Ph.D. Dissertation

Quality by Design through multivariate latent structures

Author
Daniel Gonzalo Palací López

Ph.D. Supervisors
Alberto J. Ferrer-Riquelme
Massimiliano Barolo

A doctoral thesis submitted to
Department of Applied Statistics, Operational Research, and Quality

Valencia, December 2018

A los iaios

Acknowledgements

During the course of the past five years, many have been the challenges I found both in life and for the development of what is the present Ph.D. thesis. But just as numerous, if not more, have been the people I met who helped me get to where I am now. For this, not few are those I would like to thank, and thus I will attempt to mention here as many of them as I can.

First of all to my mentor and supervisor Alberto, for showing me the importance of understanding and properly applying statistics in many aspects of life even before starting this project, but also for teaching me the way to make the best out of my own strengths without being overconfident, and to recognize my own weaknesses as an opportunity for improvement. Thank you for your patience, for making my thirst for knowledge grow without leaving me to starve, and for that healthy blend of criticism, scepticism and confidence you prepared for us every time we met to discuss any given issue. And thank you, of course, for making the office feel like a second home.

To my supervisor beyond the Mediterranean Sea, Max, because you received me and made me feel like another member of the CAPE-Lab, and even your family, while I was so far away from home (ok, maybe not *that* far away), and to Pier, for pretty much the same reason. A good deal of this thesis is what it is now thanks to your supervision, your support, and your input. Thank you, too, for not kicking me out of the department that one day I brought in some poorly cooked spaghetti with almost expired meatballs. And thank you, too, to Pierantonio's parents for their hospitality and the most wonderful lasagne I have ever tried (sorry, mum). Thanks to Andrea and Myriam, who helped me organize the trip before my first stay in Padova, and to Amir, Christopher, Fabrizio, Federico (both of them), Filippo, Gabriele, Marco, Martino, Natascia and Riccardo, with most of whom I have shared many meals and from whom I have learnt a lot of things, including what *real* pizza is like (in spite of Riccardo's attempts to bring me to the dark side of Neapolitan pizza). Special mention to Mariana, Laura, Masa and Qong, who helped me have a more varied an healthy diet during my second stay, and made coming back from work much more enjoyable that it would have been without you.

To all of the people in Leuven who made this charmingly small city look so much bigger and warmer than it may look like from the outside (I am sure those 32ºC by the end of August were just a coincidence), but most of all to Peter who, in addition, contributed greatly to my current knowledge regarding mixture design and optimal design of experiments, and forgave me for being unable to properly pronounce his last name. Thanks also to José and Wannes, for being such wonderful officemates, and also to Mito. All of you showed me the most interesting parts of the city, its history, and where to get some nice ice cream or go for a drink with you.

To those I would call my second family in the department: to José Manuel, for bearing with me and my mistimed ramblings about *N*-dimensional spaces and intersections of *r*-dimensional elements within them, among others, and Pili, because you were always there to listen, help and provide advice when needed; to José María who, in spite of your busy schedule, always managed to make some time to ask how things were doing, and to give some constructive feedback; to Abel and Raffa, who managed to put me at ease almost as easily as one can breath, and for your immeasurable help in surviving the bureaucracy I have had to face, but most of all for being who and where you were; to Eric, for the trips we've made together and all those meetings trying to fix that webpage that shall not be named; to Pedro, because of that one question that made me reconsider how to approach almost half of this thesis, and all the answers you, maybe unknowingly, gave me; to Antonio and Borja, for allowing me to try and explain to you things I myself did not fully understand at the moment, so that I could learn them; to Joan, for demanding proof, and proving me wrong, and by doing so making me improve; and to Alba, because you do not like sugar in her coffee, but you still gets one packet everyday because you know I like my sugar coffeeless.

To my friends from Chemical Engineering, Javi, Leo, Maite, Manu, María and Ruth, for being there when I needed you, and lending me an ear or a roof; to those from school, Ana, Miguel, Miguel Ángel, Patricia and Vicente, for those times you fully believed in me, and those you did not, and offered help, just in case; to Alejandro and Sara, because some things are not meant to last forever, even with the best of intentions; to Armunia, Borja and Jenny who, while far away (on foot, at least), will always have a place very close to my heart; to Eugenio, because bearing with Miguel is easier when you are close by, and Luis, my whitest African friend to date; and to all those whose names do not appear here, and should, who have come and gone.

To the people at the electrochemical laboratory (at least according to the sign outside): Enrique, Inma, Juan, José Manuel and Manuel. I hope we will be able to stay in contact for many years to come, as meeting you and working by your side, as little as it may be, is always a pleasure for me. To Maria José, José Manuel Serra and Xesc, for having motivated me, one way or another, to start this journey before I myself knew such road existed, and to Igor, who taught me to see beyond the third (and fourth) dimension.

To mum and dad, and to my brother: this, and many other journeys, I would have never been able to start nor finish without you. You deserve much more than this single line, but I do not intend to write a second thesis here, and I doubt the youngest of you would be patient enough to read it. To Juanjo, Beatriz and Ale, and to the rest of members of my family, for I have learnt many things from each of you, and I hope to learn many more in the future; to Javier and Carmen, because it is not just blood that makes a family; to Laura and Nerea, who just recently became a part of my life, for your presence beside those I care the most about let me sleep tight at night; to Chancla and Bota, because they cannot read this, but they know just how much they can brighten my days in the greyest of afternoons, and to Zuri, because your silliness is hard not to smile at.

To my grandpas, for they have been present during my whole life and, although they had to leave for a while, I am sure they will keep an eye on us to see the end of this journey.

Agradecimientos

A lo largo de los últimos cinco años, muchos han sido los retos que se me han planteado, tanto en mi vida personal como en el desarrollo de esta tesis. Pero tantas, si no más, han sido las personas que he conocido y que me han permitido llegar hasta donde estoy hoy. No a pocos he de dar las gracias por todo ello, y por tanto trataré de mencionar aquí a tantos como me sea posible.

En primer lugar a mi mentor y supervisor Alberto, por mostrarme la importancia de comprender y aplicar correctamente la estadística en muchos aspectos de la vida, incluso antes de haber comenzado este proyecto, pero también por enseñarme a aprovechar al máximo mis puntos fuertes, sin caer en la arrogancia, y a reconocer mis propias debilidades como una oportunidad para mejorar. Gracias por tu paciencia, por hacer que quiera aprender cada vez más, sin dejar de enseñarme, y por esa magistral mezcla de crítica, escepticismo y confianza que preparabas antes abordar juntos cualquier problema. Y muchas gracias, por supuesto, por hacer del despacho un segundo hogar.

A mi supervisor al otro lado del Mediterráneo, Max, porque me recibiste y me hiciste sentir como uno más del CAPE-Lab, e incluso parte de tu familia, cuando estaba tan lejos de casa (bueno, no *tan* lejos), y a Pier, por más de lo mismo. Buena parte de esta tesis no sería lo que es ahora si no fuese por vuestra supervisión, vuestro apoyo y vuestros consejos. Gracias también por no haberme echado del departamento aquel día que llevé un táper de espaguetis blandos con albóndigas casi caducadas. Y muchas gracias, también, a los padres de Pierantonio, por su hospitalidad y la lasaña más deliciosa que he probado jamás (lo siento, mamá). Gracias a Andrea y a Myriam, por haberme ayudado a organizar mi primer viaje a Padua, y a Amir, Christopher, Fabrizio, Federico (a los dos), Filippo, Gabriele, Marco, Martino, Natascia y Riccardo, con quienes he compartido numerosas comidas, y de los que he aprendido muchas cosas, incluyendo cómo es una pizza *de verdad* (a pesar de los intentos de Riccardo de llevarme al lado oscuro de la pizza napolitana). Especial mención a Mariana, Laura, Masa y Qong, pues me ayudasteis a tener una dieta más sana y variada durante mi segunda estancia, e hicisteis las vueltas del trabajo mucho más amenas de lo que habrían sido sin vosotros.

A todos aquellos en Lovaina que convirtieron a esta pequeña y encantadora ciudad en un lugar mucho más grande y cálido de lo que aparenta desde el exterior (estoy seguro de que esos 32ºC hacia finales de Agosto sólo fueron una coincidencia), pero especialmente a Peter, pues no sólo me enseñaste buena parte de lo que ahora sé sobre diseño de mezclas y de experimentos, sino que también pudiste perdonarme por no saber pronunciar tu apellido. Gracias también a José y Wannes, por ser excelentes compañeros de despacho, y a Mito. Entre todos me mostrasteis los lugares más interesantes de la ciudad, su historia, y dónde tomar un buen helado o beber algo a vuestro lado.

A aquellos a los que me atrevería a llamar mi segunda familia en el departamento: a José Manuel, por soportarnos a mí y a mis charlas inoportunas sobre, entre otras cosas, espacios *N*-dimensionales e intersecciones de elementos *r*-dimensionales dentro de ellos, y a Pili, porque siempre habéis estado ahí para escuchar, ayudar y aconsejar si hacía falta; a José María, pues tu apretada agenda no te ha impedido pasarte de vez en cuando a preguntar cómo van las cosas y aportar sugerencias; a Abel y Raffa, por saber tranquilizarme cuando lo necesitaba y por ayudarme a sobrevivir a la burocracia que se ha puesto en mi camino, pero sobre todo por ser como sois; a Eric, por los viajes que hemos hecho juntos, y por aquellas reuniones para intentar arreglar la página web que no debe ser nombrada; a Pedro, por esa pregunta que me hizo replantearme casi media tesis, y por las dudas que, tal vez sin saberlo, resolviste; a Antonio y Borja, por haberme permitido que intentara explicaros cosas que yo mismo no entendía en su momento y, con ello, comprenderlas; a Joan, por exigir pruebas, demostrar lo equivocado que estaba, y permitirme mejorar al hacerlo; y a Alba, porque prefieres el café sin azúcar, y aun así siempre coges un sobre porque sabes que no me gusta echarle café a mi azúcar.

A mis amigos de Ingeniería Química, Javi, Leo, Maite, Manu, María y Ruth, por estar ahí cuando lo necesité, para prestarme vuestro oído o tejado; para los del colegio, Ana, Miguel, Miguel Ángel, Patricia y Vicente, por aquellos momentos en que creísteis plenamente en mí, y esos otros que no, y me ofrecisteis ayuda, sólo por si acaso; a Alejandro y Sara, porque algunas cosas no están hechas para durar eternamente, incluso con la mejor de las intenciones; a Armunia, Borja y Jenny, pues aun lejos (andando al menos) siempre tendréis un lugar muy cerca de mi corazón; a Eugenio, porque soportar a Miguel contigo cerca es más fácil, y a Luis, mi amigo africano más blanco hasta la fecha.

A la gente del laboratorio de electroquímica (al menos de acuerdo con el cartel junto a la puerta): Enrique, Inma, Juan, José Manuel y Manuel. Espero que nunca perdamos el contacto, pues quedar con vosotros para hablar o trabajar, por poco que sea, es siempre un placer. A María José, José Manuel Serra y Xesc, por haberme motivado, de un modo u otro, a comenzar este viaje, antes incluso de que yo mismo supiera que el camino existía; y a Igor, por enseñarme a ver más allá de la tercera (y cuarta) dimensión.

A mamá y a papá, y al tete: éste, y muchos otros viajes, habrían quedado sin empezar ni terminar si no hubiese sido por vosotros. Merecéis mucho más que una frase, pero no tengo intención de escribir una segunda tesis, y sospecho que uno de vosotros no tendría la paciencia suficiente como para leerla. A Juanjo, Beatriz y Ale, y al resto de mi familia, pues he aprendido mucho de vosotros, y espero seguir haciéndolo en el futuro; a Javier y a Carmen, pues no es sólo la sangre lo que define a una familia; a Laura y Nerea, que hace poco pasaron a formar parte de mi vida, pues vuestra presencia junto a aquellos que más quiero me permite dormir tranquilo por las noches; a Chancla y Bota, porque no pueden leer esto, pero saben bien cuánto han iluminado mis días incluso en las tardes más grises, y a Zuri, porque es difícil no sonreír con sus payasadas.

A los iaios, pues han estado presentes toda mi vida y, aunque se han visto obligaos a ausentarse por un tiempo, estoy seguro de que están atentos al final de este viaje.

Abstract

The present Ph.D. thesis is motivated by the growing need in most companies, and specially (but not solely) those in the pharmaceutical, chemical, food and bioprocess fields, to increase the flexibility in their operating conditions in order to reduce production costs while maintaining or even improving the quality of their products. To this end, this thesis focuses on the application of the concepts of the *Quality by Design* for the exploitation and development of already existing methodologies, and the development of new algorithms aimed at the proper implementation of tools for the design of experiments, multivariate data analysis and process optimization, specially (but not only) in the context of mixture design.

Part I – Preface, where a summary of the research work done, the main goals it aimed at and their justification, are presented. Some of the most relevant concepts related to the developed work in subsequent chapters are also introduced, such as those regarding design of experiments or latent variable-based multivariate data analysis techniques.

Part II – Mixture design optimization, in which a review of existing mixture design tools for the design of experiments and data analysis via traditional approaches, as well as some latent variable-based techniques, such as Partial Least Squares (PLS), is provided. A kernel-based extension of PLS for mixture design data analysis is also proposed, and the different available methods are compared to each other. Finally, a brief presentation of the software MiDAs is done. MiDAs has been developed in order to provide users with a tool to easily approach mixture design problems for the construction of Designs of Experiments and data analysis with different methods and compare them.

Part III – Design Space and optimization through the latent space, where one of the fundamental issues within the *Quality by Design* philosophy, the definition of the so-called 'design space' (i.e. the subspace comprised by all possible combinations of process operating conditions, raw materials, etc. that guarantee obtaining a product meeting a required quality standard), is addressed. The problem of properly defining the optimization problem is also tackled, not only as a tool for quality improvement but also when it is to be used for exploration of process flexibilisation purposes, in order to establish an efficient and robust optimization method in accordance with the nature of the different problems that may require such optimization to be resorted to.

Part IV – Epilogue, where final conclusions are drawn, future perspectives suggested, and annexes are included.

Resumen

La presente tesis doctoral surge ante la necesidad creciente por parte de la mayoría de empresas, y en especial (pero no únicamente) aquellas dentro de los sectores farmacéutico, químico, alimentación y bioprocesos, de aumentar la flexibilidad en su rango operativo para reducir los costes de fabricación, manteniendo o mejorando la calidad del producto final obtenido. Para ello, esta tesis se centra en la aplicación de los conceptos del *Quality by Design* para la aplicación y extensión de distintas metodologías ya existentes y el desarrollo de nuevos algoritmos que permitan la implementación de herramientas adecuadas para el diseño de experimentos, el análisis multivariante de datos y la optimización de procesos en el ámbito del diseño de mezclas, pero sin limitarse exclusivamente a este tipo de problemas.

Parte I – Prefacio, donde se presenta un resumen del trabajo de investigación realizado y los objetivos principales que pretende abordar y su justificación, así como una introducción a los conceptos más importantes relativos a los temas tratados en partes posteriores de la tesis, tales como el diseño de experimentos o diversas herramientas estadísticas de análisis multivariado.

Parte II – Optimización en el diseño de mezclas, donde se lleva a cabo una recapitulación de las diversas herramientas existentes para el diseño de experimentos y análisis de datos por medios tradicionales relativos al diseño de mezclas, así como de algunas herramientas basadas en variables latentes, tales como la Regresión en Mínimos Cuadrados Parciales (PLS). En esta parte de la tesis también se propone una extensión del PLS basada en *kernels* para el análisis de datos de diseños de mezclas, y se hace una comparativa de las distintas metodologías presentadas. Finalmente, se incluye una breve presentación del programa MiDAs, desarrollado con la finalidad de ofrecer a sus usuarios la posibilidad de comparar de forma sencilla diversas metodologías para el diseño de experimentos y análisis de datos para problemas de mezclas.

Parte III – Espacio de diseño y optimización a través del espacio latente, donde se aborda el problema fundamental dentro de la filosofía del *Quality by Design* asociado a la definición del llamado 'espacio de diseño', que comprendería todo el conjunto de posibles combinaciones de condiciones de proceso, materias primas, etc. que garantizan la obtención de un producto con la calidad deseada. En esta parte también se trata el problema de la definición del problema de optimización como herramienta para la mejora de la calidad, pero también para la exploración y flexibilización de los procesos productivos, con el objeto de definir un procedimiento eficiente y robusto de optimización que se adapte a los diversos problemas que exigen recurrir a dicha optimización.

Parte IV – Epílogo, donde se presentan las conclusiones finales, la consecución de objetivos y posibles líneas futuras de investigación. En esta parte se incluyen además los anexos.

Resum

Aquesta tesi doctoral sorgeix davant la necessitat creixent per part de la majoria d'empreses, i especialment (però no únicament) d'aquelles dins dels sectors farmacèutic, químic, alimentari i de bioprocessos, d'augmentar la flexibilitat en el seu rang operatiu per tal de reduir els costos de fabricació, mantenint o millorant la qualitat del producte final obtingut. La tesi se centra en l'aplicació dels conceptes del *Quality by Design* per a l'aplicació i extensió de diferents metodologies ja existents i el desenvolupament de nous algorismes que permeten la implementació d'eines adequades per al disseny d'experiments, l'anàlisi multivariada de dades i l'optimització de processos en l'àmbit del disseny de mescles, però sense limitar-se exclusivament a aquest tipus de problemes.

Part I– Prefaci, en què es presenta un resum del treball de recerca realitzat i els objectius principals que pretén abordar i la seua justificació, així com una introducció als conceptes més importants relatius als temes tractats en parts posteriors de la tesi, com ara el disseny d'experiments o diverses eines estadístiques d'anàlisi multivariada.

Part II – Optimització en el disseny de mescles, on es duu a terme una recapitulació de les diverses eines existents per al disseny d'experiments i anàlisi de dades per mitjans tradicionals relatius al disseny de mescles, així com d'algunes eines basades en variables latents, tals com la Regressió en Mínims Quadrats Parcials (PLS). En aquesta part de la tesi també es proposa una extensió del PLS basada en *kernels* per a l'anàlisi de dades de dissenys de mescles, i es fa una comparativa de les diferents metodologies presentades. Finalment, s'inclou una breu presentació del programari MiDAs, que ofereix la possibilitat als usuaris de comparar de forma senzilla diverses metodologies per al disseny d'experiments i l'anàlisi de dades per a problemes de mescles.

Part III– Espai de disseny i optimització a través de l'espai latent, on s'aborda el problema fonamental dins de la filosofia del *Quality by Design* associat a la definició de l'anomenat 'espai de disseny', que comprendria tot el conjunt de possibles combinacions de condicions de procés, matèries primeres, etc. que garanteixen l'obtenció d'un producte amb la qualitat desitjada. En aquesta part també es tracta el problema de la definició del problema d'optimització com a eina per a la millora de la qualitat, però també per a l'exploració i flexibilització dels processos productius, amb l'objecte de definir un procediment eficient i robust d'optimització que s'adapti als diversos problemes que exigeixen recórrer a aquesta optimització.

Part IV– Epíleg, on es presenten les conclusions finals i la consecució d'objectius i es plantegen possibles línies futures de recerca arran dels resultats de la tesi. En aquesta part s'inclouen a més els annexos.

Contents

PART I

Preface

Chapter 1
Justification, objectives and contributions

1.1. Justification

The present Ph.D. thesis aims at providing better insight and novel algorithmic tools to address two main fields of interest, namely that concerning what are known as mixture design problems, as well as the problematic associated to the optimization of production processes through the so-called Quality by Design philosophy, which is originated by the need to guarantee the desired quality of a given product before its production while allowing as much flexibility as possible in the setting of the processing conditions, the selection of raw materials, etc.

Particularly, the present Ph.D. thesis project addresses some of the issues of practical interest in regards to both mixture design problems and process optimization via data-driven (i.e. empirical) approaches, such as:

1. efficient experimental design required given a particular purpose (e.g. data-driven exploration, prediction or optimization…) and availability of resources (e.g. time, costs, equipment…), particularly for mixture design problems;
2. comparison of data-driven model building strategies for explanatory, predictive and optimization purposes, specially for mixture design problems, with both data from a statistical Design of Experiments and historical data;
3. sensible definition of the different sets of processing conditions that allow meeting the products specifications;
4. data-driven process optimization.

Addressing these points will, potentially, guarantee:

1. reduced costs (by e.g. minimizing required experimentation, the amount of production outside of specifications, or optimizing the processing conditions);
2. increased process flexibility;
3. more consistent product.

1.2. Objectives of the thesis

In this section a more thorough description of the objectives of this Ph.D. thesis and the proposals to accomplish them is provided.

1.2.1. Traditional and latent variable-based approaches applied to mixture design problems

With a wide range of products currently used in daily life resulting from processing blends of two or more ingredients, where the properties of these products mainly depend on the raw materials being mixed and on the proportions in which they are added, mixture design problems are of great relevance. However, the special nature of these problems requires special modifications to be made to traditional approaches aimed at selecting which experimentation to perform in order to construct reliable data-based models. These are needed in most real scenarios nowadays given the large amounts of data being made available continuously and the complexity of most processes nowadays, for which the use of first principles models may be unfeasible. Furthermore, the structure and interpretation of the models traditionally selected to be fitted in mixture design also require special reformulations and tools, and even then these methods may suffer whenever a process is characterised by the presence of strong non-linear behaviours and complex relationships among the variables involved.

Part II of this Ph.D. thesis will serve as a review of some of the most widely used traditional approaches to mixture design, concerning both the experimental design and the model definition and fitting steps (Chapter 5). As an alternative, latent variable-based approaches will also be presented and compared to traditional approaches in order to assess the advantages and limitations of each of them (Chapter 6). A Matlab-powered software with graphic interface has been developed to allow users without knowledge in Matlab or programming to make use of most the tools illustrated in previous chapters in Chapter 7.

1.2.2. Latent variable-based approaches for efficient processes optimization

Extensive work has been done in the past for the optimization of production processes through the use of a wide range of tools. Some of the most recent one has been focused on the use of latent variable-based techniques in order to take advantage of the increasingly large amount of data continuously being stored regarding the processes and products on which the optimization is intended to be applied, due to the several advantages that latent variable-based methods offer in this context. However, some of the proposed tools present a series of drawbacks or aspects that leave some room for improvement, such as the lack of a standardized procedure for the optimization of quality attributes expressed as linear combinations of process outputs, the steps to follow when solving a minimization/maximization problem while avoiding involuntarily assigning too much/too little importance to certain optimization criteria, or the use of excessively complex optimization tools when simpler ones will provide equally good results with fewer risks and computational cost.

