ferrís, D. Palací-López, and C. Duchesne, "Multivariate specifications," in *The Digital Transformation of Product Formulation Concepts, Challenges, and Applications for Accelerated Innovation*, Due to 2023, Taylor & Francis, ch. 12.8.

[22] E. Tomba, P. Facco, F. Bezzo, and M. Barolo, "Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: A review," *International Journal of Pharmaceutics*, vol. 457, no. 1, pp. 283–297, 2013, ISSN: 18733476. DOI: 10.1016/j.ijpharm.2013.08.074.

[23] E. Tomba, M. Barolo, and S. García-Muñoz, "General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products," *Industrial & Engineering Chemistry Research*, vol. 51, no. 39, pp. 12 886–12 900, Oct. 2012, ISSN: 0888-5885. DOI: 10.1021/ie301214c.

[24] A. Höskuldsson, "PLS regression methods," *Journal of Chemometrics*, vol. 2, pp. 211–228, 1988. DOI: 10.1002/cem.1180020306.

[25] S. Wold, M. Sjostrom, and L. Eriksson, "PLS-Regression - A basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001. DOI: 10.1016/S0169-7439(01)00155-1.

[26] A. Ferrer, "Discussion of "A review of data science in business and industry and a future view" by Grazia Vicario and Shirley Coleman," *Applied Stochastic Models in Business and Industry*, vol. 36, pp. 23–29, 2020. DOI: 10.1002/asmb.2516.

[27] J. F. MacGregor, M. Bruwer, I. Miletic, M. Cardin, and Z. Liu, "Latent Variable Models and Big Data in the Process Industries," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 520–524, 2015, ISSN: 2405-8963. DOI: 10.1016/J.IFACOL.2015.09.020.

[28] T. Kourti and J. F. MacGregor, "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, vol. 28, no. 4, pp. 409–428, 1996. DOI: 10.1080/00224065.1996.11979699.

[29] P. Nomikos and J. F. MacGregor, "Multivariate SPC Charts for Batch Monitoring Processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995. DOI: 10.2307/1269152.

[30] N. D. Tracy, J. C. Young, and R. L. Mason, "Multivariate Control Charts for Individual Observations," *Journal of Quality Technology*, vol. 24, no. 2, pp. 88–95, 1992, ISSN: 0898-2112. DOI: 10.1080/00224065.1992.12015232.

[31] A. Ferrer, "Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process," *Quality Engineering*, vol. 19, no. 4, pp. 311–325, 2007, ISSN: 0898-2112. DOI: 10.1080/08982110701621304.

[32] G. Bano, P. Facco, N. Meneghetti, F. Bezzo, and M. Barolo, "Uncertainty back-propagation in PLS model inversion for design space determination in pharmaceutical product development," *Computers and*

Chemical Engineering, vol. 101, no. 9, pp. 110–124, 2017, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2017.02.038.

[33] K. Faber and B. R. Kowalski, "Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler," *Chemometrics and Intelligent Laboratory Systems*, vol. 34, no. 2, pp. 283–292, 1996, ISSN: 01697439. DOI: 10.1016/0169-7439(96)00022-6.

[34] L. Zhang and S. Garcia-Munoz, "A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective," *Chemometrics and Intelligent Laboratory Systems*, vol. 97, no. 2, pp. 152–158, 2009, ISSN: 01697439. DOI: 10.1016/j.chemolab.2009.03.007.

[35] C. M. Jaeckle and J. F. MacGregor, "Industrial applications of product design through the inversion of latent variable models," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 2, pp. 199–210, 2000, ISSN: 01697439. DOI: 10.1016/S0169-7439(99)00058-1.

[36] F. Yacoub and J. F. MacGregor, "Product optimization and control in the latent variable space of nonlinear PLS models," *Chemometrics and Intelligent Laboratory Systems*, vol. 70, no. 1, pp. 63–74, 2004, ISSN: 0169-7439. DOI: 10.1016/J.CHEMOLAB.2003.10.004.

[37] S. García-Muñoz, T. Kourti, J. F. MacGregor, F. Apruzzese, and M. Champagne, "Optimization of batch operating policies. Part I. Handling multiple solutions," *Industrial & Engineering Chemistry Research*, vol. 45, no. 23, pp. 7856–7866, 2006. DOI: 10.1021/ie060314g.

[38] E. Tomba, P. Facco, F. Bezzo, and S. García-Muñoz, "Exploiting historical databases to design the target quality profile for a new product," *Industrial & Engineering Chemistry Research*, vol. 52, no. 24, pp. 8260–8271, 2013, ISSN: 08885885. DOI: 10.1021/ie3032839.

[39] E. Arnese-Feffin, P. Facco, F. Bezzo, and M. Barolo, "Digital design of new products: accounting for output correlation via a novel algebraic formulation of the latent-variable model inversion problem," *Chemometrics and Intelligent Laboratory Systems*, vol. 227, 2022, ISSN: 18733239. DOI: 10.1016/j.chemolab.2022.104610.

[40] D. Palací-López, P. Facco, M. Barolo, and A. Ferrer, "New tools for the design and manufacturing of new products based on Latent Variable Model Inversion," *Chemometrics and Intelligent Laboratory Systems*, vol. 194, 2019. DOI: 10.1016/j.chemolab.2019.103848.

[41] D. Palací-López, P. Villalba, P. Facco, M. Barolo, and A. Ferrer, "Improved formulation of the latent variable model inversion–based optimization problem for quality by design applications," *Journal of Chemometrics*, no. February, pp. 1–18, 2020, ISSN: 1099128X. DOI: 10.1002/cem.3230.

[42] K. Azari, J. Lauzon-Gauthier, J. Tessier, and C. Duchesne, "Establishing multivariate specification regions for raw materials using SMB-PLS," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 1132–1137, 2015, ISSN: 24058963. DOI: 10.1016/j.ifacol.2015.09.120.

[43] L. E. Wangen and B. R. Kowalski, "A multiblock partial least squares algorithm for investigating complex chemical systems," *Journal of Chemometrics*, vol. 3, no. 1, pp. 3–20, 1989, ISSN: 0886-9383. DOI: 10.1002/cem.1180030104.

[44] T. Næs, O. Tomic, B. H. Mevik, and H. Martens, "Path modelling by sequential PLS regression," *Journal of Chemometrics*, vol. 25, no. 1, pp. 28–40, 2011, ISSN: 1099128X. DOI: 10.1002/cem.1357.

[45] J. Lauzon-Gauthier, P. Manolescu, and C. Duchesne, "The Sequential Multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability," *Chemometrics and Intelligent Laboratory Systems*, vol. 180, no. June, pp. 72–83, 2018, ISSN: 18733239. DOI: 10.1016/j.chemolab.2018.07.005.

[46] M. P. Campos, R. Sousa, and M. S. Reis, "Establishing the optimal blocks' order in SO-PLS: Stepwise SO-PLS and alternative formulations," *Journal of Chemometrics*, vol. 32, no. 8, 2018.

[47] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[48] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: 10.5281/zenodo.7037953.

[49] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2.

[50] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.5281/zenodo.6513224.

[51] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[52] F. Yates, "The analysis of replicated experiments when the field results are incomplete," *Empire Journal of Experimental Agriculture*, vol. 1, pp. 129–142, 1933.

[53] W. G. Cochran and G. M. Cox, *Experimental Designs*. New York: John Willey & Sons, 1957.

[54] N. R. Draper and D. M. Stoneman, "Estimating Missing Values in Unreplicated Two-Level Factorial and Fractional Factorial Designs," *International Biometric Society*, vol. 20, no. 3, pp. 443–458, 1964.