In Part III, Chapter 10, some of these issues are addressed, and in Chapter 11 the application of the proposed algorithms is illustrated.

1.2.3. Latent variable-based approaches applied to the Quality by Design initiative, to increase processes flexibility and guarantee the desired quality

Although slightly related to the optimization approach, one other topic of increasing relevance in the last years has been that regarding the philosophy of the so-called Quality by Design and the way it should be applied in practice. In both cases, the main objective is being able to obtain combinations of processing operating conditions and raw material properties that guarantee the desired results, with the subspace comprised by all possible combinations being frequently referred to as the Design Space. As in the optimization field, latent variable-based methods have been shown to be powerful tools when trying to accurately define this so-called Design Space. However, not much insight has been offered until recent years regarding precisely the explicit definition of such subspace of operating conditions, nor in how the accuracy in its estimation could be improved.

Chapter 9 in Part III of this Ph.D. thesis aims at providing some novel formulations of the Design Space, the subspace of process conditions most likely to contain it, and the subspace of process conditions least likely to not be fully contained by it (i.e. most likely to guarantee that the quality attributes of interest of a given product will meet their corresponding specifications). In Chapter 12 several of the concepts addressed in Chapters 9 to 11 will be resorted to in order to propose a novel sequential experimental approach to improve the accuracy in the estimation of the Design Space.

1.3. Contributions

Here is a list of all the contributions authored by the candidate during the progress of this Ph.D. thesis:

Peer-reviewed publications

1. Vitale, R., Palací-López, D., Kerkenaar, H., Postma, G., Buydens, L. & Ferrer, A. Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments. *Chemometr. Intell. Lab.* 175, 37-46 (2018).
2. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Sequential experimental approach to improve the design space estimation using latent-variable model inversion. Part I. Defining the Experimental Region. *Submitted.*
3. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Sequential experimental approach to improve the design space estimation using latent-variable model inversion. Part II. Optimization problem reformulation and sequential experimental approach. *Submitted.*

Poster communications in conferences

1. Palací-López, D. & Ferrer, A. DIseño y Análisis de Mezclas: ¿mezclado, no agitado? *Jornada de Divulgació i Aplicació de l'Estadística - JDAE 2013,* 25/9/2013, Valencia, Spain.
2. Palací-López, D. & Ferrer, A. DyAM: Software tool for mixture optimization. *3rd European conference of Process Analytics and Control Technology - EUROPACT 2014*, 09/05/2014, Barcelona, Spain.
3. Palací-López, D. & Ferrer, A. Quality by Design a través de estructuras latentes. *I Encuentro de Estudiantes de Doctorado de la Universitat Politècnica de València*, 12/06/2014, Valencia, Spain.
4. Palací-López, D. & Ferrer, A. Mixture Design: Overview and Comparison Between Different Approaches. *6th International Chemometrics Research Meeting - ICRM 2014*, 14-18/11/2014, Nijmegen, The Netherlands.
5. Palací-López, D. & Ferrer, A. Optimal Design of Experiments for Optimal Mixture Design. *14th Scandinavian Symposium on Chemometrics - SSC14*, 14-17/06/2015, Sardinia, Italy.
6. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Latent-variable model inversion and design of experiments in the latent space: a comparison oriented to process/product design. *6th International Conference Chemometrics in Analytical Chemistry - CAC 2016*, 6-10/06/2016, Barcelona, Spain.

7. Kerkenaar, H., Vitale, R., Palací-López, D., Postma, G., Buydens, L. & Ferrer, A. Kernel-Partial Least Squares regression coupled to pseudo-sample projection for the analysis of mixture designs of experiments. *XVI Chemometrics in Analytical Chemistry - CAC 2016*, 06-10/06/2016, Barcelona, Spain.

Oral communications in conferences

1. Palací-López, D., Ferrer, A. & Goos, P. Optimal Design of Experiments in the Original Space and in the Latent Space in Mixture Design. *VI Chemometrics Workshop for Young Researchers*, 02/10/2015, Valencia, Spain.
2. Palací-López, Villalba-Torán, P., Régil-Herrer, A., Núñez-Domingo, E., Facco, P., Barolo, M. & Ferrer, A. On-line Process Optimization Through PLS-model Inversion. *15th Scandinavian Symposium on Chemometrics (SSC15)*, 19-22/6/2017, Naantali, Finland.
3. Ferrer, A. & Palací-López, D. Process Optimization through PLS Model Inversion Using Historical Data (Not Necessarily from DOE). *18th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS 2018)*, 02-06/09/2018, Nancy, France.

Chapter 2
On optimal design of experiments

2.1. Introduction

Exploring, understanding, optimizing or making a production process more robust and flexible requires, in most cases if not all, building a causal model that explains how changes in input variables (e.g. materials and their properties, processing conditions…) relate to changes in the outputs (e.g. amount of product obtained, its quality, purity, value, generated pollutants…). To this purpose, deterministic (i.e. first principles) models are always desirable. However, the lack of sufficient knowledge and the generally ample need of resources required to properly construct such models makes their use unfeasible in a large number of cases. This is why data-driven models are often resorted to [1,2].

To guarantee causality when using data-driven approaches, however, independent variation in the input variables is required [3]. But, even if nowadays large amounts of data are available in most production processes, the variation in the inputs is commonly not independent. If classical polynomial model fitting by traditional methods (such as Ordinary Least Squares – OLS [4,5]) is to be resorted to, or if not enough data is yet available, this implies that controlled experimentation will have to be carried out, which usually requires extensive time and resources due to the large amount of inputs and outputs involved in most production processes at present. In such situations, defining the experiment to perform (i.e. the test or series of tests to carry out) in order to gather the necessary data [6] is commonly referred to as a design of experiment (DOE). An optimal DOE will be that which allows satisfying the purpose that motivated the experiment in the most efficient way, according to some optimality criterion.

When resorting to data-driven models, the aim of an experiment can be, as noted by Fedorov [7], characterized as one of the following in most cases:

1. Given a known (e.g. first principles) model that defines the analytical function relating the outputs of a process with its inputs, but whose parameters are unknown, a DOE may be carried out in order to obtain proper estimates of them.
2. If a process if governed by one among a series of possible functions, a DOE may be carried out in order to identify which of them is the correct one, as well as to estimate the corresponding parameters.
3. When no analytical function governing a process is known, a DOE is used to find the model (or models) that better approximates the unknown function in the region of interest for the experimenter (i.e. the experimental region).

Due to the increasingly large amounts of data and variables involved, the third of these scenarios will be the most frequent one in most real world applications in industrial processes, and therefore the identification of the experimental region becomes a crucial prior step to the DOE exercise. Any test to be performed must, then, be within this subspace, which is usually constrained by factors such as safety, equipment-operating range, and possibly additional restrictions imposed by the experimenter based on their prior knowledge of the process.

In this Ph.D. thesis, a discussion on the different optimality criteria commonly considered when resorting to DOE is presented, and algorithmic tools aimed at efficiently defining the experimental region and carrying out the DOE are provided, especially for mixture design problems and highly restricted experimental regions in general.

2.2. Optimality criteria for the design of experiments

When resorting to DOE, there are many criteria that may be taken into account in order to determine the efficiency or optimality of a given DOE, and to construct an optimal DOE [8]. Three of the most commonly considered criteria are:

a. D-optimality: this criterion is concerned with the minimization of the variance of parameter estimations. This is useful when the objective behind the DOE is factor screening, that is, determining which of the many potentially relevant manipulable inputs (commonly referred to as factors in DOE) actually have a significant effect on the output of interest. This criterion is useful in the second and third scenarios mentioned in Section 2.1.

b. I-optimality: this criterion is concerned with the response (i.e. output) estimation, and in particular with the average variance of prediction of the response over the experimental region, which is to be minimized. It is therefore a more useful criterion in scenarios such as the first one mentioned in Section 2.1.

c. G-optimality: similar to the I-optimality, this criterion refers to the maximum variance of prediction of the response over the experimental region, which again is to be minimized.

Although one of the most common criticisms of the optimal DOE approach is that different criteria will lead to different DOEs, the different criteria should be understood as seeking to provide a solution for the specific problem that motivated the experimentation. Constructing a D-optimal [7,9–11], I-optimal [12] or G-optimal [13] DOE will then depend on the problem being addressed. Furthermore, several criteria can be considered simultaneously, in order to obtain a DOE that performs well on multiple criteria at the same time [8].

2.2.1. D-optimal design of experiments

Consider a matrix of inputs $\mathbf{X}_\mathrm{o}$ [$N\times M$] and a vector of the output variable $\mathbf{y}_\mathrm{o}$ [$N\times 1$] where N is the number of samples or observations and M is the number of inputs. Consider also the inputs model matrix $\mathbf{X}$ [$N\times P$] and the outputs model vector $\mathbf{y}$ [$N\times 1$], obtained by appropriately pre-processing $\mathbf{X}_\mathrm{o}$ and $\mathbf{y}_\mathrm{o}$ for DOE or model-fitting purposes (by e.g. adding a column of ones and additional terms to $\mathbf{X}_\mathrm{o}$ corresponding to interaction terms, or second or higher order terms, and/or by centring or scaling each input/output variable), such that the matrix form of model that relates $\mathbf{X}$ and $\mathbf{y}$, referred to as a regression model, can be expressed as:

$$\mathbf{y} = \mathbf{X}\cdot\boldsymbol{\beta} + \mathbf{u} \tag{2.1}$$

Where $\boldsymbol{\beta}$ [$(P+1)\times 1$] is the vector that relates the P inputs to the output (plus an intercept or constant), and $\mathbf{u}$ [$N\times 1$] is the vector of disturbances or random errors.

The Ordinary Least Squares estimator of the vector of unknown model coefficients $\boldsymbol{\beta}$, $\mathbf{b}$, is

$$\mathbf{b} = (\mathbf{X}^\mathrm{T}\cdot\mathbf{X})^{-\mathbf{1}}\cdot\mathbf{X}^\mathrm{T}\cdot\mathbf{y} \tag{2.2}$$

and the variance-covariance matrix of this estimator, assuming the distribution of disturbances to be the same for all observations, results:

$$\mathrm{var}(\mathbf{b}) = \sigma_\mathrm{u}^2\cdot(\mathbf{X}^\mathrm{T}\cdot\mathbf{X})^{-1}$$

$$\mathrm{var}(\mathbf{b}) = \begin{pmatrix} \sigma_{b_0}^2 & cov(b_0,b_1) & \cdots & cov(b_0,b_P) \\ cov(b_0,b_1) & \sigma_{b_1}^2 & & cov(b_1,b_P) \\ \vdots & & \ddots & \vdots \\ cov(b_0,b_P) & cov(b_1,b_P) & \cdots & \sigma_{b_P}^2 \end{pmatrix} \tag{2.3}$$

Where σ_u^2 is the variance of the disturbances, which is, a priori, unknown, but can be estimated using the mean squared prediction error:

$$\hat{\sigma}_\mathrm{u}^2 = s_f^2 = \frac{1}{N-(P+1)}\cdot(\mathbf{y}-\mathbf{X}\cdot\mathbf{b})^\mathrm{T}\cdot(\mathbf{y}-\mathbf{X}\cdot\mathbf{b}) \tag{2.4}$$

When planning an experiment, however, σ_u^2 cannot yet be estimated. Therefore, only the elements of $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$ are considered. The elements of the diagonal of this matrix correspond to the relative variances of the estimation of the model parameters, which are desired to be as small as possible, and provide an estimate of how large the variances are compared to σ_u^2. If the factors are scaled so that their values in $\mathbf{X}$ range within the interval [-1, +1] (which is common practice when building a DOE), then the minimum achievable relative variance of an estimate is $1/N$, and is attained for all estimates when the DOE is orthogonal for its associated model. The relevance of a DOE being orthogonal resides in the fact that the parameters in a model can be estimated independently only when this is the case, but these estimates will be correlated to each other otherwise.

For a model matrix $\mathbf{X}$ to be orthogonal for its associated model, the sum of the element-wise products of all the pairs of columns in $\mathbf{X}$ must be zero. This means that, for a model that includes interactions of order as high as $K \leq P$, all of the $\sum_{k=1}^{K} \frac{P!}{k!\cdot(P-k)!}$ submatrices of $\mathbf{X}$ that can be defined by taking k (for $k = 1,2,..,K$) different columns of $\mathbf{X}$ (each of which will be referred to as $\mathbf{X}_{[k]}$ for simplicity) must meet that:

$$\sum_{n=1}^{N} \prod_{i=1}^{k} \mathbf{X}_{[k]}(n,i) = 0 \tag{2.5}$$

If this equality holds for every submatrix $\mathbf{X}_{[k]}$ as defined before, then $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$ is a diagonal matrix with diagonal elements $1/N$ and determinant $\left|(\mathbf{X}^T \cdot \mathbf{X})^{-1}\right| = (1/N)^P$, which is the smallest possible value for $\left|(\mathbf{X}^T \cdot \mathbf{X})^{-1}\right|$.

The inverse of the variance-covariance matrix in Equation 2.3 is the so-called information matrix, and serves as a summary of the available information on the model parameters. The larger the determinant of the information matrix is, the smaller $\left|(\mathbf{X}^T \cdot \mathbf{X})^{-1}\right|$ and the variance of the estimation of the parameters of the model. Small variances in these estimations are useful because the aim of screening experiments (where a D-optimal DOE is most useful) is to identify inputs (i.e. factors) or interactions of these have a statistically significant effect on the output/s, which is assessed via significance tests that are affected by, among other elements, the variance of the disturbances (which is not affected by the DOE) and the relative variance of the estimation of the parameters (which will depend on the DOE).

A D-optimal DOE will then be that which maximizes the determinant of the information matrix, $\left|\mathbf{X}^T \cdot \mathbf{X}\right|$ being the D-optimality criterion. This criterion, however, is only useful when comparing DOEs with the same number of tests or runs and parameters for the model to be estimated. To account for this, the D-efficiency may be considered instead:

$$\mathrm{D}-\text{efficiency}=\left(\frac{\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}}{N^{P}}\right)^{1/P}=\frac{(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{1/P}}{N} \tag{2.6}$$

However, one must be aware that the D-efficiency as defined by Equation 2.6 presents the issues of depending in the scale used for the experimental variables when constructing the DOE, and that it implicitly assumes that an orthogonal (and therefore D-optimal) experiment can be designed, which is not necessarily true in most real world applications, where the shape of the experimental region makes it impossible. The D-efficiency is useful nonetheless to compare different competing DOEs and select the one with highest D-efficiency to carry out the corresponding experimentation.

2.2.2. I-optimal and G-optimal design of experiments

Consider the model matrix as defined in Section 2.2.1, and its *n*-th row to be $\mathbf{x}_n^{\mathrm{T}}$. The variance of the prediction of the output corresponding to this set of inputs/factors, $\hat{y}_n$, can be computed as:

$$\mathrm{var}(\hat{y}|\mathbf{x}_n)=\mathrm{var}(\mathbf{b})=\sigma_{\mathrm{u}}^2\cdot\mathbf{x}_n^{\mathrm{T}}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}_n \tag{2.7}$$

Again, since σ_{u}^2 is unknown and cannot be estimated prior to the experimentation, the relative variance of prediction is computed instead. Two criteria can be evaluated then with regards to the variance of prediction:

1. the average relative value for this variance over the experimental space;
2. the maximum relative value for this variance in the experimental space.

The former corresponds to the I-optimality criterion, and can be computed as:

$$\mathrm{E}\big(\mathrm{var}(\hat{y}|\chi)\big)=\frac{\int_{\chi}\mathbf{x}^{\mathrm{T}}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}\,d\mathbf{x}}{\int_{\chi}d\mathbf{x}} \tag{2.8}$$

where $\mathrm{E}\big(\mathrm{var}(\hat{y}|\chi)\big)$ represents the expectation of the variance of the prediction of the output $\hat{y}$ over the experimental region, denoted by χ, and $\int_{\chi}d\mathbf{x}$ is the volume of the experimental region.

For the calculation of the numerator, given that $\mathbf{x}^{\mathrm{T}}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}$ is a scalar:

$$\mathbf{x}^{\mathrm{T}}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}=\mathrm{tr}[\mathbf{x}^{\mathrm{T}}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}] \tag{2.9}$$

Applying the fact that the trace of a product of matrices is the same even if they are cyclically permutated, then:

$$\mathrm{tr}[\mathbf{x}^{\mathrm{T}}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}]=\mathrm{tr}[(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{x}^{\mathrm{T}}\cdot\mathbf{x}] \tag{2.10}$$

Therefore:

$$\mathrm{E}\left(\mathrm{var}(\hat{y}|\chi)\right) = \frac{\int_{\chi} \mathrm{tr}[(\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X})^{-1} \cdot \mathbf{x}^{\mathrm{T}} \cdot \mathbf{x}]\, d\mathbf{x}}{\int_{\chi} d\mathbf{x}} = \frac{\mathrm{tr}\left[(\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X})^{-1} \cdot \int_{\chi} \mathbf{x}^{\mathrm{T}} \cdot \mathbf{x}\, d\mathbf{x}\right]}{\int_{\chi} d\mathbf{x}} \quad (2.11)$$

where $\int_{\chi} \mathbf{x}^{\mathrm{T}} \cdot \mathbf{x}\, d\mathbf{x}$ is called the moments matrix, which can be calculated exactly for cuboidal and spherical experimental regions [14].

It must be noted that $\mathrm{E}\left(\mathrm{var}(\hat{y}|\chi)\right)$ can be interpreted as the I-efficiency of a DOE given **X**, and its relative I-efficiency with respect to another DOE can be computed in the same way as the relative D-efficiency of one model with respect to another was computed.

When estimating the I-efficiency of a DOE for a model, however, being able to compute the volume of the experimental space becomes especially relevant. With respect to this, the volume can be easily calculated whenever the shape of the experimental region is that of a hyperrectangle, as is the case when all the factors are scaled to values ranging inside the interval [-1, +1] and all of them can be simultaneously set at their extreme values (in this particular case, the experimental region is said to be cuboidal, a particular case of the hyperrectangle). A spherical region is another common alternative to the cuboidal one. In most real world applications, however, the experimental region is neither one nor the other. In such cases, and specially for very irregular shapes, it may be necessary to resort to algorithms such as the Delaunay triangulation [15–17] to calculate this volume, although they may present some limitations as the dimensionality of the experimental space increases.

An alternative to the I-optimality is the G-optimality criterion, which seeks to minimize the maximum variance of prediction along the experimental region, $\max\left(\mathrm{var}(\hat{y}|\chi)\right)$, instead of $\mathrm{E}\left(\mathrm{var}(\hat{y}|\chi)\right)$. Building exact G-optimal designs, however, is not trivial, and comes at the cost of worse prediction variances over most of the region of interest [13].

2.3. Additional considerations for the design of experiments

When choosing a DOE, some aspects related to the individual parameters of the chosen model, and the chosen model itself, should be taken into account in addition to the criteria discussed in Section 2.2, namely the variance inflation, the aliasing and the principles of effect-sparsity, hierarchy and heredity.

2.3.1. Variance inflation

While it has already been stated that the relative variances of the estimation of the model parameters is minimum when an orthogonal design is resorted to, and equal to $1/N$ with [-1, +1] scaling of the factors, it is more often than not the case that the selected DOE will not be orthogonal, mainly because of budget or feasibility constraints, among possibly others. Because of this, it can be expected that the variance of the es-

timate for at least some of the parameters will be inflated. The degree to which this increase in variability takes place is called the Variance Inflation Factor (VIF). When factors are scaled to the range [-1, +1], the VIF for the *p*-th parameter of the model is equal to *N* times the relative variance of the estimation of that same parameter.

2.3.2. Aliasing

The selection of an a priori model for which a DOE is constructed implicitly assumes such model to be adequate for the problem under study. However, by doing so, the effects of some two-factor interactions or higher order terms are assumed to be negligible, while this not being necessarily the case. In such case, the estimates of the parameters of the estimated model will be biased by the parameters not included in the model.

Consider the true model to be expressed as

$$\mathbf{y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{X}_2 \cdot \boldsymbol{\beta}_2 + \mathbf{u} \tag{2.12}$$

and the selected model as

$$\mathbf{y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{u} \tag{2.13}$$

The expected value of the least squares estimation of $\boldsymbol{\beta}_1$ is:

$$\begin{aligned} \mathrm{E}(\widehat{\boldsymbol{\beta}}_1) &= \mathrm{E}\left[(\mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{X}_1)^{-1} \cdot \mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{y}\right] = (\mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{X}_1)^{-1} \cdot \mathbf{X}_1^{\mathrm{T}} \cdot \mathrm{E}[\mathbf{y}] \\ \mathrm{E}(\widehat{\boldsymbol{\beta}}_1) &= (\mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{X}_1)^{-1} \cdot \mathbf{X}_1^{\mathrm{T}} \cdot (\mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{X}_2 \cdot \boldsymbol{\beta}_2) \\ \mathrm{E}(\widehat{\boldsymbol{\beta}}_1) &= \boldsymbol{\beta}_1 + (\mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{X}_1)^{-1} \cdot \mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{X}_2 \cdot \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \mathbf{A} \cdot \boldsymbol{\beta}_2 \end{aligned} \tag{2.14}$$

$\mathbf{A}$ being the alias matrix. This means that, in general, the estimator of $\boldsymbol{\beta}_1$, $\widehat{\boldsymbol{\beta}}_1$, is biased unless $\boldsymbol{\beta}_2 = \mathbf{0}$ (i.e. the effects assumed to be negligible actually are) or $\mathbf{X}_1^{\mathrm{T}} \cdot \mathbf{X}_2$ is a matrix of zeros, i.e. none of the effects corresponding to the factors in $\mathbf{X}_2$ biases the estimate of any of the effects corresponding to the factors in $\mathbf{X}_1$.

2.3.3. Principles of effect-sparsity, hierarchy and heredity

The principle of effect sparsity (or factor sparsity) is a restatement of the Pareto principle that concerns mainly screening experimentation. According to this principle, most of the variability in the output or response variable is expected to be captured by a relatively small number of factors (although the probability of correctly identifying these "large effect" will, logically, be affected by the DOE).

The principle of hierarchy concerns model selection, in that it states that main effects (as opposed to e.g. two-factor effect or higher order interaction effects) comprise the largest source of variability in most processes, followed by two-factor interaction effects, quadratic effects, and later on by higher order interaction effects, cubic effects, etc. This means that considering main and lower-order effects into the regression mod-

el, for which the DOE is to be constructed, should precede the addition of higher-order effects.

Finally, the principle of heredity states that models with strong heredity that include e.g. two-factor interaction effects will also include the main effects of the factors involved (and the same could be said regarding higher-order-factor interactions effects with respect to lower-order-factor interaction effects). Models with weak heredity, however, may e.g. only include the main effect of one of the factors involved in a two-order interaction. Models with strong heredity present the advantage that predictions made with them are not affected by a different scaling of the factors used to fit it.

Although these principles have been validated to an extent, it must also be noted that violations of them are more common than suggested by the literature on screening experimentation [18].

Chapter 3
On latent variable and kernel-based multivariate data analysis

3.1. Introduction

Nowadays large amounts of data are being continuously generated at a high frequency in many fields thanks to the increasing variety of measurement tools and technologies rapidly being integrated in most scientific activities and production processes. This so-called data tsunami [19] could be rendered useless, however, if such a huge volume of data were not converted into valuable information through its appropriate analysis and interpretation. Consider, e.g., a present-day industrial environment, where most statistical process control schemes, used for fault detection and to monitor the evolution of the process, are based on the univariate control of one or small numbers of measured variables (temperatures, pressure, pH, etc.). While useful in the past, this strategy is completely inadequate given the huge quantities of instrumental responses being continuously and automatically registered in modern plants. This is specially true when a process is being monitored by many automated sensors, a common occurrence at present, since any undesired special events that would bring the process out of its so-called Normal Operating Conditions (NOC) and impact the quality of the final products will be reflected not just in a change in the magnitude of some of the measured process variables, but potentially also in the relationship among them. This change in their relationship (or correlation structure) is, however, difficult to detect by classical statistical approaches that can only handle independent or, at most, slightly correlated variables.

Alternative methods, based on the so-called latent variables, may be used instead to take advantage of such correlation in order to model the structure of the process space. These techniques can deal with large volumes of data that contain little useful information (i.e. low signal-to-noise ratio), and also allow the use of happenstance (i.e. 'routine' data, not from a DOE, and therefore not causal in nature [20]) data to build mod-

els for optimization purposes, by allowing causal relationships to be inferred in the space of the latent variables.

In this Ph.D. thesis, use will be made of some of these methodologies for a variety of purposes presented in the distinct chapters of the manuscript, from mixture design data analysis and model-fitting, to model inversion and optimization.

3.2. Latent variable-based multivariate data analysis techniques

By projecting the high-dimensional original variables space onto a low-dimensional, latent orthogonal (i.e. uncorrelated) variables space, multivariate latent variable-based methods (widely resorted to in the field of chemometrics) reduce the dimensionality of the data under study and permit describing their underlying sources of variation. Principal Component Analysis (PCA) and Partial Least Squares (PLS) regression are some of the most well-known and extended techniques of this kind.

3.2.1. Principal Component Analysis (PCA)

PCA [21,22] is one of the most widely used multivariate statistical tools to compress, describe and interpret large sets of data. Consider a matrix $\mathbf{X}$ [$N \times M$], defined as in Chapter 2, so that M denotes the number of measured variables[i] and N is the number of observations registered (e.g. at N time instants or for N individuals). When M is very large (maybe even $M > N$) the relevant information contained in $\mathbf{X}$ will usually be intercorrelated among a number of the variables in the dataset. This makes it possible to reduce, with a certain degree of acceptable accuracy, the original M-dimensional space spanned by these variables values for the registered observations into an A-dimensional subspace ($A \leq M$) associated to the directions of maximum variability for the available data, and onto which the N individuals can be projected as new points. Mathematically, PCA is based on the following bilinear structure model:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}} + \mathbf{1} \cdot \mathbf{m}_{\mathbf{X}}^{\mathrm{T}} + \mathbf{E} \qquad (3.1)$$

Where $\mathbf{T}$ [$N \times A$] contains in its n-th row the projection coordinates or scores values of the latent variables of the n-th row of $\mathbf{X}$ onto the A-dimensional subspace, $\mathbf{P}$ [$M \times A$] is the array of the so-called loadings which determine the A basis vectors (which are the vectors signalling the A directions of maximum variability of the data in $\mathbf{X}$, also known as principal components or factors[ii]) of the PCA subspace, and $\mathbf{E}$ [$N \times M$] stands for the matrix of unmodelled residuals, i.e. the portion of $\mathbf{X}$ outside of the PCA subspace not

[i] M is used here for simplicity although, as in Chapter 2, this matrix may be extended to have P columns including interaction and second- or higher order-effects. In that case, when using PCA in this manuscript, the matrix containing the values for the main

[ii] To avoid confounding the meaning of 'factor' here with that in Chapter 2, only the term 'principal component' will be used from this point onwards when referring to the basis vectors in PCA

explained for the chosen rank, A; $\mathbf{m_X}$ [M×1] and $\mathbf{D_{s_X}}$ [M×M] are, respectively, the column vector of centring factors (usually the mean values of the variables in $\mathbf{X}$) and the diagonal matrix of scaling factors (usually the standard deviation of the variables in $\mathbf{X}$) applied to the M input variables before performing the PCA; and $\mathbf{1}$ [N×1] is a vector of ones.

The PCA solution may also be formulated in many equivalent ways and attained by Singular Value Decomposition (SVD) [23], among other algorithms, such that:

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^{\mathrm{T}} + \mathbf{E}_{\mathrm{o}} \tag{3.2}$$

with the columns of $\mathbf{U}$ [N×A] and $\mathbf{V}$ [M×A] being the left and right singular vectors of $\mathbf{X}$, respectively, and $\mathbf{S}$ [A×A] a diagonal matrix with the squares of the first A non-zero singular values of $\mathbf{X}$, $\sqrt{\boldsymbol{\lambda}_A}$, as its diagonal elements. By comparing Equations 3.1 and 3.2 It can be easily seen that $\mathbf{T} = \mathbf{U} \cdot \mathbf{S}$, $\mathbf{P} = \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{V}$ and $\mathbf{E}_{\mathrm{o}} = \mathbf{1} \cdot \mathbf{m}_{\mathbf{X}}^{\mathrm{T}} + \mathbf{E}$. The following properties are assumed:

$$\begin{aligned} \mathbf{P}^{\mathrm{T}} \cdot \mathbf{P} &= \mathbf{I}_A \\ \mathbf{T}^{\mathrm{T}} \cdot \mathbf{T} &= \mathrm{diag}(\boldsymbol{\lambda}_A) \end{aligned} \tag{3.3}$$

where $\mathbf{I}_A$ is the [A×A] identity matrix, and the a-th element of $\boldsymbol{\lambda}_A$ is the eigenvalue of the a-th PCA component.

3.2.2. Partial Least Squares regression (PLS)

PLS [24–26] is a latent variable-based approach used to model the inner relationships between a matrix of inputs or predictors, $\mathbf{X}$ [N×M], and a matrix of outputs or response variables, $\mathbf{Y}$ [N×L], through few, uncorrelated Latent Variables (LV) that identify the underlying causal relationship between $\mathbf{X}$ and $\mathbf{Y}$. PLS is usually resorted to in order to predict $\mathbf{Y}$ from the A-dimensional subspace associated to $\mathbf{X}$ such that its covariance with $\mathbf{Y}$ is maximised. The PLS regression model structure can be expressed as follows:

$$\begin{aligned} \mathbf{X} &= \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}} + \mathbf{1} \cdot \mathbf{m}_{\mathbf{X}}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} + \mathbf{1} \cdot \mathbf{m}_{\mathbf{Y}}^{\mathrm{T}} + \mathbf{F} \\ \mathbf{T} &= \left(\mathbf{X} - \mathbf{1} \cdot \mathbf{m}_{\mathbf{X}}^{\mathrm{T}}\right) \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{W} \cdot (\mathbf{P}^{\mathrm{T}} \cdot \mathbf{W})^{-1} = \left(\mathbf{X} - \mathbf{1} \cdot \mathbf{m}_{\mathbf{X}}^{\mathrm{T}}\right) \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{W}^{*} \end{aligned} \tag{3.4}$$

being $\mathbf{T}$ [N×A], $\mathbf{P}$ [M×A] and $\mathbf{E}$ [N×M] the $\mathbf{X}$ scores, loadings and residuals matrices, respectively; $\mathbf{Q}$ [L×A] and $\mathbf{F}$ [N×L] the respective $\mathbf{Y}$ loadings and residuals matrices, and $\mathbf{W}$ [M×A] the weighting matrix, such that $A < r_{\mathbf{X}}$, $r_{\mathbf{X}}$ being the rank of $\mathbf{X}$, i.e. the maximum number of latent variables that can be considered when fitting the PLS regression model; $\mathbf{m_X}$ and $\mathbf{D_{s_X}}$ being the [M×1] column vector of centring factors and the [M×M] diagonal matrix with the scaling factors applied to the M input variables before fitting the PLS-regression model, respectively, and $\mathbf{m_Y}$ and $\mathbf{D_{s_Y}}$ the corresponding [L×1] column vector and [L×L] diagonal matrix associated to the output variables, and

$\mathbf{1}$ a vector of ones as defined in Section 3.2.1.

Alternatively, $\mathbf{Y}$ can be expressed as a function of $\mathbf{X}$:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} \cdot \mathbf{B} + \mathbf{1} \cdot \mathbf{b}_0^{\mathrm{T}} + \mathbf{1} \cdot \mathbf{m}_{\mathbf{Y}}^{\mathrm{T}} + \mathbf{F} \\ \mathbf{B} &= \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{W}^{*} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \\ \mathbf{b}_0 &= \mathbf{m}_{\mathbf{Y}} - \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \mathbf{W}^{*\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{m}_{\mathbf{X}} \end{aligned} \tag{3.5}$$

where $\mathbf{B}$ [$M \times L$] is an array of regression coefficients, and $\mathbf{b}_0$ [$L \times 1$] a vector of intercepts of the regression model.

PLS presents several advantages over most classical statistical predictive methods such as Ordinary Least Squares regression (OLS) [4,5], such as not needing to assume linearly independent regressors and being able to simultaneously model several outputs while simultaneously taking into account the correlation structure not only among the inputs but also the outputs. Furthermore, since causal relationships can be inferred in the latent space [27], this permits the use of historical datasets for optimization purposes, which reduces the amount of experimentation required, or even prevents it altogether [28]. Lastly, since the initial number of variables involved (M) has been reduced to a smaller number of uncorrelated LV (A), the computational cost of any optimization problem in the latent space will be reduced when compared to the equivalent problem in the original space, even more so as the number of variables in the original space increases with the complexity of the problem addressed while the number of LV remains relatively low.

3.3. Kernel-based techniques

Although latent variable-based multivariate data analysis techniques such as PCA and PLS have proven to be very powerful tools for the analysis and interpretation of multivariate data, they a priori assume the underlying structure of the datasets to which they are applied to be linear. While several approaches have been proposed to address situations where this assumption is far from appropriate, such as non-linear PLS [29–36] or artificial neural networks [37], they often require optimising many adjustable parameters, with the additional risk of being affected by local minima and show overfitting. Kernel-based techniques [38] are then a good alternative [39], as can be seen by their broad application in fields such as chemistry [40,41], biology [42], informatics [43,44] and continuous process monitoring [45,46]. Kernel-based approaches usually suffer from an important disadvantage when it comes to the interpretation of the influence of the original variables or its importance on the model, however, since this information is lost after the kernel transformation. Although possibilities to recover this information exist their implementation is not straightforward, most of them do not permit graphical nor intuitive interpretation, and their use for continuous process monitoring requires an appropriate database containing past identified failures for comparison, which is usually not available in real case studies [41,47–50]. An extension of the principles of non-

linear bi-plots and pseudo-sample projection [51], however, have been proposed in recent years to overcome these limitations [52–55].

An adaptation of these kernel techniques will be presented here for mixture data analysis, and therefore the basic principles behind them as well as a brief explanation regarding the pseudo-samples and pseudo-sample projection are succinctly explained in the following sections.

3.3.1. Basic principles of kernel-based techniques

All of the kernel-based data analysis techniques are based on the so-called kernel transformation, given by:

$$K(\mathbf{x}_n, \mathbf{x}_{n'}) = < \phi(\mathbf{x}_n), \phi(\mathbf{x}_{n'}) > \tag{3.6}$$

where $\mathbf{x}_n^{\mathrm{T}}$ [1×M] and $\mathbf{x}_{n'}^{\mathrm{T}}$ [1×M] represent the n-th and n'-th rows of the original data matrix $\mathbf{X}$ [N×M], ϕ the mapping function applied to them (which is not needed to be known a priori with kernel-based approaches), and $<$ and $>$ denote the inner product.

Applying this transformation to every possible couple of row vectors that constitute $\mathbf{X}$, a squared symmetric matrix $\mathbf{K}$ [N×N] is obtained whose elements represent the dissimilarity (or distance measurements) between every two observations. All the many generic kernel functions that can be applied to obtain $\mathbf{K}$ present the two fundamental properties:

i. They allow the original data to be projected onto the so-called feature space, a space of higher dimension than the original one, such that non-linear relationships in $\mathbf{X}$, if they exist, may be described in a linear way.

ii. They permit the inner product between observations to be performed in the feature space, which makes it possible to apply in this space algorithms of classical multivariate linear methodologies based on the calculation of the inner product matrix of $\mathbf{X}$ (e.g. PCA and PLS) [39].

Only three types of kernel functions will be applied in the second part of this thesis: the linear (first-order), the q^{th}-order polynomial, and the Gaussian or Radial Basic Function (RBF) kernel. Table 3.1 gathers their mathematical formulations and, for the last one, its possible adjustable parameter. It must be noted that, although only the PLS technique will be used in this manuscript together with the kernel transformation, once the kernel matrix $\mathbf{K}$ has been computed, any classical bilinear technique can be applied to it generating, e.g., a Kernel-PCA (K-PCA) or Kernel-PLS (K-PLS) model when a PCA or PLS, respectively, is applied to $\mathbf{K}$.

Table 3.1. Kernel functions used in the second part of this thesis and their adjustable parameters

Kernel type	Kernel function	Adjustable parameter
Linear (first-order)	$\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_{n'}$	-
q^{th}-order polynomial	$(\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_{n'})^q$	-
Gaussian/RBF	$exp\left(-\frac{\lVert\mathbf{x}_n-\mathbf{x}_{n'}\rVert^2}{2\cdot\sigma}\right)$	σ

3.3.2. Pseudo-samples and pseudo-sample projection

As stated in [56][iii], the term pseudo-sample is used to refer to an observation whose weight is put in one single variable, such that, for example, a vector of zeros of all of its elements except for a one in its m-th position, $\mathbf{g}$ [M×1], would represent one of the many possible pseudo-samples associated the m-th variable of a dataset $\mathbf{X}$. The score of such a sample, when projected onto the latent space of a PLS model fitted with a single latent variable[iv], can be calculated as in Equation 3.7:

$$t_{\mathbf{g}^{\mathrm{T}}} = \mathbf{g}^{\mathrm{T}} \cdot \mathbf{w}^* = w_m^* \tag{3.7}$$

Which is the m-th value of the weighting vector $\mathbf{w}^*$ and provides information regarding the contribution of the m-th variable, x_m, to the model. Instead of a vector like $\mathbf{g}$, a matrix $\mathbf{V}_m$ [V×M] can be created that contains V samples instead of a single one, such that its entries are as follows:

$$\mathbf{V}_m = \begin{bmatrix} 0 & 0 & \cdots & \min(x_m) & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \vdots & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \max(x_m) & \cdots & 0 & 0 \end{bmatrix} \tag{3.8}$$

Its projection onto the latent space will, therefore, define a trajectory of the form:

$$\mathbf{V}_m \cdot \mathbf{w}^* = \begin{bmatrix} \min(x_m) \cdot w_m^* \\ \vdots \\ \max(x_m) \cdot w_m^* \end{bmatrix} \tag{3.9}$$

In a more general case, where A latent variables ($A \geq 1$) are considered when fitting the PLS regression model (or more than one component chosen for PCA), the matrix resulting from the previous operation provides the geometrical locus of all the points

[iii] Note that most of the contents of Section 3.3.2, as well as the appendices in Sections 14.1.1 and 14.1.2, can be found in [56], and have been put together here to avoid making the reader look for the original source.

[iv] The presented mathematical derivation can be easily extended to PCA

along the direction determined by the origin of the latent space and each point, whose coordinates represent the weights of x_m on the A latent variables. While the representation of these trajectories does not provide, per se, additional information to that given by the PLS (or PCA) model, some insight can still be obtained regarding the evolution of the original variables in the latent space when kernel-based methods are resorted to. As demonstrated by Postma et al. [53], pseudo-sample projection does in fact permit recovering information related to the contribution of the original variables when a Euclidean distance matrix, **D**, is dealt with. This strategy can be resorted to when using K-PCA/K-PLS, since **D**, if double-centred, is directly generated through a linear kernel transformation of a mean-centred dataset (see Section 14.1.1). By transforming each pseudo-sample array into a pseudo-sample kernel one, as done with **X**, a new $V \times N$ array is then obtained that contains information about the dissimilarity between the V pseudo-samples and the N original observations (see Section 14.1.2). Furthermore, the pseudo-sample projection can be used with any kernel transformation as long as they generate sets of distances which may be embedded in a Euclidean space [51].

3.4. Important additional notions: cross-validation and jackknifing

In some parts of the manuscript concerning model-fitting statistical techniques, the concepts of cross-validation and jackknifing may appear, thus shall be defined.

Cross-validation refers to a model validation technique commonly resorted to in order to select the most appropriate number latent variables to extract when a latent variable-based technique (e.g. PLS) is used to fit a model for, e.g., predictive purposes. Cross-validation is performed as follows:

1. The data under study are split into several complementary subsets (i.e. no two subsets contain the same observation)
2. The analysis (e.g. via PLS) is carried out using of all but one of these subsets
3. The remaining subset is exploited for testing (e.g. to evaluate the predictive capability of the regression model fitted with all other subsets)

Multiple rounds of cross-validation are usually performed under different partitions and the final results are averaged to reduce variability. When used to assess the performance of the model for prediction, this is generally done by evaluating the evolution of the prediction error as the complexity of the model (e.g. number of latent variables extracted in PLS) increases.

Jackknifing, on the other hand, is a resampling approach used to quantify a particular statistic (e.g. variance of a prediction, bias…) by iteratively removing an individual from a data matrix and compute the estimate of interest, to then calculate the mean or a percentile of the resulting distribution of values for the corresponding estimate.

Chapter 4
Materials and methods

4.1. Hardware

Most of the computations executed for the elaboration of this Ph.D. thesis were run on a MacBook Pro equipped with a 2.9 GHz Intel Core i5 and 8 GB 1867 MHz DDR3 RAM.

The testing of the software MiDAs, presented in Section 7, and the simulations performed with the software PRO/II, necessary to illustrate some of the examples in Sections 9 and 12, were run on an HP-Pavilion with a 2.33 GHz Core2 Quad Q8200 and 5 GB 800 MHz DDR2 RAM.

4.2. Software

The software packages exploited here are:

- macOS X Yosemite, Version 10.10.5;
- MATLAB R2014B, Version 8.4.0.150421
- Microsoft Windows 10 Pro
- PRO/II, Version 9.3

4.3. Datasets and methods

To facilitate a more friendly reading of this manuscript, the information regarding datasets, materials and methods presented and used along this manuscript can be found in the corresponding chapter where they are resorted to or applied.

A short summary can be seen below nonetheless.

- In Chapter 6, Section 6.1, datasets from the following sources were used:
  - R.D. Snee, Developing Blending Models for Gasoline and Other Mixtures, Technometrics. 23 (1981) 119–130.
  - J.A. Cornell, Experiments with Mixtures, Second Edi, Wiley, New York, USA, 1990.
- In Chapter 6, Section 6.2, datasets from the following sources were used:
  - L. Eriksson, E. Johansson, C. Wikstrom, Mixture design — design generation , PLS analysis , and model usage, (1998).
  - D. Alman, C. Pfeifer, Empirical colorant mixture models, Color Res. Appl. 12 (1987) 210–222.
  - J.A. Cornell, How to run mixture experiments for product quality, Asq Press, 1990.
- In Chapter 9, datasets where generated by means of:
  - The mathematical model in

    P. Facco, F. Dal Pastro, N. Meneghetti, F. Bezzo, M. Barolo, Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development, Ind. Eng. Chem. Res. 54 (2015) 5128–5138.
  - The procedure shown in

    F. Arteaga, A. Ferrer, Building covariance matrices with the desired structure, Chemom. Intell. Lab. Syst. 127 (2013) 80–88.

    F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemom. Intell. Lab. Syst. 101 (2010) 38–42.
  - The simulated production process in PRO/II according to

    PRO/II Casebook #1 Vinyl Chloride Monomer Plant, in: PRO/II Caseb. #1, SIMULATION SCIENCES INC, 1992.
- The code required for the Quadratic Programming (QP) and Linear Programming (LP) optimization algorithms described in Chapter 10, and resorted to in Chapters 11 and 12, has been implemented in Matlab (MATLAB R2014B). As a part of this code, the functions *fmincon*, *fminunc* and *linprog* are used for the constrained QP, unconstrained QP, and LP, respectively.
- In Chapter 12, datasets where generated following the same procedure as with the first an third cases in Chapter 9.

PART II

Mixture design optimization

Chapter 5
Traditional approaches to mixture design

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

1. Vitale, R.[v], Palací-López, D.[v], Kerkenaar, H., Postma, G., Buydens, L. & Ferrer, A. Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments. *Chemometr. Intell. Lab.* 175, 37-46 (2018).

5.1. Introduction

A wide range of products currently used in daily life results from processing blends of two or more ingredients. Hence, the physicochemical properties of these products mainly depend on the raw materials being mixed and on the proportions in which they are added. Alloys, as well as drugs and foodstuffs, are just some of the numerous examples where this applies, and their manufacturing can be considered a so-called mixture problem [57]. Traditionally, mixture problems are defined as those in which i) the proportions x_q of the Q different constituents are related to the aforementioned properties, ii) these proportions are of at least as much relevance as their absolute quantities, and iii) their sum must be a fixed value (usually 1 or 100%):

$$\sum_{q=1}^{Q} x_q = 1 \qquad 0 \leq x_q \leq 1 \tag{5.1}$$

[v] These authors had equal contributions

This perfect collinearity restriction makes it impossible to modify the composition of any one of the ingredients independently from the rest, which significantly affects i) the shape of the experimental region within which a DOE, whenever necessary, is to be built and in which experimentation should be performed, and ii) the way the data is to be analysed, and models built and interpreted if classical polynomial fitting by traditional methods, like Ordinary Least Squares (OLS) or Generalised Least Squares (GLS) [58–60], are to be resorted to.