[55] T. J. Mitchell, "An Algorithm for the Construction of "D-Optimal" Experimental Designs," *Technometrics*, vol. 16, no. 2, pp. 203–210, 1974, ISSN: 15372723. DOI: 10.1080/00401706.1974.10489175.

[56] R. Xampeny, P. Grima, and X. Tort-martorell, "Which runs to skip in two level factorial designs when not all can be performed," *Quality Engineering*, vol. 30, no. 4, pp. 594–609, 2018. DOI: 10.1080/08982112.2018.1428751.

[57] A. Folch-Fortuny, F. Arteaga, and A. Ferrer, "PLS model building with missing data: New algorithms and a comparative study," *Journal of Chemometrics*, vol. 31, no. 7, pp. 1–12, 2017, ISSN: 1099128X. DOI: 10.1002/cem.2897.

[58] I. S. Helland, "Partial Least Squares Regression and Statistical Models," *Scandinavian Journal of Statistics*, vol. 17, pp. 97–114, 1990, ISSN: 17549469.

[59] H. Martens, L. Izquierdo, M. Thomassen, and M. Martens, "Partial least-squares regression on design variables as an alternative to analysis of variance," *Analytica Chimica Acta*, vol. 191, pp. 133–148, 1986, ISSN: 00032670. DOI: 10.1016/S0003-2670(00)86303-5.

[60] M. Zarzo and A. Ferrer, "Batch process diagnosis: PLS with variable selection versus block-wise PCR," *Chemometrics and Intelligent Laboratory Systems*, vol. 73, no. 1 SPEC. ISS. Pp. 15–27, 2004, ISSN: 01697439. DOI: 10.1016/j.chemolab.2003.11.009.

[61] R. V. Lenth, "Quick and easy analysis of unreplicated factorials," *Technometrics*, vol. 31, no. 4, pp. 469–473, 1989, ISSN: 15372723. DOI: 10.1080/00401706.1989.10488595.

[62] B. Efron and G. Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983, ISSN: 00295493. DOI: 10.2514/6.2008-1716. arXiv: arXiv:1011.1669v3.

[63] H. Martens and M. Martens, "Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)," *Food Quality and Preference*, vol. 11, no. 1-2, pp. 5–16, 2000, ISSN: 09503293. DOI: 10.1016/s0950-3293(99)00039-7.

[64] H. Martens, M. Høy, F. Westad, D. Folkenberg, and M. Martens, "Analysis of designed experiments by stabilised PLS regression and jack-knifing," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 151–170, 2001, ISSN: 01697439. DOI: 10.1016/S0169-7439(01)00157-5.

[65] R. S. Kenett, E. Rahav, and D. M. Steinberg, "Bootstrap analysis of designed experiments," *Quality and Reliability Engineering International*, vol. 22, no. 6, pp. 659–667, 2006, ISSN: 07488017. DOI: 10.1002/qre.802.

[66] T. N. Goh, "Economical Experimentation via 'Lean Design'," *Quality and Reliability Engineering International*, vol. 12, pp. 383–388, 1996. DOI: 10.1002/0470062002.ch19.

[67] C. R. Kaplan and J. W. Gentry, "Use of Condition Numbers for Shortcut Experimental Design," *AIChE Journal*, vol. 33, no. 4, pp. 681–685, 1987, ISSN: 15475905. DOI: 10.1002/aic.690330418.

[68] G. E. P. Box, "A Simple way to deal with missing observations from DOE," *Quality Engineering*, vol. 20, pp. 443–458, 1990.

[69] D. A. Belsey, "Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise," *Journal of Econometrics*, vol. 20, pp. 211–253, 1982.

[70] J. Camacho and A. Ferrer, "Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects," *Chemometrics and Intelligent Laboratory Systems*, vol. 131, pp. 37–50, 2014, ISSN: 01697439. DOI: 10.1016/j.chemolab.2013.12.003.

[71] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019, ISBN: 9780471183860. DOI: 10.1002/9781119013563.

[72] M. F. Rempel and J. Zhou, "On exact K-optimal designs minimizing the condition number," *Communications in Statistics - Theory and Methods*, vol. 43, no. 6, pp. 1114–1131, 2014, ISSN: 03610926. DOI: 10.1080/03610926.2012.670352.

[73] J. A. Cornell, *Experiments with Mixtures*, 2n ed. New York: John Willey & Sons, 1990.

[74] N. Kettaneh-Wold, "Analysis of mixture data with partial least squares," *Chemometrics and Intelligent Laboratory Systems*, vol. 14, no. 1-3, pp. 57–69, 1992, ISSN: 01697439. DOI: 10.1016/0169-7439(92)80092-I.

[75] L. Eriksson, E. Johansson, and C. Wikström, "Mixture design - Design generation, PLS analysis, and model usage," *Chemometrics and Intelligent Laboratory Systems*, vol. 43, no. 1-2, pp. 1–24, 1998, ISSN: 01697439. DOI: 10.1016/S0169-7439(98)00126-9.

[76] R. Vitale, D. Palací-López, H. H. Kerkenaar, G. J. Postma, L. M. Buydens, and A. Ferrer, "Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of ex-

periments," *Chemometrics and Intelligent Laboratory Systems*, vol. 175, no. May 2017, pp. 37–46, 2018, ISSN: 18733239. DOI: 10.1016/j.chemolab.2018.02.002.

[77] H. Wold, "Soft modelling: The basic design and some extensions," *Systems Under Indirect Observation, Part II*, 1982.

[78] C. Duchesne and J. F. MacGregor, "Establishing Multivariate Specification Regions for Incoming Materials," *Journal of Quality Technology*, vol. 36, no. 1, pp. 78–94, 2004, ISSN: 0022-4065. DOI: 10.1080/00224065.2004.11980253.

[79] J. A. De Smet, "Development of Multivariate Specification Limits Using Partial Least Squares Regression," Ph.D. dissertation, McMaster University, Hamilton, Ontario, Canada, 1993.

[80] S. García-Muñoz, S. Dolph, and H. W. Ward, "Handling uncertainty in the establishment of a design space for the manufacture of a pharmaceutical product," *Computers and Chemical Engineering*, vol. 34, no. 7, pp. 1098–1107, 2010, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2010.02.027.

[81] S. García-Muñoz, "Establishing multivariate specifications for incoming materials using data from multiple scales," *Chemometrics and Intelligent Laboratory Systems*, vol. 98, no. 1, pp. 51–57, 2009, ISSN: 01697439. DOI: 10.1016/j.chemolab.2009.04.008.

[82] J. F. MacGregor, Z. Liu, M. J. Bruwer, B. Polsky, and G. Visscher, "Setting simultaneous specifications on multiple raw materials to ensure product quality and minimize risk," *Chemometrics and Intelligent Laboratory Systems*, vol. 157, pp. 96–103, 2016, ISSN: 18733239. DOI: 10.1016/j.chemolab.2016.06.021.

[83] A. Paris, C. Duchesne, and É. Poulin, "Establishing Multivariate Specification Regions for Incoming Raw Materials Using Projection to Latent Structure Models: Comparison Between Direct Mapping and Model Inversion," *Frontiers in Analytical Science*, vol. 1, no. November, pp. 1–15, 2021. DOI: 10.3389/frans.2021.729732.

[84] P. Facco, F. Dal Pastro, N. Meneghetti, F. Bezzo, and M. Barolo, "Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development," *Industrial & Engineering Chemistry Research*, vol. 54, no. 18, pp. 5128–5138, 2015, ISSN: 15205045. DOI: 10.1021/acs.iecr.5b00863.