Furthermore, it must be noted that not every process involving a blend should necessarily be treated as a mixture design problem, nor does a problem not involving a mixture prevent it from having to be addressed as such. Consider the following two cases:

a) When one of the ingredients of the blend, whose effect on the quality attribute of interest is null, represents most of the bulk of the mixture itself, as in e.g. a children's cough medicine [61], any changes in the proportions of the other constituents will remain almost insignificant, and therefore the *absolute* quantities of those components will be more relevant than their *relative* amounts in the blend. Such case should not be approached as a mixture design problem.

b) Goos & Jones [62] present an interesting case of a supplier of a milling operation aimed at reducing the thickness of sheet aluminium down to a specific value in three steps (i.e. using three mills). Although each mill may contribute to the reduction of thickness to different extents, the total amount of reduction in thickness is fixed. This is equivalent to the so-called 'mixture-amount problem', where the 'amount' (i.e. absolute reduction in thickness of the sheet aluminium) may vary depending on the required girth, but the sum of the *percentage of required reduction performed by each mill* is constant (100%). Therefore, although technically no blend is involved, this case should be approached as a mixture problem (more specifically a mixture-amount problem where the quality of the product depends on both the relative and absolute quantities of the 'ingredients').

Finally, dealing with 'mixture variables' adds additional complexity to most real world production processes where 'process variables' (e.g. temperatures, pressures…) must also be accounted for. These problems usually require fitting so-called mixture-process variable models, which significantly increases the amount of data required to do so and complicates its interpretation. Additionally, although not addressed in the present PhD. Thesis, taking into consideration also the information regarding the ingredients/raw materials properties in L or T-shaped data structures require the use of more complex, latent variable-based approaches [63–65].

In this section of the manuscript, insight is provided regarding how the nature of the mixture design problems affects the shape of the experimental region (also referred to as *Mixture Space*) and the form and interpretation of mixture models, which ultimately impact the way a mixture design of experiments has to be constructed and carried out.

For the sake of simplicity terms such as ‘mixtures’, ‘blends’ and ‘ingredients’ will be used even though it has already been stated that some ‘mixture design problems’ do not necessarily concern an actual mixture, blend, or any ‘ingredient’.

5.2. The mixture space

Given a series of ingredients of a blend, and the process conditions under which the blend is to be produced/processed, the mixture space is constituted by all feasible combinations of ratios of its constituents. By ‘feasible’ it is understood that restrictions may be imposed on the ratios of some or all of the ingredients to e.g. guarantee the physicochemical stability of the blend or assure that it can be safely or easily manipulated under those processing conditions. Lower bounds on the proportions of the ingredients are usually set above zero, and upper bounds below one, in order to guarantee a small amount of all/most of the constituents to be present in the blend, and to avoid ‘pure’ mixtures constituted by a single ingredient, respectively. Two different kinds of mixture spaces must then be differentiated, depending on their shape: the so-called simplex mixture spaces, which retain the original shape (but not necessarily size) it had before any restrictions were imposed (other than the ones expressed in Equation 5.1), and irregular mixture spaces for which this is no longer true.

5.2.1. Assessing the shape of the mixture space

A mixture design problem is said to be unconstrained if the only restrictions imposed on the blend constituents proportions are i) the perfect collinearity restriction defined in Equation 5.1, and ii) them being able to vary from 0 to 1. In such case, the mixture space is said to be that of a simplex, and can be visualized as the convex hull enveloped by the $(Q-1)$-dimensional generalization of a tetrahedron, with Q being the number of mixture ingredients. In this case, and for $Q=2$, the mixture space is the line segment connecting the ‘pure mixtures’ with only the first and the second constituent in it, respectively; for $Q=3$, the mixture space is the convex hull of the equilateral triangle; for $Q=4$, it is the convex hull of a tetrahedron, and so on.

A mixture design problem is said to be constrained if at least one active/non-redundant restriction is imposed on at least one of the blend constituents, in addition to those for ‘unconstrained’ mixture problems. A restriction is said to be redundant if it can be discarded without modifying the shape/size of the convex hull. Otherwise, such restriction is active. By imposing additional active univariate inequality constraints on the proportions of the blend ingredients, the resulting mixture space may remain a (differently sized) simplex or not but, as a rule of thumb, imposing active multivariate constraints on them (i.e. restrictions on linear combinations of the ratios/proportions) will make the mixture space no longer a simplex.

The importance of determining if the mixture space is a simplex or not lies on the fact that i) there exist pre-defined optimal DOEs that relay on the mixture design being a simplex and ii) the interpretation of the mixture model's parameters (discussed in Section 5.3) may only have a practical sense as long as the mixture space is a simplex.

When only univariate restrictions are imposed on the ingredients' proportions in the blend, $\mathbf{lb_x} \leq \mathbf{x} \leq \mathbf{ub_x}$, assessing if the mixture space is a simplex is done as follows:

a. If the vector of lower bounds for the proportions, $\mathbf{lb_x}$ [Q×1], is the [Q×1] vector of zeros [Q×1], $\mathbf{0}_Q$, and the vector of upper bounds, $\mathbf{ub_x}$ [Q×1], is the [Q×1] vector of ones [Q×1], $\mathbf{1}_Q$, the mixture space is 'unconstrained' and therefore a simplex.

b. If $\mathbf{ub_x} = \mathbf{1}_Q$, three different scenarios must be considered:

   i. If $\sum_{q=1}^{Q} lb_q < 1$, the mixture space is a simplex, but some of the upper bounds may be inconsistent (i.e. unachievable) given $\mathbf{lb_x}$. This occurs whenever $r_q = ub_q - lb_q > R_\mathrm{L}$, where $R_\mathrm{L} = 1 - \sum_{q=1}^{Q} lb_q$. In that case, the corresponding consistent upper bound, ub_q^*, is $ub_q^* = lb_q + R_\mathrm{L}$

   ii. If $\sum_{q=1}^{Q} lb_q = 1$, the mixture space is a single point, and the consistent upper bounds $\mathbf{ub_x^*} = \mathbf{lb_x}$.

   iii. If $\sum_{q=1}^{Q} lb_q > 1$, the mixture space is the null set, since no blend can simultaneously satisfy all lower bounds on its ingredients' proportions.

c. If $\mathbf{lb_x} = \mathbf{0}_Q$, four scenarios must be considered:

   i. If $\sum_{q=1}^{Q} ub_q > 1$ and $\sum_{q=1}^{Q} ub_q - \min(\mathbf{ub_x}) \leq 1$, the mixture space is a simplex, but some of the lower bounds may be inconsistent given $\mathbf{ub_x}$. This occurs if $r_q > R_\mathrm{U}$, where $R_\mathrm{U} = \sum_{q=1}^{Q} ub_q - 1$. In that case, the corresponding consistent lower bound, lb_q^*, is $lb_q^* = ub_q - R_\mathrm{U}$.

   ii. If $\sum_{q=1}^{Q} ub_q = 1$ and $\sum_{q=1}^{Q} ub_q - \min(\mathbf{ub_x}) \leq 1$, the mixture space is a single point, and the consistent lower bounds $\mathbf{lb_x^*} = \mathbf{ub_x}$.

   iii. If $\sum_{q=1}^{Q} ub_q > 1$, but $\sum_{q=1}^{Q} ub_q - \min(\mathbf{ub_x}) > 1$, the mixture space is not a simplex.

   iv. If $\sum_{q=1}^{Q} ub_q < 1$, the mixture space is the null set, since no blend can simultaneously satisfy all upper bounds on its ingredients' proportions

d. If $\mathbf{lb_x} \neq \mathbf{0}_Q$ and $\mathbf{ub_x} \neq \mathbf{1}_Q$ (the most general and realistic scenario), the consistent bounds ought to be calculated as in scenarios 'b' and 'c', and then R_L

and R_{U} computed with the updated bounds. For the mixture space to not be void it must be met that $\sum_{q=1}^{Q} lb_q \leq 1$ and $\sum_{q=1}^{Q} ub_q \geq 1$. This being the case:

i. If $R_{\mathrm{L}} \neq R_{\mathrm{U}}$ and $r_q = \min(R_{\mathrm{L}}, R_{\mathrm{U}})\ \forall q \in \{1, 2, \ldots, Q\}$, the mixture space is a simplex. If $R_{\mathrm{L}} < R_{\mathrm{U}}$, the simplex is referred to as an L-simplex, and its facets are defined by the lower bounds imposed on the ingredients, such that the simplex retains its original orientation. On the other hand, If $R_{\mathrm{L}} > R_{\mathrm{U}}$, the simplex is referred to as an U-simplex, and its facets are defined by the upper bounds imposed on the ingredients, such that the orientation of the simplex is inverted with respect to the 'unconstrained' simplex.

ii. If $R_{\mathrm{L}} = R_{\mathrm{U}}$, the mixture space is not a simplex.

Whenever the mixture space is a simplex in scenarios 'b', 'c' and 'd', it is not uncommon to substitute the so-called *components*, x_q, whose values will typically not range from 0 to 1, by the corresponding so-called *pseudocomponentes*, x_q^*, whose values will:

$$\begin{aligned} x_q^* &= \frac{x_q - lb_q^*}{R_{\mathrm{L}}}\ \forall q\ \in \{1, 2, \ldots, Q\} \text{ if } R_{\mathrm{L}} < R_{\mathrm{U}} \\ x_q^* &= \frac{ub_q^* - x_q}{R_{\mathrm{U}}}\ \forall q\ \in \{1, 2, \ldots, Q\} \text{ if } R_{\mathrm{L}} > R_{\mathrm{U}} \end{aligned} \tag{5.2}$$

Figure 5.1 illustrates three examples of an unconstrained (L-simplex) mixture space, a constrained U-simplex mixture space, and a constrained irregular mixture space for a blend with 3 ingredients.

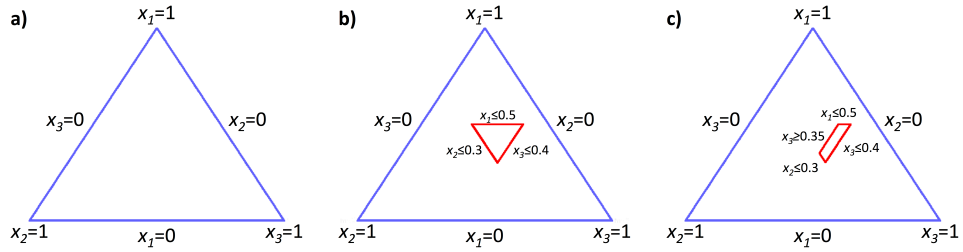


Figure 5.1 Visualization of a) a L-simplex mixture space without restrictions imposed on the ingredients' proportions ($\mathbf{lb}_{\mathbf{x}}^* = \mathbf{0}_3$; $\mathbf{ub}_{\mathbf{x}}^* = \mathbf{1}_3$), b) a U-simplex mixture space where the restrictions $\mathbf{lb}_{\mathbf{x}} = [0.1; 0.1; 0.2]$; $\mathbf{ub}_{\mathbf{x}} = [0.5; 0.3; 0.4]$ are imposed on the ingredients' proportions ($\mathbf{lb}_{\mathbf{x}}^* = [0.3; 0.1; 0.2]$; $\mathbf{ub}_{\mathbf{x}}^* = \mathbf{ub}_{\mathbf{x}}$), and c) an irregular mixture space where the restrictions $\mathbf{lb}_{\mathbf{x}} = [0.1; 0.1; 0.35]$; $\mathbf{ub}_{\mathbf{x}} = [0.5; 0.3; 0.4]$ are imposed on the ingredients' proportions ($\mathbf{lb}_{\mathbf{x}}^* = [0.3; 0.1; 0.35]$; $\mathbf{ub}_{\mathbf{x}}^* = \mathbf{ub}_{\mathbf{x}}$)

When multivariate restrictions are also imposed on the ingredients' proportions in the blend, and the previous assessment indicates that the mixture space is still a simplex when only the univariate restrictions are accounted for, it must be evaluated if any of the multivariate restrictions is redundant or not. To do so, the Q vertices of the simplex

can be easily obtained from the consistent bounds. Unless *all* vertices of the simplex meet *all* of the multivariate restrictions, at least one of such constraints is active (i.e. not redundant) and the mixture space is no longer a simplex.

5.2.2. Identifying the envelope of the mixture space

Section 5.2.1 focused on assessing the shape of the mixture space in order to recognize if it is a simplex or not. However, this may not suffice to properly build a DOE, specially when the mixture space is not a simplex and/or process variables are also involved. 'Identifying the envelope of the mixture space' here means being able to define which constraints of those initially imposed actually delimit the convex hull that is the mixture space (i.e. are active/non-redundant constraints), and how these constraints relate to every *r*-dimensional ($r < Q - 1$) element that is part of that envelope (vertices, edges, planes and *r*-dimensional hyperplanes). This is required to increase the computational efficiency of DOE construction algorithms, may they be point-exchange [10] or coordinate-exchange-based ones [66]. The identification of this envelope is easily done when the mixture space is a simplex and the consistent bounds are known, since all of its facets, which are the $(Q - 2)$-dimensional elements of the envelope, are defined directly by the lower bound (for the L-simplex) or upper bounds (for the U-simplex). Every vertex is then a 0-dimensional element of the envelope that can be obtained by defining the different feasible combinations of lower and upper bounds. Two distinct algorithms for the identification of the envelope are proposed, given that only linear restrictions (either univariate or multivariate) are imposed, based on the following principles:

i. Each linear restriction (univariate or multivariate) imposed on the variables of a *M*-dimensional space is associated to a $(M - 1)$-dimensional hyperplane.
ii. Each inequality restriction will divide the original *M*-dimensional space into two half-M-dimensional spaces, but not reduce its dimensionality.
iii. If it exists, the intersection of *R* half-M-dimensional spaces defines an *M*-dimensional convex space whose *R* facets are associated to the *R* restrictions that defined those half-*M*-dimensional spaces.
iv. A minimum of *M*+1 compatible (i.e. they can be met simultaneously) inequality restrictions is required to define the envelope of a convex hull in a *M*-dimensional space (e.g. the triangle and the tetrahedron are the geometric shapes with less sides/faces that can be described as linear restrictions in a 2 and 3 dimensional space, respectively).
v. Each equality restriction will reduce the dimensionality of the originally M-dimensional space by one. Therefore, at most *M* equality restrictions can be imposed simultaneously, as long as they are compatible with each other.
vi. Any *r*-dimensional element that is part of the envelope of a convex hull can be defined as the intersection of $M - r$ of the linear restrictions associated to the facets of such envelope.

For the application of the proposed algorithms, all restrictions will be expressed in matrix form as follows:

$$\begin{matrix} \mathbf{A_x} \cdot \mathbf{x} \leq \mathbf{d_x} \\ \mathbf{F_x} \cdot \mathbf{x} = \mathbf{f_x} \end{matrix} \tag{5.3}$$

$\mathbf{A_x}$ [$I_1 \times Q$] being a matrix whose *i*-th row contains the coefficients of the *i*-th linear combination of inputs (in this case, mixture ingredients' proportions in the blend) to which an inequality constraint is applied, and $\mathbf{d_x}$ [$I_1 \times 1$] is the column vector whose *i*-th element indicates the maximum permitted value for that linear combination. Similarly, $\mathbf{F_x}$ [$I_2 \times Q$] and $\mathbf{f_x}$ [$I_2 \times 1$] are associated to the equality restrictions imposed on different linear combinations of inputs. In accordance with the nomenclature used here, I_1 inequality restrictions and I_2 equality restrictions ($I_2 \leq Q$) are assumed to be imposed. Note that, for the particular case of an 'unconstrained' mixture space:

$$\begin{matrix} \mathbf{A_x} = \begin{bmatrix} -\mathbf{I}_Q \\ \mathbf{I}_Q \end{bmatrix} \; ; \; \mathbf{d_x} = \begin{bmatrix} \mathbf{0}_Q \\ \mathbf{1}_Q \end{bmatrix} \\ \mathbf{F_x} = \mathbf{1}_Q^{\mathrm{T}} \; ; \; \mathbf{f_x} = 1 \end{matrix} \tag{5.4}$$

The first of the proposed algorithms is a slight variant of the so-called CONSIM algorithm [67], and the steps one must follow for its application are the following:

1. Evaluate if all of the equality restrictions are compatible with each other by solving the corresponding system of linear equations.
    a. If there is no solution, the mixture space is void.
    b. If a single solution exists, check if it meets all inequality constraints. If not, the mixture space is void. Otherwise it is just one mixture.
    c. If more than one solution exist, continue to the next step.
2. Generate an arbitrarily large hyperrectangle and define a [$2^Q \times Q$] matrix of initial vertices so that none of the vertices meet any restriction. Define also a symmetric [$2^Q \times 2^Q$] 'matrix of edges' that serves to indicate which of these initial vertices are connected to each other, which has a one in its *i*-th row, *j*-th column if the *i*-th and *j*-th vertices are connected, and a zero otherwise.
3. Starting with the equality restrictions, evaluate which vertices are at one side of the hyperplane defined by each restriction, which at the other side, and which, if any, meet the restriction at its limit (i.e. as if it was an equality constraint). The initial matrices of vertices and edges are then updated according to the following rules:
    a. A new vertex results from the intersection of the hyperplane corresponding to the equality/inequality restriction and the edge connection two vertices located each at one side of such hyperplane.

   b. Vertices that do not meet the restriction being assessed at the moment are discarded

   c. The new vertices are connected to the remaining of the two vertices checked in the current iteration, as well as to any new vertices generated during the same step that meet the same Q-1 current active restrictions at their limit/as an equality.

4. Check active and redundant restrictions among the ones applied so far. A restriction is active if at least M vertices meet it as if it was an equality constraint (even if it is an inequality restriction). It is redundant otherwise.

5. Repeat steps 3 and 4 until all the restrictions have been used to 'cut' the cuboid. The final matrix of restrictions will contain only those that are not redundant

6. If, during any step, all vertices are discarded, then no subspace exists that meets all imposed restrictions

The second of the proposed algorithms will be referred to as 'Segmentation by defining and discarding vertices', and the steps for its application are the following:

1. Evaluate if all of the equality restrictions are compatible with each other by solving the corresponding system of linear equations, and proceed identically as in the first step in the previous algorithm. Proceed to the next step only if appropriate.

2. Define all possible combinations of non-parallel inequality restrictions (usually $I_1/2$) taken $(Q - I_2)$ at a time. This requires generating an auxiliary matrix of at most dimension $\left[\left(\frac{I_1!}{(Q-I_2)!\cdot(I_1+I_2-Q)!}\right)\times(Q - I_2)\right]$.

3. For every combination defined in step 2, solve the system of linear equations that results from concatenating the equality constraints with the corresponding combination of inequality constraints. Check if this solution (if it is a single point) meets all other inequality restrictions not considered to obtain it. Unless it does, discard it. Otherwise, add it to a matrix of 'final vertices'.

4. Once all 'final vertices' have been obtained, check active and redundant restrictions among the initial ones. A restriction is active if at least M vertices meet it as if it was an equality constraint. It is redundant otherwise.

5. If no vertices are kept at any step, then no subspace exists that meets all imposed restrictions, and the mixture space is therefore void.

It is important to note that the computational efficiency of each algorithm will depend on the number of constraints imposed and the dimensionality of the mixture space. However, the 'Segmentation by defining and discarding vertices' is expected to be more efficient in most scenarios, since the redundancy of the restrictions is only as-

sessed at the very last step, and no 'edges matrix' or a initial matrix of vertices need to be defined, both of which dramatically increase in size with Q. Algorithms for the definition of extreme vertices DOEs such as the ones proposed by Snee & Marquadt [68] or McLean & Anderson [69] may prove to be a sensible alternative to obtain the vertices of an irregular mixture space when no multivariate constraints are imposed, although they may present issues when applied to problems with high dimensionality.

5.3. Regression model structures in mixture design

Let y denote the output of interest being studied in a mixture problem, and let x_q $(i \in \{1,2,\dots,Q\})$ be the proportion of the i-th constituent in a blend with at most Q ingredients. The variable y is considered as a random variable, whose distribution depends on the values of the x_i, such that:

$$E(y) = f(x_1, x_2, \dots, x_Q) \tag{5.5}$$

Because of the perfect collinearity restriction in Equation 5.1, caution is advised when trying to model the response function by means of classical polynomials. Consider, as an example, the first-order polynomial to be used as an approximation of the response function for a mixture problem with Q constituents involved:

$$E(y) = \alpha_0 + \sum_{i=1}^{Q} \alpha_i \cdot x_i \tag{5.6}$$

The interpretation of the parameters of the model in Equation 5.6 is:

α_0: expected average value of y when all the constituents of the mixture are absent $(x_i = 0 \; \forall i \in \{1,2,\dots,Q\})$

α_i: expected increase of the average value of y when the proportion of the i-th blend ingredient is increased by 1, while keeping the proportions of all other components unchanged

As can be observed, the interpretation of the coefficients makes no practical sense, since i) no blend can exist if all of its constituents are absent, and ii) it is impossible to modify one of the ingredient's proportion without altering the proportion of at least one other ingredient. Alternative model structures are therefore required that account for the particular nature of mixture variables. This still holds true when process variables are involved.

5.3.1. The Scheffé models

One of the most common reparametrization of the classical polynomials to model mixtures is that of the Scheffé canonical polynomials [70], which result from applying the

restriction in Equation 5.1 to the classical polynomials. Here the first-, second- and third-order Scheffé canonical are presented, as they are the most commonly resorted to.

- Canonical form of the linear model/first-order polynomial:

Given a blend with Q constituents, the first-order Scheffé canonical polynomial can be obtained by introducing the restriction in Equation 5.1 into the first-order polynomial in Equation 5.6. By doing so:

$$E(y) = \alpha_0 + \sum_{i=1}^{Q} \alpha_i \cdot x_i = \alpha_0 \cdot \sum_{i=1}^{Q} x_i + \sum_{i=1}^{Q} \alpha_i \cdot x_i \tag{5.7}$$

Reorganizing the different terms:

$$E(y) = \sum_{i=1}^{Q} \beta_i \cdot x_i \qquad \beta_i = \alpha_0 + \alpha_i \tag{5.8}$$

Which is the first-order Scheffé canonical polynomial. What makes it different from its classical polynomial counterpart, in terms of structure, is only the absence of an intercept. The greatest difference, though, lies in its interpretation. Consider a 'pure mixture' with only the i-th ingredient present in it. In that case, $x_i = 1$ and $x_j = 0\ \forall j \neq i$, and $E(y|x_i = 1) = \beta_i$. Therefore, β_i represents the expected average value for the response variable y for the 'pure mixture' constituted solely by its i-th ingredient. A more visual interpretation of the parameters of the first-order Scheffé canonical polynomial is illustrated in Figure 5.2.

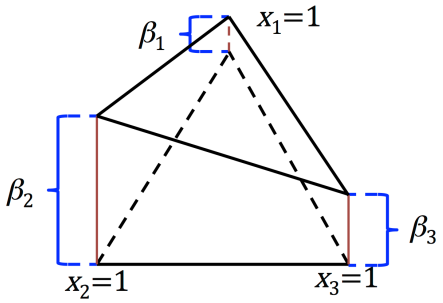


Figure 5.2 Interpretation of the β_i parameters of the Scheffé canonical polynomial for a mixture with Q=3

This model is quite simple and implies that for any mixture $E(y)$ is a weighted average of the expected responses for the different 'pure mixtures', the weighting coefficients

being the proportions x_i of the respective components in the mixture. Although this may seem an inconvenience, it is extremely useful in screening studies, when Q is very high, and in cases where the constraints on the constituents' proportions delimit the mixture space to a very small region of the complete simplex. However, as already mentioned and observed in Figure 5.2, β_i corresponds to the expected average value for the response variable y for the 'pure mixture' with only the i-th ingredient. This may make sense whenever the proportion of the i-th constituent can vary from 0 to 1, but this is rarely the case in practice since the allowed proportion of each component in the mixture is usually restricted. Therefore, in most real case studies the estimation of β_i, b_i, will hold more theoretical than practical sense, since it would be an estimator of the expected average response for a mixture outside of the mixture space and/or without practical interest.

Once the first-order Scheffé canonical polynomial model is fitted, it may be of interest to study the effect that each component of the mixture has 'by itself' on the response, that is, if in general it can be said that a given component actually improves or worsens the value of the response when added to a greater or lesser extent, or if it actually does not really have a significant effect on it. This assessment can be done by computing the so-called total or orthogonal effect of the q-th component, which is defined as the difference in the expected response between the q-th 'pure mixture' and the mixture at the centroid of the facet of the simplex opposite to the q-th vertex:

$$q - \text{th total effect} = \beta_q - \sum_{\substack{j=1 \\ j \neq q}}^{Q} \beta_j \Big/ (Q-1) \tag{5.9}$$

Since usually restrictions are imposed on the proportions of the constituents in the mixture, it is very likely that the proportion of each component will not vary between 0 and 1, but along a narrower interval. Then computing the adjusted orthogonal effect of each component is advised instead of their total effects. The adjusted orthogonal effect is obtained as

$$q - \text{th adj. orth. effect} = r_q \cdot \left(\beta_q - \sum_{\substack{j=1 \\ j \neq q}}^{Q} \beta_j \Big/ (Q-1) \right) \tag{5.10}$$

r_q being defined in Section 5.2.1, where it was obtained to assess if a mixture space was a simplex or not.

It should be noted that the calculation of both the total and adjusted effects of the blend constituents implicitly takes the centroid of a simplex as the reference mixture. Therefore their interpretation will be useful (even more so for the adjusted effects) whenever the mixture space is effectively a simplex. Otherwise, which actually concerns most

real problems, the orthogonal effects of the different components will generally have little meaning, if any.

- Canonical form of the quadratic/second-order polynomial:

The classical form of the second-order polynomial is:

$$E(y) = \alpha_0 + \sum_{i=1}^{Q} \alpha_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \alpha_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \alpha_{ii} \cdot x_i^2 \tag{5.11}$$

The quadratic terms can be reformulated as:

$$x_i^2 = x_i \cdot \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^{Q} x_j \right) \tag{5.12}$$

Substituting Equations 5.1 and 5.12 in Equation 5.11, and reorganizing terms:

$$\begin{aligned} E(y) \quad &= \sum_{i=1}^{Q} \beta_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \beta_{ij} \cdot x_i \cdot x_j \\ \beta_i &= \alpha_0 + \alpha_i + \alpha_{ii} \;\; ; \;\; \beta_{ij} = \alpha_{ij} - \alpha_{ii} - \alpha_{jj} \end{aligned} \tag{5.13}$$

The interpretation of the parameters β_i is the same as for the linear model, and the parameter β_{ij} indicates a synergistic interaction between the *i*-th and *j*-th constituents its value is positive, an antagonistic interaction if it is negative. A synergistic interaction between the *i*-th and *j*-th ingredients implies that, given a mixture that contains both, the expected value of the response y will be greater than the weighted averages of the individual responses corresponding to the respective 'pure mixtures', whereas if there is a antagonistic interaction the opposite would be expected. Therefore, the expected response for a sample of a given composition will be the weighted average of the simple effects, plus the effects of any significant synergistic/antagonistic pair-wise interactions.

It must be noted, however, that this model assumes these interactions to remain synergistic/antagonistic regardless of the proportions in which the corresponding two constituents are found in the blend. To better understand the interpretation of the β_{ij} parameters, consider the simplified scheme shown in Figure 5.3, where all but the *i*-th and *j*-th components of a mixture are absent, and the proportions of these two components vary from 0 to 1.

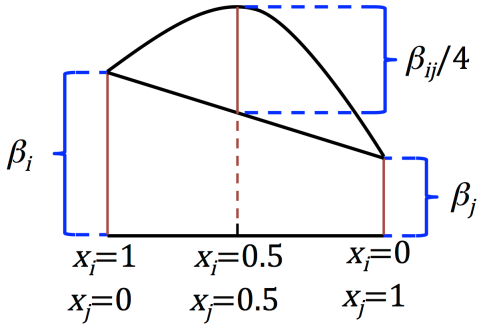


Figure 5.3 Interpretation of the β_{ij} parameters of the Scheffé canonical polynomials

In Figure 5.3 a case is shown for which the parameter β_{ij} is positive, so that the curve represented for the expected response is always above the straight line connecting the expected values of *y* for the *i*-th and *j*-th 'pure mixtures' (which the curve would coincide with if $\beta_{ij} = 0$). In this model it is assumed that the effect of the interaction between each pair of components is constant, that is, it is always of the same magnitude and in the same direction (synergistic and antagonistic). Then, the maximum (if the interaction is synergistic) or minimum (if it is antagonistic) of the curve is achieved when $x_i = x_j = 0.5$, and $E(y|x_i = x_j = 0.5) = \frac{\beta_i+\beta_j}{2} + \frac{\beta_{ij}}{4}$.

As with the interpretation of β_i for the first-order canonical polynomial, this interpretation of β_{ij} is valid in theory, but generally not in practice, since the presence of restrictions on the proportion of the different components in the mixture implies that, in most cases, binary mixtures will not belong to the mixture space, or will not be of practical interest. Thus, although the interpretation of the sign of the parameters β_{ij} may still make theoretical sense, the strict interpretation of its value, without the aforementioned considerations, will often lack practicality.

- Canonical form of the cubic/third-order polynomial:

As with the previous cases, consider the classical form of the third-order polynomial:

$$E(y) = \alpha_0 + \sum_{i=1}^{Q} \alpha_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \alpha_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \alpha_{ii} \cdot x_i^2 + \sum_{i=1}^{Q-2} \sum_{j=i+1}^{Q-1} \sum_{k=j+1}^{Q} \alpha_{ijk} \cdot x_i \cdot x_j \cdot x_k + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \alpha_{iij} \cdot x_i^2 \cdot x_j + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \alpha_{ijj} \cdot x_i \cdot x_j^2 + \sum_{i=1}^{Q} \alpha_{iii} \cdot x_i^3 \quad (5.14)$$

The third-order Scheffé canonical polynomial is:

$$E(y) = \sum_{i=1}^{Q} \beta_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \left[\beta_{ij} \cdot x_i \cdot x_j + \gamma_{ij} \cdot x_i \cdot x_j \cdot (x_i - x_j)\right] + \sum_{i=1}^{Q-2} \sum_{j=i+1}^{Q-1} \sum_{k=j+1}^{Q} \beta_{ijk} \cdot x_i \cdot x_j \cdot x_k \tag{5.15}$$

$$\beta_i = \alpha_0 + \alpha_i + \alpha_{ii} + \alpha_{iii}$$
$$\beta_{ij} = \alpha_{ij} - \alpha_{ii} - \alpha_{jj} - \alpha_{iii} - \alpha_{jjj}$$
$$\gamma_{ij} = \alpha_{iij} + \alpha_{ijj} + \alpha_{iii} + \alpha_{jjj}$$
$$\beta_{ijk} = \alpha_{ijk} - \alpha_{iij} - \alpha_{ijj} - \alpha_{iik} - \alpha_{ikk} - \alpha_{jjk} - \alpha_{jkk} - \alpha_{iii} - \alpha_{jjj} - \alpha_{kkk}$$

In this model, the interpretation of the parameters of the form β_i and β_{ij} remains, while the β_{ijk} parameters allow the quantification of the effect of possible ternary interactions, and the γ_{ij} ones reflect, if significant, more complex types of binary interactions between pairs of components with respect to the quadratic model, such that their intensity or even the sign may vary (from synergistic to antagonistic) depending on the proportion in which both constituents are found in the blend. While the interpretation of the parameters β_{ijk} is analogous to that of β_{ij}, although hard to illustrate in a single image, the interpretation that should be given of the parameters γ_{ij} may be clarified by looking at Figure 5.4.

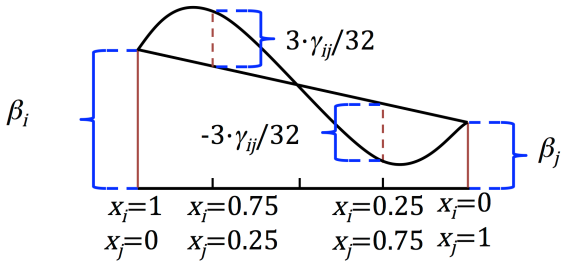


Figure 5.4 Interpretation of the $\boldsymbol{\gamma_{ij}}$ parameters of the Scheffé canonical polynomials

In Figure 5.4 the expected value for the response are shown for different compositions of a binary mixture in which only the *i*-th and *j*-th constituents in the blend are present. As can be seen, a positive value of $\boldsymbol{\gamma}_{ij}$ indicates that the interaction between the *i*-th and *j*-th ingredients is synergistic when $x_i > x_j$ (higher proportion of the *i*-th constituent), and antagonistic otherwise. As with the parameters β_i, β_{ij} and β_{ijk}, caution is advised when interpreting the parameters $\boldsymbol{\gamma}_{ij}$ must be carried out carefully, since in

many situations the binary mixtures may not be a part of the mixture space, or will not be of practical interest. Therefore, strictly adhering to the interpretation here presented could be meaningless in practice.

Since the number of parameters of this model increases rapidly with Q, even with a relatively small number of blend constituents, an alternative one, the special cubic model, is often formulated instead. Consider the following variant of the classical third-order polynomial in Equation 5.14, where terms of the form $\alpha_{iij} \cdot x_i^2 \cdot x_j$ and $\alpha_{ijj} \cdot x_i \cdot x_j^2$ are no longer taken into account:

$$E(y) = \alpha_0 + \sum_{i=1}^{Q} \alpha_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \alpha_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \alpha_{ii} \cdot x_i^2 + \sum_{i=1}^{Q-2} \sum_{j=i+1}^{Q-1} \sum_{k=j+1}^{Q} \alpha_{ijk} \cdot x_i \cdot x_j \cdot x_k + \sum_{i=1}^{Q} \alpha_{iii} \cdot x_i^3 \tag{5.16}$$

The special cubic model results:

$$E(y) = \sum_{i=1}^{Q} \beta_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \beta_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q-2} \sum_{j=i+1}^{Q-1} \sum_{k=j+1}^{Q} \beta_{ijk} \cdot x_i \cdot x_j \cdot x_k \tag{5.17}$$

$$\beta_i = \alpha_0 + \alpha_i + \alpha_{ii} + \alpha_{iii}$$

$$\beta_{ij} = \alpha_{ij} - \alpha_{ii} - \alpha_{jj} - \alpha_{iii} - \alpha_{jjj}$$

$$\beta_{ijk} = \alpha_{ijk} - \alpha_{iii} - \alpha_{jjj} - \alpha_{kkk}$$

This model disregards the terms associated with the more complex effects of binary interactions, and therefore does not allow quantifying these effects. However, the reduced number of parameters to be estimated also permits a smaller number of tests to be performed, whenever a DOE is necessary, and provides better estimations of the coefficients with the same number of samples required to fit the complete third-degree Scheffé canonical polynomial.

With regards to these so-called Scheffé models, it must be highlighted that:

i. The reparametrization performed on the classical polynomials allows eliminating the perfect collinearity created by the mixture constraint (Equation 5.1). However, in most real problems there are restrictions that generate imperfect collinearity, which these models are unable to deal with.

ii. These models present complicated forms even for uncomplicated problems, and the interpretation of their parameters is neither intuitive nor practical in most mixture problems, where the mixture space is not a simplex.

iii. The absence of an intercept makes centring the data around an average impossible, which in practice may lead to unreliable estimates of the parameters of the model (i.e. highly dependant on the dataset used to fit the model).

5.3.2. The Cox models

The Cox models [71] are a reparametrization of the Scheffé models, one of the objectives of which is to allow a more intuitive interpretation of the effects of each blend constituent in the response variable. For this purpose, a formulation that is visually very similar to the standard one is used, since they are polynomials with a constant. However, a restriction relative to a reference mixture is imposed on these coefficients. The first-order Cox model can be formulated as:

$$\begin{gathered} E(y) = \beta'_0 + \sum_{i=1}^{Q} \beta'_i \cdot x_i \\ s.t. \quad \sum_{i=1}^{Q} \beta'_i \cdot s_i = 0 \end{gathered} \tag{5.18}$$

where s_i is the proportion of the *i*-th blend constituent in a pre-specified reference mixture, such that β'_0 is the expected value of the response variable for this same reference mixture. Thus the Cox linear model is the projection of a model in a Q-dimensional subspace onto a $(Q-1)$-dimensional subspace whose centre is the reference mixture.

For the interpretation of the parameters β'_i, consider that the proportion of the *i*-th component in the mixture is increased by Δ_i from x_i, while decreasing the proportions of all other components proportionally to their ratios in the reference mixture (this is the direction of the so-called Cox axis for the *i*-th component). This implies that:

$$\begin{gathered} x_i \to x_i + \Delta_i \\ x_j \to x_j - \Delta_i \cdot \frac{s_j}{(1-s_i)} \quad \forall j \neq i \end{gathered} \tag{5.19}$$

And the variation in the expected response will be:

$$\begin{gathered} \Delta E(y) = \beta'_0 + \beta'_i \cdot (x_i + \Delta_i) + \sum_{\substack{j=1 \\ j \neq i}}^{Q} \beta'_j \cdot \left(x_j - \frac{\Delta_i \cdot s_j}{1-s_i} \right) - \left[\beta'_0 + \sum_{i=1}^{Q} \beta'_i \cdot x_i \right] \\ \Delta E(y) = \beta'_i \cdot \Delta_i - \frac{\Delta_i}{1-s_i} \cdot \sum_{\substack{j=1 \\ j \neq i}}^{Q} \beta'_j \cdot s_j \end{gathered} \tag{5.20}$$

Because of the restriction in Equation 5.21:

$$\Delta E(y) = \beta'_i \cdot \Delta_i - \frac{\Delta_i}{1 - s_i} \cdot \left[-\beta'_i \cdot s_i + \sum_{i=1}^{Q} \beta'_i \cdot s_i \right]$$
$$\Delta E(y) = \beta'_i \cdot \Delta_i + \beta'_i \cdot s_i \cdot \frac{\Delta_i}{1 - s_i} = \frac{\beta'_i \cdot \Delta_i}{1 - s_i} \tag{5.21}$$

Therefore:

$$\beta'_i = \frac{\Delta E(y)_{\left[\begin{smallmatrix} x_i \to x_i + \Delta_i \\ x_j \to x_j - \Delta_i \cdot s_j/(1-s_i) \end{smallmatrix}\right]}}{\Delta_i/(1 - s_i)} \tag{5.22}$$

Consequently, if $\Delta_i = (1 - s_i)$, it can be seen that β'_i is the variation on the expected response when moving from the reference mixture to the vertex corresponding to the 'pure mixture' constituted solely by the *i*-th ingredient.

The second-order Cox model can be formulated as:

$$E(y) = \beta'_0 + \sum_{i=1}^{Q} \beta'_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \beta'_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \beta'_{ii} \cdot x_i^2$$
$$s.t. \begin{cases} \sum_{i=1}^{Q} \beta'_i \cdot s_i = 0 \\ \sum_{j=1}^{Q} c_{ij} \cdot \beta'_{ij} \cdot s_j = 0 \;\; \forall\, i \in \{1,2,\dots,Q\} \\ c_{ij} = \begin{cases} 1/2 & if\ i \neq j \\ 1 & if\ i = j \end{cases} \end{cases} \tag{5.23}$$

The interpretation of the parameters β'_{ij} is similar to the corresponding one for classical polynomials, while the interpretation of the parameters β'_{ii} is associated to the curvature of the response surface along the *i*-th Cox axis, i.e. the direction of the line that connects the reference mixture with the vertex corresponding to the *i*-th 'pure mixture'.

It must be noted that the Scheffé model coefficients can be computed from the Cox ones in the same way as done when relating the model coefficients for the classical polynomials to the coefficients of the Scheffé canonical polynomials (see Section 14.2.1). However, the Cox coefficients cannot be directly obtained from the Scheffé ones, and instead a reference mixture must be defined and a system of linear equations solved to do so. Alternatively, the Cox model may be obtained as the solution of an OLS problem with restrictions imposed on the model parameters (more specifically, those described in e.g. Equation 5.23 for the second-order polynomial). Given the complexity of estimating the Cox coefficients, and both the Scheffé and Cox models

(of the same degree) providing the same predictions of the response, resorting to one formulation or the other will depend mainly on how relevant the interpretability of the model's parameters is for a given problem.

5.3.3. Mixture-process variable models

Both the Scheffé and Cox models presented in Sections 5.3.1 and 5.3.2 are useful when the mixture problem being addressed involve exclusively mixture variables (i.e. blend constituents proportions). This is not a realistic expectation, however, in many situations, and therefore different models are needed when the effect of mixture factors and process variables on the response variable is to be quantified simultaneously. Still, classical polynomial fitting by traditional methods, like Ordinary or Generalised Least Squares (OLS/GLS), require treating both kinds of variables differently: while mixture data analysis requires the formulation of special models, such as the Scheffé or Cox ones, classical polynomials can still be resorted to when fitting models associated to process variables. Historically, the models used to analyse data that include both mixture factors and process variables has been carried out by formulating two models separately, each one concerning mixture factors (a Scheffé model) and process variables (a classical polynomial), respectively, and then combining them. To exemplify this procedure, a common mixture-process variable model can be obtained by combining the second-order Scheffé model for proportions of the Q mixture constituents, x_q (Equation 5.13), with the polynomial model that includes main effects and two-factor-interaction terms for the M process variables, z_m:

$$E(y) = \alpha_0 + \sum_{k=1}^{M} \alpha_k \cdot z_k + \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \alpha_{kl} \cdot z_k \cdot z_l \tag{5.24}$$

The combined model results:

$$\begin{aligned} E(y) = & \sum_{i=1}^{Q} \gamma_{0,i} \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \gamma_{0,ij} \cdot x_i \cdot x_j \\ & + \sum_{k=1}^{M} \left[\sum_{i=1}^{Q} \gamma_{k,i} \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \gamma_{k,ij} \cdot x_i \cdot x_j \right] \cdot z_k \\ & + \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \left[\sum_{i=1}^{Q} \gamma_{kl,i} \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \gamma_{kl,ij} \cdot x_i \cdot x_j \right] \cdot z_k \cdot z_l \end{aligned} \tag{5.25}$$

Terms in the first row in Equation 5.25 correspond to the linear and non-linear properties of the mixture constituents and binary blends. Terms in the second row contain the linear effect of the *k*-th process variable, z_k, in such properties, while terms in the third row quantify the effects of the interactions among the *k*-th and *l*-th process variables on

the mixture ingredients' blending properties. The issue of its interpretability aside, one of the most severe drawbacks of this model is the large number of parameters whose estimation is required, which requires at least $[Q + Q \cdot (Q-1)/2] \cdot [1 + M + M \cdot (M-1)/2]$ runs if a DOE is resorted to. An alternative model is proposed by Kowalski et al. [72]:

$$E(y) = \sum_{i=1}^{Q} \gamma_{0,i} \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \gamma_{0,ij} \cdot x_i \cdot x_j + \sum_{k=1}^{M} \sum_{i=1}^{Q} \gamma_{k,i} \cdot x_i \cdot z_k + \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \alpha_{kl} \cdot z_k \cdot z_l + \sum_{k=1}^{M} \alpha_k \cdot z_k^2 \tag{5.26}$$

This model assumes that significant effects exist neither of the process variables nor the two-factor interactions between the *k*-th and *l*-th process variables on the mixture ingredients' blending properties. It does, however, consider the additive effect of two-factor-interactions and quadratic effects of the process variables on $E(y)$ (as opposed to the multiplicative effect in Equation 5.25). This results in more parsimonious model for which $Q + Q \cdot (Q-1)/2 + Q \cdot M + M \cdot (M-1)/2 + M$ have to be estimated. The advantages and disadvantages that come from defining a more or less parsimonious model are discussed in the literature, and different proposals are provided by e.g. Prescott [73], who also presents a slightly less parsimonious model than that in Equation 5.26, with $Q \cdot (Q+1) \cdot (Q+2)/6 + M \cdot Q \cdot (M+Q+2)/2$ parameters to be estimated. To better visualize the complexity of these models, the number of parameters whose estimation is required depending on *Q* and *M*, for the three models, is shown in Table 5.1. Shading has been added to the cells corresponding to the combinations of *Q* and *M* for which the model proposed by Prescott is more parsimonious than the one in Equation 5.25.

Table 5.1. Number of parameters of the model in Equation 5.25, and the ones proposed by Kowalski et al. [72] and Prescott [73] (separated by a slash from each other), as a function of *Q* and *M*

		Q					
		2	3	4	5	6	7
M	1	6/6/9	12/10/19	20/15/34	30/21/55	42/28/83	56/36/119
	2	12/10/16	24/15/31	40/21/52	60/28/80	84/36/116	112/45/161
	3	21/15/25	42/21/46	70/28/74	105/36/110	147/45/155	196/55/210
	4	33/21/36	66/28/64	110/36/100	165/45/145	231/55/200	308/66/266
	5	48/28/49	96/36/85	160/45/130	240/55/185	336/66/251	448/78/329
	6	66/36/64	132/45/109	220/55/164	330/66/230	462/78/308	616/91/399

In short, the study of even the simplest, yet realistic, mixture-process variables problems by traditional methods requires fitting excessively complex and hard to interpret models, which in turn demands large amounts of experimentation.