[85] J. J. Peterson and M. Yahyah, "A Bayesian Design Space Approach to Robustness and System Suitability for Pharmaceutical Assays and Other Processes," *Statistics in Biopharmaceutical Research*, vol. 1, no. 4, pp. 441–449, 2009, ISSN: 1946-6315. DOI: 10.1198/sbr.2009.0037.

[86] G. Bano, P. Facco, F. Bezzo, and M. Barolo, "Probabilistic Design space determination in pharmaceutical product development: A Bayesian/latent variable approach," *AIChE Journal*, vol. 64, no. 7, pp. 2438–2449, 2018, ISSN: 15475905. DOI: 10.1002/aic.16133.

[87] E. del Castillo and M. S. Reis, "Bayesian predictive optimization of multiple and profile response systems in the process industry: A review and extensions," *Chemometrics and Intelligent Laboratory Systems*, vol. 206, no. June, p. 104 121, 2020, ISSN: 18733239. DOI: 10.1016/j.chemolab.2020.104121.

[88] E. Rozet, P. Lebrun, B. Debrus, B. Boulanger, and P. Hubert, "Design Spaces for analytical methods," *Trends in Analytical Chemistry*, vol. 42, pp. 157–167, 2013, ISSN: 18793142. DOI: 10.1016/j.trac.2012.09.007.

[89] S. García-Muñoz and J. Mercado, "Optimal Selection of Raw Materials for Pharmaceutical Drug Product Design and Manufacture using Mixed Integer Nonlinear Programming and Multivariate Latent Variable Regression Models," *Industrial & Engineering Chemistry Research*, vol. 52, pp. 5934–5942, 2013. DOI: 10.1021/ie3031828.

[90] S. Wold, M. Josefson, J. Gottfries, and A. Linusson, "The utility of multivariate design in PLS modeling," *Journal of Chemometrics*, vol. 18, no. 34, pp. 156–165, 2004, ISSN: 0886-9383. DOI: 10.1002/cem.861.

[91] R. S. Kenett, C. Gotwalt, L. Freeman, and X. Deng, "Self-supervised cross validation using data generation structure," *Applied Stochastic Models in Business and Industry*, vol. 38, no. 5, pp. 750–765, 2022, ISSN: 15264025. DOI: 10.1002/asmb.2701.

[92] J. F. MacGregor and M.-J. Bruwer, "A Framework for the Development of Design and Control Spaces," *Journal of Pharmaceutical Innovation*, vol. 3, no. 1, pp. 15–22, 2008, ISSN: 1872-5120. DOI: 10.1007/s12247-008-9023-5.

[93] T. Kourti and J. F. MacGregor, "Process analysis, monitoring and diagnosis, using multivariate projection methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 28, pp. 3–21, 1995, ISSN: 01697439. DOI: 10.1016/0169-7439(95)80036-9.

[94] C. Duchesne and J. F. MacGregor, "Multivariate analysis and optimization of process variable trajectories for batch processes," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, pp. 125–137, 2000.

[95] D. De-Felipe and E. Benedito, "A review of univariate and multivariate process capability indices," *International Journal of Advanced Manufacturing Technology*, vol. 92, no. 5-8, pp. 1687–1705, 2017, ISSN: 14333015. DOI: 10.1007/s00170-017-0273-6.

[96] A. Ferrer, "Latent structures-based multivariate statistical process control: A paradigm shift," *Quality Engineering*, vol. 26, no. 1, pp. 72–91, 2014, ISSN: 08982112. DOI: 10.1080/08982112.2013.846093.

[97] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, no. 3, pp. 403–414, 1995, ISSN: 09670661. DOI: 10.1016/0967-0661(95)00014-L.

[98] *WHO Coronavirus (COVID-19) Dashboard* /, https://covid19.who.int/, 2023.

[99] M. Ciotti, M. Ciccozzi, A. Terrinoni, W. C. Jiang, C. B. Wang, and S. Bernardini, "The COVID-19 pandemic," *Critical Reviews in Clinical Laboratory Sciences*, pp. 365–388, 2020, ISSN: 1549781X. DOI: 10.1080/10408363.2020.1783198.