The definition of combined mixture-process variables models, such as the one in Equation 5.25, adds further complexity to both the constructions of a DOE to gather the required data to fit them and to the process of estimating the model parameters themselves. This is because many of the columns in the matrix **X** corresponding to interactions among mixture and process variables will be a linear combination of columns corresponding to lower-order interactions and simple effects. The procedure to estimate such parameters is illustrated with some examples in [57], and several proposals for the construction of appropriate DOEs for this sort of problems have also been provided in the literature [74–76].

5.4. Mixture design of experiments

In this section the problematic associated to the construction of mixture designs of experiments will be addressed, first concerning mixture problems where the mixture space is a simplex, and then the case where the mixture space is irregular. The approach for the construction of DOE in irregular mixtures spaces that will be proposed in Section 5.4.3 is formulated in such a way that it can also be resorted to for addressing mixture-amount, mixture-process variables, and mixture-amount-process variables problems.

5.4.1. Simplex-based DOE

As long as the mixture space is a simplex, constructing a design of experiments is relatively simple. The following are some basic designs of experiments commonly used in mixing problems in this situation:

- Simplex lattice experimental design $\{Q, m\}$:

These designs allow the estimation of m-order mixture model. In this design, m +1 equidistant proportions varying from 0 to 1 $\left(0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}, 1\right)$ are defined for each of the Q blend constituents (the so-called pseudo-components can be used for restricted simplex spaces, and the 'real' proportions computed once the DOE has been constructed) and then all feasible mixture resulting from the combinations of these proportions for the different blend constituents are defined. For a Q-dimensional mixture space, the simplex lattice DOE generates $m - r$ equidistant points on each r-dimensional element ($r \leq m - 1$) that is part of the envelope of the mixture space (not accounting for the extreme vertices, in the case of the edges). This type of design is especially suitable to estimate the parameter of e.g. a cubic model. However, if Q is not small and $m > 2$, a

very high number of experimental runs $\left(C_m^{Q+m+1}\right)$ is required. Two examples of simplex lattice experimental designs appear in Figure 5.5.

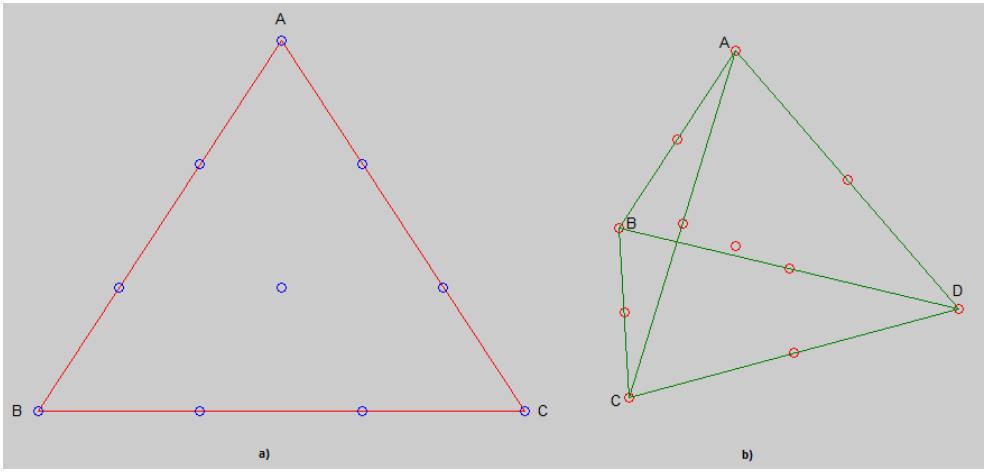


Figure 5.5 Examples of a) a simplex lattice design {3,3}, and b) a simplex lattice design {4,2}, to which the overall centroid of the simplex has been added

- Simplex centroid design:

This sort of design consists of 2^{Q-1} points corresponding to the centroids of all the *r*-dimensional elements of the envelope of the mixture space, from r=0 (vertices) to $r = Q - 1$ (overall simplex centroid). Each blend from this DOE meets that, for the corresponding *r*-dimensional element they are the centroid of, the proportion of $r + 1$ components in it is $\frac{1}{r+1}$, while all other constituents are absent. These designs are suitable for problems where the number of components is not very high, and to estimate models such as the special cubic one, which can be especially useful for studying the shape of the response surface and identify the model that better fits it. Figure 5.6 illustrates two examples of this sort of design.

It can easily be observed that, in the particular case of a problem with 3 components, the simplex centroid design is identical to the simplex lattice {3,2} when the general centroid of the mixture space is added to the latter.

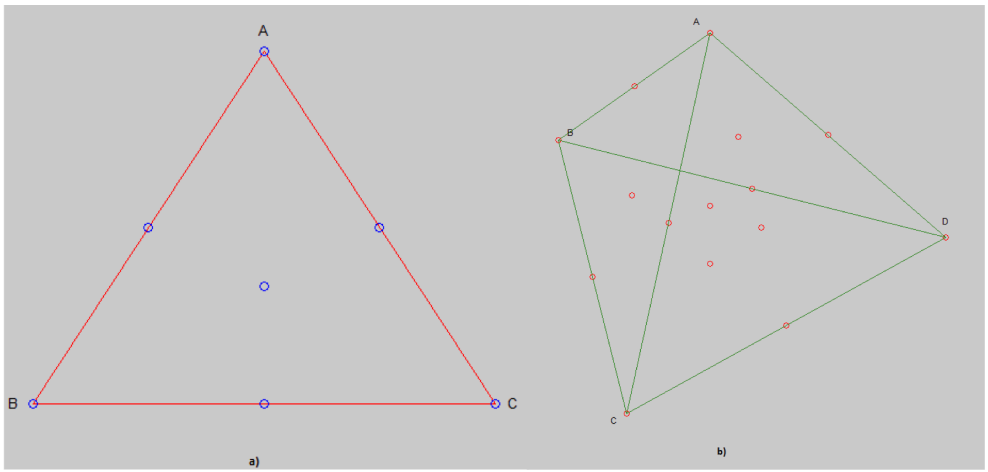


Figure 5.6 Examples of simplex centroid designs for a) Q=3, and b) Q=4

- Simplex axial designs:

Unlike simplex lattice or simplex centroid designs, these designs include points inside the mixture space (other than its overal centroid), and not just on its envelope. More specifically, these points are located on the axis (or Cox axis) associated with each blend constituent. As a reminder, any blend located on the axis associated with the i-th component meets that, if the proportion of such component in the mixture is x_i, then the proportions of all other ingredients are $x_j = (1 - x_i)/(Q - 1) \forall j \neq i$. There are two main variants of these designs:

- *Simplex axial 2Q+1 design*:

  This design includes the Q vertices of the simplex, its overall centroid (which belongs to the Q axes) and the Q interior points resulting from taking, for the i-th axis, the midpoint between the overall simplex centroid and the i-th vertex, for which the proportions of the blend ingredients are $x_i = \frac{(Q+1)}{2Q}$ and $x_j = \frac{1}{2Q} \forall j \neq i$. Figure 5.7 illustrates two examples of this kind of design.

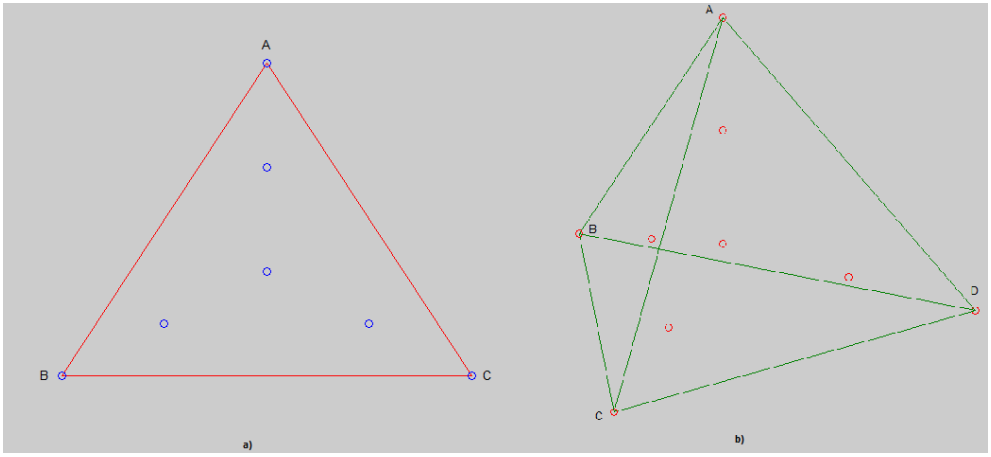


Figure 5.7 Examples of simplex axial 2Q+1 designs for a) Q=3, and b) Q=4

- *Simplex axial 3Q+1 design*:

  This design results from adding to the previous one the extreme points of the axes (i.e. the centroids of the facets of the simplex opposite to its vertices passing through its overall centroid) for which the proportions of the blend ingredients are $x_i = 0$ and $x_j = \frac{1}{Q-1} \forall j \neq i$. Figure 5.8 illustrates two examples of this kind of design.

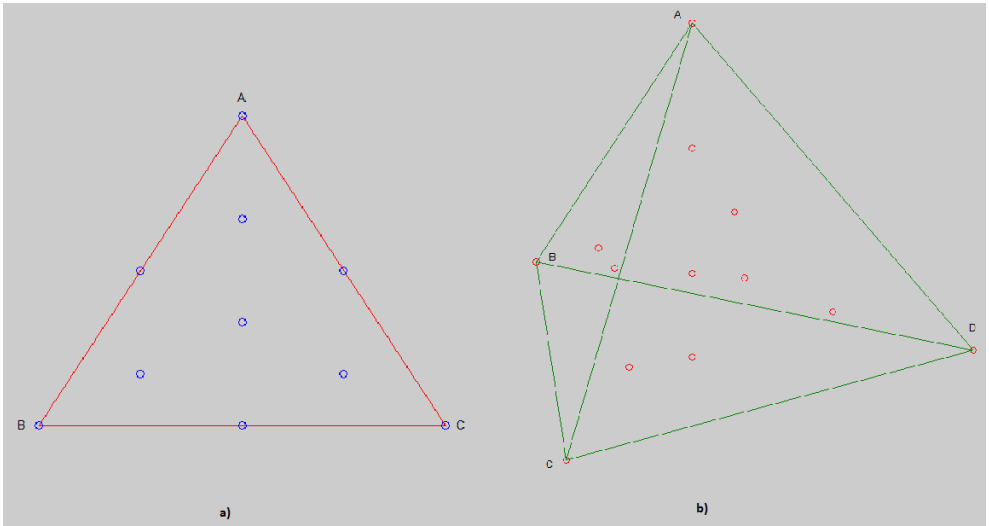


Figure 5.8 Examples of simplex axial 3Q+1 designs for a) Q=3, and b) Q=4

This type of designs are used mainly for screening purposes, or to study the nature of the effect of the components, since the estimates of the parameters of the model fitted with data from an axial design are less affected (compared to the simplex lattice and simplex centroid designs) by the bias associated to assuming a model less complex than the real one, whenever this occurs. However, axial designs do not allow the adjustment of, for example, the quadratic Scheffé model except when Q is small.

It should be stressed nonetheless that a proper DOE is not usually limited to the points suggested by these designs. On the contrary, it is common practice to add replications or additional runs to those covered by these basic designs, either to study the lack of fit of the model obtained or to obtain a better estimate of its parameters. An example of this is the augmented simplex centroid design, which combines the simplex centroid and the simplex axial $2Q$+1 designs. This design presents excellent properties to estimate the parameters of a special quartic model, which is very useful to study the shape of the response surface of the model. When $Q = 3$, this design is identical to the axial $3Q + 1$.

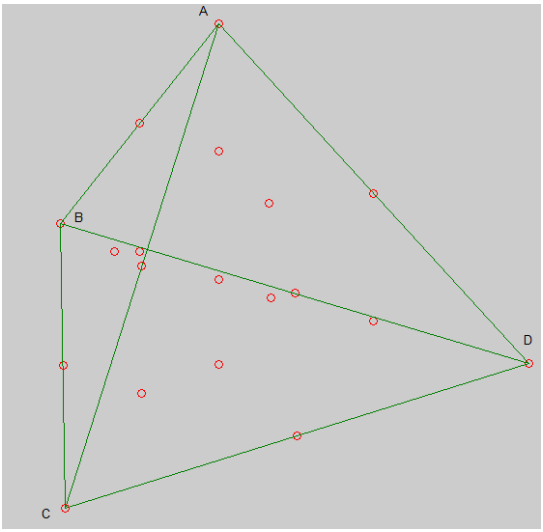


Figure 5.8 Example of an augmented simplex centroid design for Q=4

In general, whenever the number of points obtained with any of these designs is too high compared to the number of model parameters to be estimated, and/or excessively resources-consuming, a selection of the most appropriate runs to carry out is done by generating an optimal DOE according to some criteria (see Section 2.2 for additional information). Regarding the I-optimality criterion in particular, the volume of the simplex is necessary for its computation. Such volume can be obtained by means of the Cayley-Menger Determinant [77].

5.4.2. DOE in irregular mixture spaces

Whenever the mixture space is not a simplex, some of the most commonly resorted to designs are based on the use of vertices and centroids of edges, facets and other r-dimensional ($r \leq Q-1$) elements that constitute the mixture space as candidate points for the DOE matrix. Algorithms for the definition of extreme vertices such as the ones proposed by Snee & Marquadt [68] or McLean & Anderson [69] are good options to obtain all of the vertices of the envelope of the mixture space, from which all other centroids may be determined. The former of this algorithms, also known as the XVERT algorithm, is computationally more efficient than the last, and operates as follows:

1. The Q blend constituents are arranged from lowest to highest r_q (see Section 5.2.1).
2. A 2^{Q-1} factorial design is constructed using the lower and upper consistent bounds of the first $Q-1$ constituents as two levels of each factor.
3. The proportion of the remaining constituent is computed as $x_Q = 1 - \sum_{j=1}^{Q-1} x_j$. If $lb_Q^* \leq x_Q \leq ub_Q^*$, the new point is a vertex.
4. If no vertex was obtained in the previous step, then:
    a. If $x_Q < lb_Q^*$, make $x_Q = lb_Q^*$. Otherwise $x_Q = ub_Q^*$
    b. Q-1 additional points are generated by adjusting the proportions of the remaining Q-1 components so that $\sum_{j=1}^{Q} x_j = 1$. If the sum of the q components is 1. If $lb_j^* \leq x_j \leq ub_j^*$, a new vertex will have been obtained.
5. Possible duplicate points are eliminated.

Despite this algorithm being able to find all of the vertices of an irregular mixture space as long as no multivariate constraints are imposed on the proportions of the blend constituents (other than the one defined in Equation 5.1), the computational cost may increase excessively for very large values of Q. In that case, the modified XVERT or MXVERT algorithm can be resorted to instead. The MXVERT algorithm does not provide all vertices of the mixture space, but a 2^{Q-1-k} subset of it. However, this subset is easily obtained and usually has good statistical properties as a DOE or candidate set to construct one. The steps to apply it are:

1. The Q blend constituents are arranged from lowest to highest r_q (see Section 5.2.1).
2. A 2^{Q-1-k} factorial design is constructed using the lower and upper consistent bounds of the first $Q-1$ constituents as two levels of each factor.

3. The proportion of the remaining constituent is computed as $x_Q = 1 - \sum_{j=1}^{Q-1} x_j$. If $lb_Q^* \leq x_Q \leq ub_Q^*$, the new point is a vertex.
4. If no vertex was obtained in the previous step, then:
    a. If $x_Q < lb_Q^*$, make $x_Q = lb_Q^*$. Otherwise $x_Q = ub_Q^*$
    b. Adjust x_{Q-1} as much as the restriction $lb_{Q-1}^* \leq x_{Q-1} \leq ub_{Q-1}^*$ allows it to make $\sum_{j=1}^{Q} x_j$ as close to one as possible. If $\sum_{j=1}^{Q} x_j = 1$ after this adjustment, a new vertex has been obtained.
    c. If it was not possible to achieve $\sum_{j=1}^{Q} x_j = 1$ in 'b', repeat this step with x_{Q-2}, and if necessary with x_{Q-3}, x_{Q-4} … until x_1 if necessary. Once $\sum_{j=1}^{Q} x_j = 1$ is achieved, a new vertex has been obtained.
5. Possible duplicate points are eliminated.

Both the XVERT and MXVERT algorithm are limited to mixture problems without multivariate constraints on the constituents other that Equation 5.1. Furthermore, they cannot be used in mixture-amount, mixture-process variables nor mixture-amount-process variables problems when constraints on non-mixture variables are also imposed. Good alternatives in such case are the proposed variation of the CONSIM algorithm or the 'Segmentation by defining and discarding vertices' presented in Section 5.2.2, followed by a variation of the coordinate-exchange algorithm for mixture design proposed by Piepel et al. [66]. Consider that all inequality and equality restrictions have been checked, and it has been found that any $\mathbf{x}$ point within the experimental space (this can be applied to any problem where the experimental space is a convex hull) must meet that

$$\begin{aligned} \mathbf{A}_{\mathbf{x}}^* \cdot \mathbf{x} &\leq \mathbf{d}_{\mathbf{x}}^* \\ \mathbf{F}_{\mathbf{x}}^* \cdot \mathbf{x} &= \mathbf{f}_{\mathbf{x}}^* \end{aligned} \tag{5.27}$$

where $\mathbf{A}_{\mathbf{x}}^*$, $\mathbf{d}_{\mathbf{x}}^*$, $\mathbf{F}_{\mathbf{x}}^*$, and $\mathbf{f}_{\mathbf{x}}^*$ are differentiated of $\mathbf{A}_{\mathbf{x}}$, $\mathbf{d}_{\mathbf{x}}$, $\mathbf{F}_{\mathbf{x}}$ and $\mathbf{f}_{\mathbf{x}}$ in that they refer exclusively to the active, non-redundant restrictions that define the envelope of the experimental space. If a design with *N* runs is desired, the steps one should follow to apply this generalized coordinate-exchange algorithm is the following:

1. Generate a starting design constituted by *N* random feasible design points.
2. Sequentially modify the coordinates of each point, starting from the first one, and focusing on modifying one variable at a time, starting again with the first one. Since some equality restrictions may prevent the coordinate associated to the *m*-th the variable to be changed independently, the procedure to modify the coordinates of a given $\mathbf{x}_n$ is:

i. Generate a vector $\mathbf{o}_m$ [$M \times 1$], which is a vector of zeros except for its m-th element, which is a one. This vector would provide the direction of maximum variability for the m-th variable if no equality constraints were imposed that affected it.

ii. Project $\mathbf{o}_m$ onto the subspace generated by the active, non-redundant equality restrictions. It can be demonstrated (see Section 14.2.2) that this projection, $\mathbf{o}_m^*$, is:

$$\mathbf{o}_m^* = \left[\mathbf{I}_m - \mathbf{F}_\mathbf{x}^{*\mathrm{T}} \cdot \left(\mathbf{F}_\mathbf{x}^* \cdot \mathbf{F}_\mathbf{x}^{*\mathrm{T}}\right)^{-1} \cdot \mathbf{F}_\mathbf{x}^*\right] \cdot \mathbf{o}_m + \mathbf{F}_\mathbf{x}^{*\mathrm{T}} \cdot \left(\mathbf{F}_\mathbf{x}^* \cdot \mathbf{F}_\mathbf{x}^{*\mathrm{T}}\right)^{-1} \cdot \mathbf{f}_\mathbf{x}^* \tag{5.28}$$

iii. Vary the coordinates of $\mathbf{x}_n$ following the direction of $\mathbf{o}_m^*$ and $-\mathbf{o}_m^*$ without moving outside of the experimental region (i.e. making sure that the inequality restrictions are always met), and assess the value for the optimality criteria used to construct the DOE.

iv. Repeat steps i-iii for the next coordinates, until all M variables have been modified following their feasible direction of maximum variability $\mathbf{o}_m^*$ (or $-\mathbf{o}_m^*$).

3. Repeat step 2 for the N design points.
4. Repeat steps 2 and 3 until no significant improvement in the DOE optimality criteria is observed.

Regarding the I-optimality criteria, again the volume of the experimental space needs to be calculated. To do so, it is proposed to sequentially divide each r-dimensional subspace ($2 \leq r \leq M$) whose union constitutes the whole experimental space into as many $(r-1)$-dimensional hyperpyramids as facets delimit it, such that these facets constitute their bases, and the opposite vertex is the centroid of the r-dimensional subspace. By sequentially calculating the volume of these hyperpyramids, and then calculating their sum, the volume of the r-dimensional can be computed exactly. Although there is no need to use this procedure in such a situation, consider a cube as an example. The steps to follow would be:

1. Divide the cube into 6 pyramids (3-dimensional hyperpyramids) whose bases are the facets of the cube, and the centroid of the cube is the vertex opposite to those bases, common to all the pyramids.
2. Divide the base of each pyramid into 4 triangles (2-dimensional hyperpyramids) whose bases are the edges of the cube/sides of the squares, and the centroid of one face is the vertex opposite to the base of each triangle in that face, which is common to all of the triangles in the same facet.
3. Calculate the areas of the triangles, and sum them up to obtain the area of each side of the cube, which is the area of the base of each one of the pyramids generated in the first step.

4. Calculate the volumes of the 6 pyramids. Their sum is the volume of the cube.

The volume of an r-dimensional hyperpyramid is equal to the volume of its $(r-1)$-dimensional base multiplied by its height, and divided by r. Its height is equal to the distance from the base to the opposite vertex, which can be calculated by projecting that vertex on the $(r-1)$-dimensional base by means of Equation 5.28, and computing the distance between that vertex and its projection.

Chapter 6
Latent variable-based methods for mixture data analysis

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

1. Vitale, R.[vi], Palací-López, D.[vi], Kerkenaar, H., Postma, G., Buydens, L. & Ferrer, A. Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments. *Chemometr. Intell. Lab.* 175, 37-46 (2018).