[100] E. Burn *et al.*, "The natural history of symptomatic COVID-19 during the first wave in Catalonia," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021, ISSN: 20411723. DOI: 10.1038/s41467-021-21100-y.

[101] J. N. Gustine and D. Jones, "Immunopathology of Hyperinflammation in COVID-19," *American Journal of Pathology*, vol. 191, no. 1, pp. 4–17, 2021, ISSN: 15252191. DOI: 10.1016/j.ajpath.2020.08.009.

[102] S. R. Knight *et al.*, "Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score," *BMJ*, vol. 370, 2020, ISSN: 17561833. DOI: 10.1136/bmj.m3339.

[103] Geoffrey J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition," in *Wiley Series in Probability and Statistics*, Hoboken, NJ, USA: John Wiley & Sons, Inc, 1992, ISBN: 9780471725299. DOI: 10.1002/0471725293.

[104] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003, ISSN: 08869383. DOI: 10.1002/cem.785.

[105] B. Schölkopf and A. J. Smola, *Learning with Kernels support vector machines, regularization, optimization, and beyond*. The MIT Press, 2018, ISBN: 9780262536578.

[106] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[107] S. Wold, N. Kettaneh-Wold, and B. Skagerberg, "Nonlinear PLS Modeling," *Chemometrics and Intelligent Laboratory Systems*, vol. 7, pp. 53–65, 1989.

[108] R. Rosipal, "Nonlinear Partial Least Squares : An Overview," *Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques*, pp. 169–189, 2011. DOI: 10.4018/978-1-61520-911-8.ch009.

[109] S. J. Pocck, *Clinical Trials: A Practical Approach*. John Wiley & Sons, 2013, ISBN: 9781118793916. DOI: 10.1002/9781118793916.

[110] E. Alvarez, F. Lamagna, M. Szewc, C. Miguelete, and S. C. D. Bariloche, "A Machine Learning alternative to placebo-controlled clinical trials upon new diseases : A primer," 2020. arXiv: arXiv:2003.12454v1.

[111] J. A. Westerhuis *et al.*, "Assessment of PLSDA cross validation," *Metabolomics*, vol. 4, pp. 81–89, 2008. DOI: 10.1007/s11306-007-0099-6.

[112] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24–38, 2006, ISSN: 01697439. DOI: 10.1016/j.chemolab.2005.05.004.

[113] K. Linderman, R. G. Schroeder, S. Zaheer, and A. S. Choo, "Six Sigma: A goal-theoretic perspective," *Journal of Operations Management*, vol. 21, no. 2, pp. 193–203, 2003, ISSN: 02726963. DOI: 10.1016/S0272-6963(02)00087-6.

[114] A. Ismail, S. B. Mohamed, H. Juahir, M. E. Toriman, and A. Kassim, "DMAIC Six Sigma Methodology in Petroleum Hydrocarbon Oil Classification," *International Journal of Engineering & Technology*, vol. 7, no. July, pp. 98–106, 2018. DOI: 10.14419/ijet.v7i3.14.16868.

[115] R. Santana Peruchi, P. Rotela Junior, T. G. Brito, A. P. Paiva, P. P. Balestrassi, and L. M. Mendes Araújo, "Integrating Multivariate Statistical Analysis Into Six Sigma DMAIC Projects : A Case Study on AISI 52100 Hardened Steel Turning," *IEEE Access*, vol. 8, pp. 1–10, 2020. DOI: 10.1109/ACCESS.2020.2973172.

[116] J. De Mast and J. Lokkerbol, "An analysis of the Six Sigma DMAIC method from the perspective of problem solving," *International Journal of Production Economics*, vol. 139, no. 2, pp. 604–614, 2012, ISSN: 09255273. DOI: 10.1016/j.ijpe.2012.05.035.