6.1. Mixture design data analysis with Partial Least Squares

Since Partial Least Squares regression (PLS) is a method based on projection to latent structures, it does not assume independence of the factors, and is therefore capable of operating in cases where there is collinearity, either perfect or imperfect, or both simultaneously. This means that, as opposed to classical polynomial model fitting approaches like OLS, no special model is required in order to use PLS-regression techniques when dealing with mixture data. Furthermore, additional sources of correlation are automatically accounted with PLS, and there is no need to treat mixture and process variables differently, since there is actually no difference in practice between the existing perfect collinearity relationship among mixture factors and amount or process factors upon which an equality constraint is imposed, or which are affected by collinearity. It must be taken into account, however, that PLS is a method of projection to latent structures and, therefore, it is necessary to properly choose the dimensionality of the model, that is, the number of latent variables A to extract. To do this a cross-validation test may be sufficient to make a decision. When analysing mixture data with PLS, the unscaled, uncentred regression coefficients $\boldsymbol{\beta}$ that relate the ingredients' proportions in

[vi] These authors had equal contributions

the blend, **X**, with the output variable, **y**, can be interpreted in the same way as the parameters of the Cox model. This means that the Scheffé parameters can be estimated from the coefficients obtained through PLS regression. Furthermore, These regression coefficients are identical to those for the Cox model if the number of selected latent variables to fit the PLS model, *A*, is equal to the number of non-zero singular values of **X**.

In this section some examples available in the literature are analysed using both OLS and PLS regression techniques in order to better illustrate some of the advantages of using PLS for mixture design data analysis.

6.1.1. Methods and datasets

To assess the differences in the interpretability of the results obtained by using OLS and PLS for the study of mixture problems, both techniques will be applied to the two real cases studies presented below.

6.1.1.1 Example 1: seven-component octane blending experiment of Cornell

The dataset used for this study is available in [78]. In this simple example the linear model will be fitted, and used to illustrate one of the major drawbacks of using classical regression techniques for mixture data analysis, which is the difficulty in detecting similarly behaving blend components, and the benefits of using PLS in this case.

Table 6.1 shows the different blend constituents and the lower and upper bounds for their proportions in the mixture, which define the mixture space.

Table 6.1. Example 1: blend ingredients and lower and upper bounds for their proportions in the mixture

Blend constituent	Lower bound	Upper bound
Straight run	0	0.21
Reformate	0	0.62
T.C. naphtha	0	0.12
C.C. naphtha	0	0.62
Polymer	0	0.12
Alkylate	0	0.74
Natural gasoline	0	0.08

6.1.1.2 Example 2: gasoline blending data of Snee

The dataset used for this study can be found in [79]. In this example the potential relationship between the research octane number (RON) and the proportions of the blend constituents in it is evaluated. Both univariate and multivariate restrictions are imposed on the composition of the blend, resulting in an irregular mixture space. The quadratic model will be fitted via OLS and PLS, and the results compared.

Table 6.2 shows the different blend constituents and the lower and upper bounds for their proportions in the mixture.

Table 6.2. Example 2: blend ingredients and lower and upper bounds for their proportions in the mixture

Blend constituent	Lower bound	Upper bound
Butane (B)	0	0.15
Isopentane (I)	0	0.30
Reformate (R)	0	0.35
Cat. cracked (C)	0	0.60
Alkylate (A)	0	0.60

In addition to these bounds, the following restrictions are imposed on the blend constituents' rates in the mixture, which also affect the shape of the mixture space (i.e. none of them are redundant):

$$\begin{gathered} B + I \leq 0.30 \\ C + A \leq 0.70 \\ 97 \leq 101.8 \cdot B + 99.6 \cdot I + 112.4 \cdot R + 94.2 \cdot C + 99.8 \cdot A \leq 101 \end{gathered} \tag{6.1}$$

6.1.2. Results and discussion

6.1.2.1 Example 1: seven-component octane blending experiment of Cornell

In this example, the linear Scheffé polynomial is fitted for screening purposes using a reduced number of observations corresponding to a series of blends with the 7 constituents within the mixture space described in Section 6.1.1.1. The response variable of interest is the octane of the blend. The summary of the results from fitting this model via OLS is shown in Table 6.3.

Table 6.3. (First) linear Scheffé polynomial parameters' estimation via OLS for example 1

Parameter	Estimate	SE
β_1	34.32	209.13
β_2	85.923	1.2482
β_3	141.25	375.33
β_4	77.18	9.2135
β_5	87.75	5.8157
β_6	100.3	3.4738
β_7	116.92	81.096

As commented in Section 5.3.1, the *i*-th linear parameter β_i corresponds to the expected average value for the response variable *y* (in this case the RON) for the 'pure mixture' with only the *i*-th ingredient. Therefore, studying the statistical significance of the β_i coeficients is meaningless because there is no point in testing whether they are zero or not. Instead, the statistical significance of the adjusted effects (calculated as in Section 5.3.1) as a whole and of each one of them is assessed as illustrated in Section 5.8 in [57]. The results are shown in Table 6.4.

Table 6.4. Adjusted effects calculation and statistical significance assessment from the (first) linear Scheffé polynomial via OLS for example 1 (* on the right indicates p-value<0.05)

Blend constituent	Adjusted effect estimate	p-value	
Overall	-	$7.5 \cdot 10^{-5}$	*
x_1 (Straight run)	-14.12	0.81	
x_2 (Reformate)	-4.36	0.76	
x_3 (T.C. naphtha)	6.90	0.90	
x_4 (C.C. naphtha)	-10.68	0.44	
x_5 (Polymer)	-0.59	0.83	
x_6 (Alkylate)	7.21	0.66	
x_7 (Natural gasoline)	2.33	0.79	

Table 6.4 shows apparently contradictory results, since, although there seem to exist an overall statistical significant effect of the ingredients of the mixture on the octane, none of the tests performed to assess the statistical significance of the effect of each individual blend constituent permits identifying any of them as having a statistically significant effect on the RON. However, a high degree of correlation between components x_1 (Straight run) and x_3 (T.C. naphtha) should be noted, as can be seen in Table 6.5, where cells with shading correspond to correlation coefficients whose absolute value is greater than or equal to 0.7.

Table 6.5. Correlation matrix for the parameters of the (first) linear Scheffé polynomial fitted via OLS for example 1

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
β_1	1	0.0360	-0.9983	-0.5766	-0.3137	-0.5653	0.5987
β_2	0.0360	1	-0.0586	-0.1862	-0.1122	-0.3226	0.2133
β_3	-0.9983	-0.0586	1	0.6201	0.3358	0.6099	-0.6432
β_4	-0.5766	-0.1862	0.6201	1	0.4394	0.9535	-0.9924
β_5	-0.3137	-0.1122	0.3358	0.4394	1	0.3241	-0.4753
β_6	-0.5653	-0.3226	0.6099	0.9535	0.3241	1	-0.9649
β_7	0.5987	0.2133	-0.6432	-0.9924	-0.4753	-0.9649	1

Table 6.6 provides the results from combining both constituents and treating them as a new one to fit a new linear Scheffé model.

Table 6.6. (Second) linear Scheffé polynomial parameters' estimation via OLS for example 1

Parameter	Estimate	SE
$\beta_1 + \beta_3$	72.57	7.2776
β_2	85.934	1.1418
β_4	76.16	6.7208
β_5	87.401	5.032
β_6	99.923	2.5591
β_7	126.24	57.834

On the other hand, the results from assessing the statistical significance of the adjusted effects for this new scenario can be seen in Table 6.7.

Table 6.7. Adjusted effects calculation and statistical significance assessment from the (second) linear Scheffé polynomial via OLS for example 1 (* on the right indicates p-value<0.05)

Blend constituent	Adjusted effect estimate	p-value	
(Overall)	-	$5.5 \cdot 10^{-6}$	*
$x_1 + x_3$	-7.45	0.22	
x_2	-4.04	0.44	
x_4	-11.32	0.30	
x_5	-0.57	0.69	
x_6	7.59	0.39	
x_7	3.35	0.52	

As with Table 6.4, Table 6.7 does not allow identifying the adjusted effect of any of the mixture ingredients, individually, as statistically significant. However, a high degree of correlation between the estimated effects of $(x_1 + x_3)$ and x_7 should be noted, as can be seen in Table 6.8, where cells with shading correspond to correlation coefficients whose absolute value is greater than or equal to 0.7.

Table 6.8. Correlation matrix for the parameters of the (second) linear Scheffé polynomial fitted via OLS for example 1

	$\beta_1 + \beta_3$	β_2	β_4	β_5	β_6	β_7
$\beta_1 + \beta_3$	1	-0.3820	0.9179	0.3939	0.9310	-0.9589
β_2	-0.3820	1	-0.1957	-0.1013	-0.3643	0.2335
β_4	0.9179	-0.1957	1	0.3204	0.9275	-0.9884
β_5	0.3939	-0.1013	0.3204	1	0.1701	-0.3664
β_6	0.9310	-0.3643	0.9275	0.1701	1	-0.9453
β_7	-0.9589	0.2335	-0.9884	-0.3664	-0.9453	1

Table 6.9 provides the results from combining the three constituents and treating them as a new one to fit a third (and last) linear Scheffé model.

Table 6.9. (Third) linear Scheffé polynomial parameters' estimation via OLS for example 1

Parameter	Estimate	SE
$\beta_1 + \beta_3 + \beta_7$	78.40	1.80
β_2	85.70	1.08
β_4	81.64	1.15
β_5	88.95	4.57
β_6	101.93	0.80

R^2=0,9916 R^2-adj=0,9868

Finally, the results from assessing the statistical significance of the adjusted effects for this third scenario can be seen in Table 6.10.

Table 6.10. Adjusted effects calculation and statistical significance assessment from the (third) linear Scheffé polynomial via OLS for example 1 (* on the right indicates p-value<0.05)

Blend constituent	Adjusted effect estimate	p-value	
(Overall)	-	$4.8 \cdot 10^{-7}$	*
$x_1 + x_3 + x_7$	-4.57	$1.5 \cdot 10^{-3}$	*
x_2	-1.26	0.27	
x_4	-4.41	0.01	*
x_5	0.24	0.68	
x_6	13.51	$1.8 \cdot 10^{-5}$	*

As has been observed, three consecutive adjustments of a Scheffé model via OLS have been required, coupled with a careful inspection of the existing collinearity among the different components, in order to conclude that three of them could be considered, for the purposes of assessing their influence on the response, equivalent.

If, instead, the linear Cox polynomial is fitted via PLS regression (A=3), the estimations of the parameters shown in Table 6.11 are obtained.

Table 6.11. Linear Cox polynomial parameters' estimation via PLS (A=3) for example 1

Coefficient	Scaled/centred estimate	Unscaled/uncentred estimate
β'_0	0	92.6759
β'_1	-0.1391	-9.8281
β'_2	-0.2087	-6.9602
β'_3	-0.1376	-16.6662
β'_4	-0.2932	-8.4218
β'_5	-0.0384	-4.3887
β'_6	0.4564	10.1615
β'_7	-0.1434	-34.5288

R^2=0.99056 R^2-adj=0.97922 Q^2=0.90208

In this case it is enough to observe that, for the components x_1 (Straight run), x_3 (T.C. naphta) and x_7 (Natural Gasoline), the centred and scaled coefficients are very similar. The estimated coefficients (without centring or scaling) are very different due to the difference in the ranges of possible variation of the composition of each of the components in the mixture. In fact, the highest (absolute) value is detected for x_7, the constituent with the narrowest allowed range of compositions.

In addition, it is possible to observe from the scaled and centred coefficients that the largest positive effect corresponds to x_6 (alkylate), the next most important components being x_4 (C.C. naphtha) and then the combination of x_1 (Straight run), x_3 (T.C. naphta) and x_7 (Natural Gasoline). The range of variation in composition for x_2 (reformate) is very high, and therefore its total effect will be lower in spite of the value for its estimated centred and scaled parameter being higher. On the other hand, x_5 (polymer) is the one constituent that presents a smaller effect on the response, and can be considered practically null.

6.1.2.2 Example 2: gasoline blending data of Snee

In this example, the effect of 5 blend constituents on the RON of a gasoline mixture was studied. The estimations of the parameters from fitting a quadratic Scheffé model with the data provided via OLS is shown in Table 6.12.

From the Table 6.12 and Figure 6.1 it may seem that the fitted model is quite appropriate to explain the response variable. It should be noted, however, that although many of the terms corresponding to two-factor interactions appear as statistically not significant, the correlation matrix (see Table 6.13) indicates a high collinearity among several of the estimated coefficients (cells with shading correspond to correlation coefficients whose absolute value is greater than or equal to 0.7).

Furthermore, and according to the model fitted, the effect of butane (x_1; B in Figure 6.2) seems to show a slight degree of non-linearity in the mixture, as is also reflected in the trace plot in Figure 6.2. This is in agreement with the statistical significance of the parameters of the model corresponding to the two-factor interactions, since only those in which the butane intervenes appear as clearly statistically significant. Additionally, these interactions are all negative, that is, the interaction of butane with the rest of the components of the mixture is antagonistic. The trace graph also allows us to observe this clearly, since the curve corresponding to the butane is convex with respect to the origin.

Table 6.12. Quadratic Scheffé polynomial parameters' estimation via OLS for example 2 (* on the right indicates p-value<0.07)

Parameter	Estimate	SE	tStat	p-value	
β_1	158.508	22.690	-	-	
β_2	97.384	7.139	-	-	
β_3	106.935	8.029	-	-	
β_4	92.342	2.365	-	-	
β_5	100.057	1.897	-	-	
β_{12}	-56.029	27.468	-2.040	0.0661	*
β_{13}	-88.109	27.483	-3.206	0.0084	*
β_{14}	-62.866	28.776	-2.185	0.0514	*
β_{15}	-57.709	26.594	-2.170	0.0528	*
β_{23}	-5.978	10.297	-0.581	0.5732	
β_{24}	7.005	14.155	0.495	0.6304	
β_{25}	6.054	14.126	0.429	0.6765	
β_{34}	10.913	17.632	0.619	0.5486	
β_{35}	1.670	12.434	0.134	0.8956	
β_{45}	3.368	2.720	1.238	0.2414	

R^2=0.98262 R^2-adj=0.9605

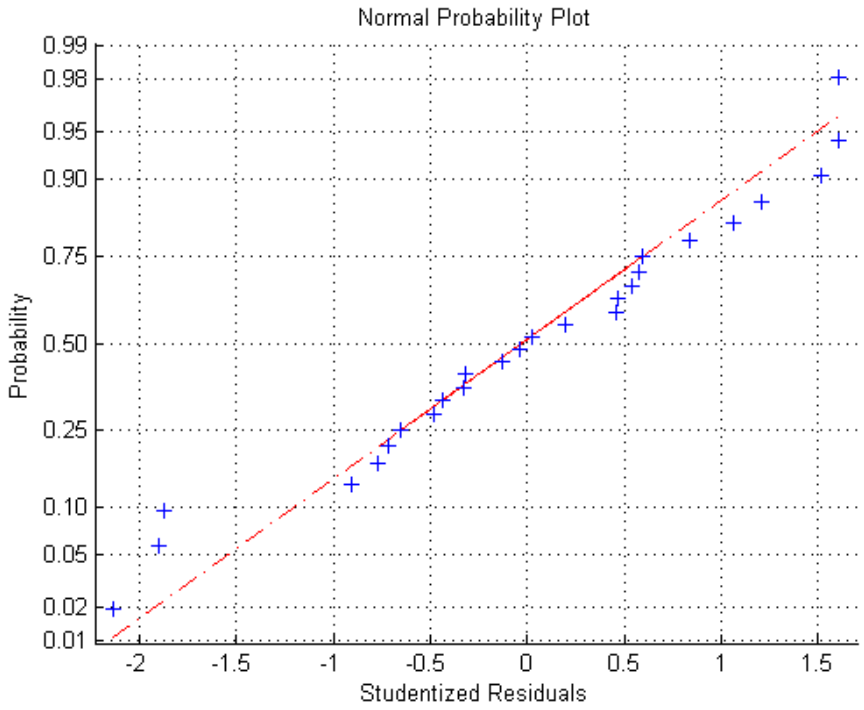


Figure 6.1 Studentized residual plots for the quadratic Scheffé model fitted via OLS for example 2

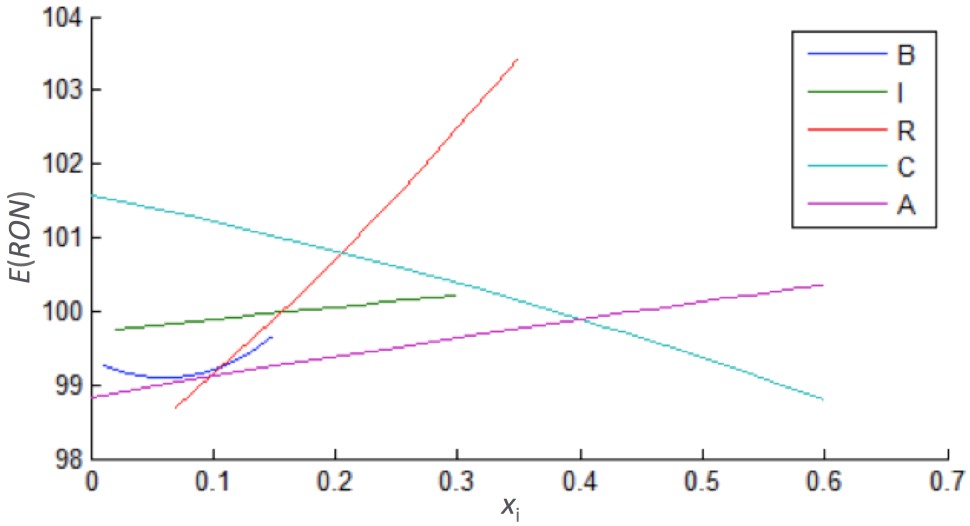


Figure 6.2. Trace plot for the quadratic Scheffé model fitted via OLS for example 2

Table 6.13. Correlation matrix among the estimates of the quadratic Scheffé polynomial parameters fitted via OLS for example 2

	B	I	R	C	A	B·I	B·R	B·C	B·A	I·R	I·C	I·A	R·C	R·A	C·A
B	1	-0.222	0.522	0.499	0.296	-0.038	0.040	0.381	0.332	-0.395	-0.766	-0.620	-0.157	-0.624	-0.412
I	-0.222	1	0.003	-0.113	-0.071	-0.958	-0.949	-0.921	-0.914	0.090	0.174	0.095	0.071	0.145	0.062
R	0.522	0.003	1	0.329	0.664	-0.160	-0.294	0.181	0.164	-0.527	-0.461	-0.700	-0.089	-0.789	-0.844
C	0.499	-0.113	0.329	1	0.071	-0.034	-0.011	0.212	0.199	-0.066	-0.935	-0.895	-0.596	-0.244	-0.173
A	0.296	-0.071	0.664	0.071	1	-0.012	-0.087	0.112	-0.003	-0.447	-0.181	-0.319	-0.451	-0.878	-0.959
B·I	-0.038	-0.958	-0.160	-0.034	-0.012	1	0.964	0.811	0.827	-0.021	0.038	0.092	-0.024	0.024	0.057
B·R	0.040	-0.949	-0.294	-0.011	-0.087	0.964	1	0.796	0.783	0.031	-0.007	0.133	-0.055	0.079	0.159
B·C	0.381	-0.921	0.181	0.212	0.112	0.811	0.796	1	0.912	-0.127	-0.315	-0.266	-0.055	-0.257	-0.157
B·A	0.332	-0.914	0.164	0.199	-0.003	0.827	0.783	0.912	1	-0.065	-0.282	-0.245	0.05	-0.184	-0.076
I·R	-0.395	0.090	-0.527	-0.066	-0.447	-0.021	0.031	-0.127	-0.065	1	0,199	0.233	0.175	0.375	0.505
I·C	-0.766	0.174	-0.461	-0.935	-0.181	0.038	-0.007	-0.315	-0.282	0.199	1	0.911	0.498	0.434	0.304
I·A	-0.620	0.095	-0.700	-0.895	-0.319	0.092	0.133	-0.266	-0.245	0.233	0.911	1	0.415	0.536	0.487
R·C	-0.157	0.071	-0.089	-0.596	-0.451	-0.024	-0.055	-0.055	0.05	0.175	0.498	0.415	1	0.284	0.347
R·A	-0.624	0.145	-0.789	-0.244	-0.878	0.024	0.079	-0.257	-0.184	0.375	0.434	0.536	0.284	1	0.923
C·A	-0.412	0.062	-0.844	-0.173	-0.959	0.057	0.159	-0.157	-0.076	0.505	0.304	0.487	0.347	0.923	1

Alternatively, the quadratic Cox model can be obtained by fitting a PLS-regression model (A=4), which will provide the estimations for the Cox coefficients shown in Table 6.14. It should be noted that, in this case, the values of the parameters associated with the quadratic term for butane (β'_{11}) and its interactions with other components remain high in comparison with the rest. The same is apparently no longer true for other interactions, while interactions such as the ones between isopentane (I) and reformate (R), and between isobutane and the catalytically cracked constituent of the blend (C) show relatively high values for their parameters estimation.

These observed differences between the results obtained by OLS and PLS show how some interactions with a significant effect on the response have not been detected as such by OLS, due to the issues this regression methods suffers from when analysing mixture data in highly restricted or irregular mixture spaces, as is the case. By means of PLS, however, values have been obtained that seem to be more in line with the importance of these interactions.

To make a fairer comparison between OLS and PLS, one must consider that the Cox polynomial parameters estimated via PLS with $A = r_{\mathbf{X}}$ (A being the number of LV extracted to adjust the model and $r_{\mathbf{X}}$ the rank of $\mathbf{X}$) are identical to those of the Cox model obtained via OLS (from the Scheffé one) if the overall centroid of the dataset (i.e. the centre of projection in PLS) is chosen as the reference mixture.

Table 6.14. Quadratic Cox polynomial parameters' estimation via PLS (A=4) for example 2

Parameter	Scaled/centred estimate	Unscaled/uncentred estimate
β'_0	0	99.5439
β'_1	-0.0620	-1.3059
β'_2	-0.0993	-1.8248
β'_3	0.4579	5.4746
β'_4	-0.5798	-3.2736
β'_5	0.3792	2.0461
β'_{12}	0.0985	13.0510
β'_{13}	-0.0557	-7.1456
β'_{14}	-0.0553	-2.2676
β'_{15}	0.0138	1.6159
β'_{23}	0.0384	9.2189
β'_{24}	-0.0414	-2.0978
β'_{25}	-0.0215	-0.4896
β'_{34}	0.1657	4.5111
β'_{35}	-0.1210	-7.2758
β'_{45}	0.0674	0.8566
β'_{11}	0.1168	17.7973
β'_{22}	0.0274	0.1538
β'_{33}	-0.0615	-5.5928
β'_{44}	-0.1161	-0.6575
β'_{55}	0.0016	0.4425

R^2=0.96228 R^2-adj=0.92746 Q^2=0.7323

6.1.3. Conclusions and additional considerations

In this section one of the major drawbacks of using classical regression techniques for mixture design data analysis, consisting in the difficult identification of blend constituents whose influence in the mixture properties is similar to each other, has been shown, as well as how resorting to PLS may be useful in avoiding this problem.

Furthermore, as already stated, the model fitted with PLS is equivalent to the reparametrization of the Scheffé model into the Cox polynomial (of the same order) if the number of LV extracted to fit the PLS model is the maximum feasible, and the centre of projection of the latent subspace is taken as the reference mixture (in Section 6.2 different examples will be illustrated in which the same Scheffé model is obtained in different scenarios through OLS and PLS, given these conditions are met). Because of this, PLS-regression can be resorted to in order to obtain the same model as with OLS when analysing mixture data while, at the same time, bypassing the limitations of this approach. This being of interest or not in practice is left to debate, given that e.g. one of the advantages of latent variable-based methods is the reduction in dimensionality/compression of the information, which is lost if as many latent variables as possible are extracted.

From a more practical point of view, in terms of the available software for mixture design data analysis, it is also important to mention that not just any statistical software prepared for data analysis via OLS/GLS can be resorted to for this purpose, while any software aimed at data analysis via latent variable-based methods can also be used for mixture design model fitting. This is to say, specific/specialized software is required if traditional approaches are to be resorted to when dealing with mixture data, while no specific software is required when using latent variable-based algorithms.

Finally, it must also be pointed out that most latent-variable based methods present an important limitation (although not just when it comes to mixture design), which is the lack of a standardized approach to determine the statistical significance of e.g. the effects/estimations of the parameters of the fitted regression model, specially in absence of replicates in the dataset available to fit such model. This is the reason why such tests did not appear in Tables 6.11 and 6.14.

6.2. Kernel-PLS and pseudo-sample trajectories

As previously discussed, PLS regression-based techniques can be resorted to in order to avoid some of the issues associated to the use of classical polynomial fitting by traditional methods such as OLS or GLS. In fact, PLS regression based-techniques have proven to guarantee satisfactory performance even when highly restricted mixture spaces have been dealt with and allow variables of different nature (e.g. component proportions and physicochemical properties, as well as production process conditions) to be fused and simultaneously analysed. Nevertheless, if the mixture data under study are affected by strong non-linear relationships (which is rather common in e.g. indus-

trial scenarios), applying classical PLS (even taking into account additional interaction and/or higher-degree terms) may not constitute an appropriate modelling strategy since it assumes their underlying structure is linear [39]. A good alternative may be represented by the combination of K-PLS regression and pseudo-sample trajectories, but the described way of defining the different pseudo-sample matrices (see Section 3.2.2) is not suitable when mixture problems are concerned, because it violates the constraint in Equation 5.1 (i.e. it is impossible to vary the composition of any one of the constituents independently from the rest), and then needs to be slightly adapted.

The main aim of this section is to evaluate the potential of such a combination in this particular field of interest, by comparing it with well-established methodologies, i.e. Scheffé model fitting by means of OLS and Cox model fitting by means of PLS. Both simulated and real case studies will be investigated.

6.2.1. Methods

PLS and K-PLS regression have been presented in Sections 3.2 and 3.3, and the basic principles of the Scheffé and Cox models in Sections 5.3.1 and 5.3.2. The extension of the pseudo-sample projection approach for mixture data handling is now illustrated.

6.2.1.1 Pseudo-sample trajectories for mixture data

In order to account for the perfect collinearity mixture constraint in Equation 5.1, the pseudo-samples matrices $\mathbf{V}_q$ (similar to $\mathbf{V}_m$ in Section 3.3.2) should be structured in such a way that the V values in their q-th column range from the minimum to the maximum proportion of the q-th ingredient and all the elements of each one of their rows sum up to 1. E.g. if a ternary mixture problem is faced, a hypothetical $\mathbf{V}_1$ may have such an aspect:

$$\mathbf{V}_1 = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.2 & 0.4 & 0.4 \\ 0.4 & 0.3 & 0.3 \\ 0.6 & 0.2 & 0.2 \\ 0.8 & 0.1 & 0.1 \\ 1 & 0 & 0 \end{bmatrix} \tag{6.2}$$

More generally:

$$\mathbf{V}_q = \begin{bmatrix} \frac{1-v_{1,q}}{Q-1} & \frac{1-v_{1,q}}{Q-1} & \cdots & \min(\mathbf{x}_q) & \cdots & \frac{1-v_{1,q}}{Q-1} & \frac{1-v_{1,q}}{Q-1} \\ \frac{1-v_{2,q}}{Q-1} & \frac{1-v_{1,q}}{Q-1} & \cdots & \vdots & \cdots & \frac{1-v_{1,q}}{Q-1} & \frac{1-v_{1,q}}{Q-1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \frac{1-v_{V,q}}{Q-1} & \frac{1-v_{1,q}}{Q-1} & \cdots & \max(\mathbf{x}_q) & \cdots & \frac{1-v_{1,q}}{Q-1} & \frac{1-v_{1,q}}{Q-1} \end{bmatrix} \quad (6.3)$$

where $\mathbf{x}_q$ is the q-th column vector of $\mathbf{X}$ and $v_{v,q}$ refers to the entry in the v-th row and q-th column of $\mathbf{V}_q$. As shown in Figure 6.3, this would mean spanning the mixture space (a simplex in this case) in the direction connecting the vertex associated to the 'pure mixture' composed by only the q-th constituent ([1, 0, 0], if q = 1) and the mid-point of its opposite side ([0, 0.5, 0.5], if q = 1)[vii], which for the complete simplex is also de so-called Cox's direction given the reference mixture [1/3, 1/3, 1/3]. As will be highlighted in Sections 6.2.3.3 and 6.2.3.4, the representation of the corresponding pseudo-sample trajectories yields the so-called trace plot, traditionally used in mixture design analysis to get an approximate idea of the linear and non-linear effects the change in the proportion of every q-th ingredient may have on the quality attribute of interest. However, because these effects are confounded with those due to the simultaneous variation of the proportion of the other blend constituents, a precise identification of each of the Scheffé polynomial coefficients cannot be achieved this way.

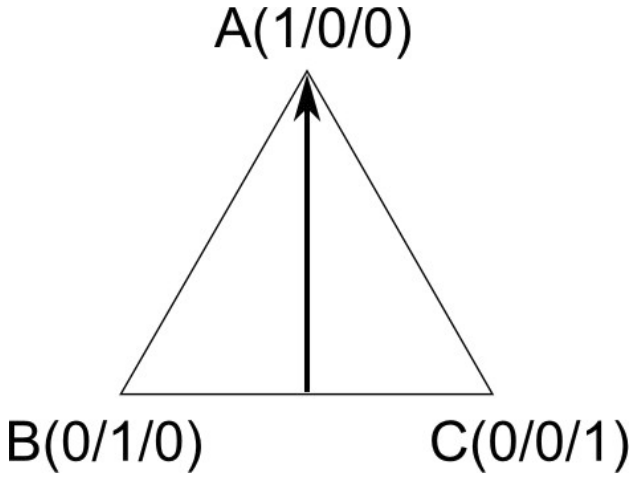


Figure 6.3. Graphical representation of the direction spanned by the pseudo-sample trajectory associated to the constituent A in a generic ternary mixture design space, coincident with the Cox's direction for the reference mixture [1/3, 1/3, 1/3]

[vii] Equation 6.2 is not valid if the design space is not a simplex or if it is a simplex but the ingredient proportions do not vary from 0 to 1, but the extension of the described methodology to such situations is straightforward.

6.2.1.2 Pseudo-sample-based response surface and Scheffé model coefficients

Alternatively, by using a combination of multiple pseudo-sample trajectories and graphing them in a contour plot, the response surface for the full mixture space can be retrieved. To do this, every pseudo-sample matrix has to be constructed by i) fixing the proportions of all but two constituents, ii) increasing the proportion of one of these two constituents, and iii) decreasing the proportion of the other accordingly (so that Equation 5.1 is met). This procedure is iteratively repeated for different values of the fixed proportions of the rest of the ingredients, which graphically implies moving over the design space in a particular direction, as displayed in Figure 6.4.

Notice that a measure of the Scheffé model coefficients for the first-order effects of the ingredients B and C and for their interaction can be derived from the trajectory covering the BC side of the simplex, as seen in Figure 5.3 and will be illustrated in Section 6.2.3.1. This is also valid for the trajectories covering the AB and the AC side of the triangle, not represented in Figure 6.4.

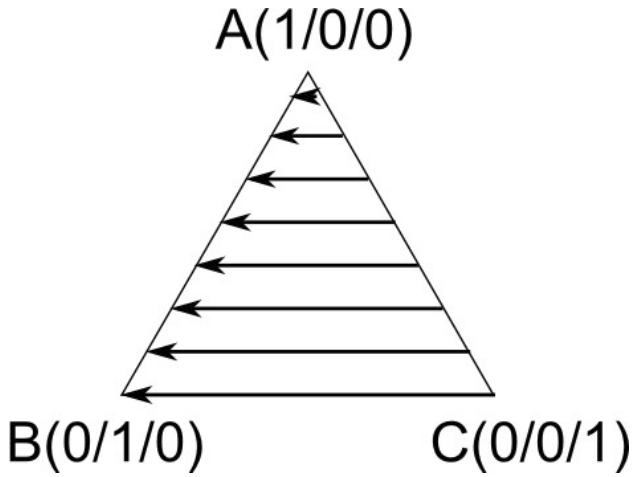


Figure 6.4. Graphical representation of the direction spanned by the pseudo-sample trajectories used to retrieve the response surface for a generic simplex for a ternary mixture. The resolution of the final plot will logically increase with the number of such trajectories. In this specific case, in every single pseudo-sample matrix, the proportion of A is fixed, while those for B and C vary

6.2.2. Datasets

Two simulated and real datasets from mixture designs of experiments will be object of this study.

6.2.2.1 Data simulated according to a second-order polynomial model

66 artificial samples (with no replicates) of a ternary mixture homogeneously distributed inside a simplex and a single response variable were simulated according to the following second-order Scheffé model:

$$\begin{gathered} y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_{1,2} \cdot x_1 \cdot x_2 + \beta_{1,3} \cdot x_1 \cdot x_3 + \beta_{2,3} \cdot x_2 \cdot x_3 \\ \beta_1 = 1.89\ ;\ \beta_2 = -1.33\ ;\ \beta_3 = 0.67 \\ \beta_{1,2} = -2.89\ ;\ \beta_{1,3} = 0.54\ ;\ \beta_{2,3} = -1.33 \\ x_i \epsilon \{0,1\} \qquad s.t. \sum_{i=1}^{3} x_i = 1 \end{gathered} \tag{6.4}$$

whose reformulation as a Cox model for a reference mixture where $s_1 = s_2 = s_3 = \frac{1}{3}$ can be written as (see Section 5.3.2):

$$\begin{gathered} y = \alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \alpha_3 \cdot x_3 + \alpha_{1,2} \cdot x_1 \cdot x_2 + \alpha_{1,3} \cdot x_1 \cdot x_3 + \\ + \alpha_{2,3} \cdot x_2 \cdot x_3 + \alpha_{1,1} \cdot x_1^2 + \alpha_{2,2} \cdot x_2^2 + \alpha_{3,3} \cdot x_3^2 \\ \alpha_0 = 0\ ;\ \alpha_1 = 2\ ;\ \alpha_2 = -2.67\ ;\ \alpha_3 = 0.67\ ;\ \alpha_{1,2} = -1.67 \\ \alpha_{1,3} = 0.44\ ;\ \alpha_{2,3} = 0\ ;\ \alpha_{1,1} = -0.11\ ;\ \alpha_{2,2} = 1.33\ ;\ \alpha_{3,3} = 0 \\ x_i \epsilon \{0,1\} \qquad s.t. \begin{cases} \sum_{i=1}^{3} x_i = 1 \\ \sum_{i=1}^{3} \alpha_i \cdot s_i = 0 \\ \sum_{j=1}^{3} c_{i,j} \cdot \alpha_{i,j} \cdot s_j = 0 \\ s_1 = s_2 = s_3 = 1/3 \end{cases} \end{gathered} \tag{6.5}$$

According to Equation 6.5, a positive first-order and a small negative second-order term characterize the apparent (note that the terms in this model are not independent from one another) effect of the first constituent on the output. Conversely, the second one features a negative first-order and a positive second-order term. The third ingredient exhibits a small positive first-order and no second-order term. Positive interaction terms were generated for both $x_1 \cdot x_2$ and $x_1 \cdot x_3$, while no interaction was assumed to involve x_2 and x_3. No noise was added after the data simulation.

6.2.2.2 Tablet data

This dataset was first described in [61], where 10 pharmaceutical tablets resulting from distinct blends of cellulose, lactose and phosphate were prepared to assess the influence of these substances on the release time of the active ingredient of the final manufactured drug. No replicates were performed.

6.2.2.3 Bubble data

The bubbles data relate to an experiment also reported in [61]. Here different proportions of two dish-washing liquids (DWL1 and DWL2), water and glycerol were combined to produce 24 soap mixtures (21 unique samples and 3 replicates) in order to determine which composition would yield the most durable bubbles (i.e. longest bubble lifetime) of a minimum acceptable size.

6.2.2.4 Colorant data

This dataset was described in [80]. In it, 49 blends (46 unique samples and 3 replicates) of different proportions of white ($\boldsymbol{C_w}$), black ($\boldsymbol{C_b}$), violet ($\boldsymbol{C_v}$) and magenta ($\boldsymbol{C_m}$) paints were manufactured to optimise the values of three specific colour responses: lightness ($\boldsymbol{L}^*$), red-green tone ($\boldsymbol{a}^*$) and yellow-blue tone ($\boldsymbol{b}^*$).

6.2.2.5 Gasoline data

An augmented simplex-centroid DOE was resorted to in order to generate the dataset used in this example, where different proportions of three gasoline constituents, catalytically cracked, C_5-isomer and reformate, were mixed to produce 10 distinct blend [81] to evaluate the effect of these constituents on the octane rating of the final product, and possibly maximise it.

6.2.2.6 Data simulated according to a highly non-linear model

12 artificial samples of a ternary mixture simulated according to an augmented simplex centroid design of experiments (with 2 replicates for the design centroid) were generated to obtain the dataset used in this example, based on the following model:

$$\begin{gathered} y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot log(x_1 + 0.01) + \\ + \beta_5 \cdot x_3^4 + \beta_6 \cdot sin[(1.01 - x_2)\pi] \\ \beta_1 = 4.87\ ;\ \beta_2 = -1.35\ ;\ \beta_3 = 5.67 \\ \beta_4 = -1.52\ ;\ \beta_5 = -0.35\ ;\ \beta_6 = 8.00 \\ x_i \epsilon \{0, 1\} \qquad s.t. \sum_{i=1}^{3} x_i = 1 \end{gathered} \tag{6.6}$$

Normally distributed random noise was added in this case to the response variable estimated by Equation 6.6.

6.2.3. Results

Both the simulated and real data were used for addressing an exploratory comparison among Scheffé polynomial fitting by means of OLS, Cox polynomial fitting by means of PLS, and K-PLS in terms of goodness-of-fit in calibration (R^2), goodness-of-fit in leave-one-out cross validation (Q^2), and Root Mean Square Error in leave-one-out Cross-Validation (RMSECV)[viii] [82], and for illustrating that under certain conditions K-PLS can guarantee improved prediction and interpretation. Moreover, a way of retrieving the coefficients of a Scheffé polynomial (when they hold) from the pseudo-sample trajectories yielded by a K-PLS model with the same complexity was derived. The whole set of routines resorted to for data processing and analysis was self-coded in MATLAB R2012b (Version 8.0.0.783) and is available on request.

6.2.3.1 Data simulated according to a second-order polynomial model

This section will be focused on demonstrating how the pseudo-sample trajectories can be resorted to in order to recover the coefficients of the Scheffé model in Equation 6.4 given dataset generated following the procedure outlined in Section 6.2.2.1. Figure 6.5 shows the curves for the response variable *y* (predicted by means of a 3-latent variable second-order K-PLS model) corresponding to the trajectories spanning the three sides of the ternary mixture space of the simulated dataset, by moving from a vertex (pure blend) to another vertex of the simplex which is like the one in Figure 6.4.

Note that, since every β_i $(\forall\, i \in \{1,3\})$ measures the expected *y* for the pure mixture composed by the only *i*-th constituent, each one of such parameters should match the predicted response at one of the two extremes of the respective pseudo-sample trajectory. As indicated in Figure 6.4, since the data at hand are noiseless, an exact correspondence was here observed for β_1, β_2 andβ_3. Analogously, the coefficients for the interaction terms $x_1 \cdot x_2$, $x_1 \cdot x_3$ and $x_2 \cdot x_3$ can be computed as:

$$\begin{aligned}
\beta_{1,2} &= \frac{\hat{y}_{0.5,0.5,0} - 0.5 \cdot \beta_1 - 0.5 \cdot \beta_2}{0.25} = \frac{-0.44 - 0.5 \cdot 1.89 - 0.5 \cdot (-1.33)}{0.25} = -2.89 \\
\beta_{1,3} &= \frac{\hat{y}_{0.5,0,0.5} - 0.5 \cdot \beta_1 - 0.5 \cdot \beta_3}{0.25} = \frac{1.42 - 0.5 \cdot 1.89 - 0.5 \cdot 0.67}{0.25} = 0.54 \\
\beta_{2,3} &= \frac{\hat{y}_{0,0.5,0.5} - 0.5 \cdot \beta_2 - 0.5 \cdot \beta_3}{0.25} = \frac{-0.44 - 0.5 \cdot (-1.33) - 0.5 \cdot 0.67}{0.25} = -1.33
\end{aligned} \tag{6.7}$$

where $\hat{y}_{0.5,0.5,0}$, $\hat{y}_{0.5,0,0.5}$ and $\hat{y}_{0,0.5,0.5}$ denote the estimated average y value for the binary blends with compositions $x_1 = x_2 = 0.5$, $x_1 = x_3 = 0.5$ and $x_2 = x_3 = 0.5$, respectively (i.e. the mid-points of the three trajectories in Figure 6.5.a, 6.5.b and 6.5.c).

[viii] Notice that when extreme observations are left out of the original data, responses for mixtures that are outside of the calibration experimental domain are predicted (extrapolation). However, as this is the case for all the approaches under study, a fair comparison of the RMSECV values is still guaranteed. Furthermore, for K-PLS, the objects/samples to be iteratively left out are removed from the datasets before the kernel transformation

6.2.3.2 Tablet data

Second-order Scheffé, Cox and K-PLS models were fitted for the analysis of the tablet dataset[ix]. The number of extracted PLS and K-PLS latent variables was selected by leave-one-out cross-validation. As Tables 6.15 and 6.16 point out, the three modelling strategies returned comparable performance indices and regression coefficients, respectively. Figure 6.6 displays their corresponding response surface plot, which are almost identical to one another (including the RBF K-PLS, not shown). They also enable a similar interpretation of the effects of the single constituents on the active ingredient release time, and allow assessing that high contents of phosphate, moderate contents of cellulose and low contents of lactose clearly led to high values of such property of interest. More concretely, binary mixtures composed by roughly $\frac{2}{3}$ of phosphate and $\frac{1}{3}$ of cellulose are expected to exhibit the longest release time. Short release times are instead yielded by blends consisting of e.g. $\frac{2}{3}$ of lactose and $\frac{1}{3}$ of cellulose. Thus, it is quite reasonable to assume the presence of a positive contribution for the interaction phosphate/cellulose and a negative contribution for the interaction lactose/cellulose. As illustrated in Section 6.2.3.1, one can look at the pseudo-sample trajectories spanning the sides of the triangle in Figure 6.6.c for an accurate determination of the Scheffé model first-order and interaction parameters.

Table 6.15. Tablet data: R^2, Q^2 and Root Mean Square Error in Cross-Validation (RMSECV) values resulting from second-order Scheffé model fitting by means of OLS, second- order Cox model fitting by means of PLS, and second-order K-PLS

	# LV	R^2	Q^2	RMSECV
Scheffé second-order model (OLS)	-	0.98	0.84	38.86
Cox second-order model (PLS)	5	0.99	0.83	39.69
K-PLS second-order model	5	0.98	0.83	38.86

6.2.3.3 Bubble data

As with the previous example, the second-order Scheffé, Cox and K-PLS models adjusted for the bubbles dataset rendered very close R^2, Q^2 and RMSECV values (see Tables 6.17 and 6.18). Since this particular mixture problem embraces up to four constituents, the proportion of one of them has to be fixed to allow the response surfaces to be graphed as in Section 6.2.3.2. Given that glycerol presented a much more positive effect on the bubble lifetime and a much higher cost than any other ingredient, as also suggested in [61], its relative amount was set at 0.4. The results (virtually indistin-

[ix] The use of second-order models was originally suggested in [61]

guishable) are represented in Figure 6.7. Figure 6.8 shows instead the corresponding trace plots. As one can easily see, although the effect of DWL2 on the response of interest seems to be more positive than that of DWL1 and water, the interaction of these latter is crucial to guarantee high bubble lifetimes (i.e. more equilibrated blends of DWL1, DWL1 and water would feature more durable bubbles).

The pseudo-sample trajectories spanning the sides of the simplex in Figure 6.7.c cannot be directly resorted to for the estimation of the related Scheffé model coefficients in this situation owing to the fact that the experimental space of the bubbles data is just a portion of a whole simplex, and then they do not reflect the evolution of the predicted response while moving from a pure mixture to another. On the other hand, if these trajectories are constructed so that they exactly overlap the entire edges of this hypothetical tetrahedron, the methodology proposed in Section 6.2.3.1 for the retrieval of the first-order and binary interaction parameters is still valid (assuming that any effect involving the two constituents of the concerned binary mixture do not vary outside the actual data space) [57].

Table 6.16. Tablet data: Scheffé model coefficients estimated by Scheffé polynomial fitting by means of OLS, Cox polynomial fitting by means of PLS and K-PLS

	Scheffé model fitting by OLS	Cox model fitting by PLS	K-PLS
β_1	198.16	198.10	198.16
β_2	114.06	111.94	114.06
β_3	328.97	326.21	328.97
$\beta_{1,2}$	-403.26	-404.98	-403.26
$\beta_{1,3}$	350.56	347.54	350.56
$\beta_{2,3}$	330.37	323.14	330.37

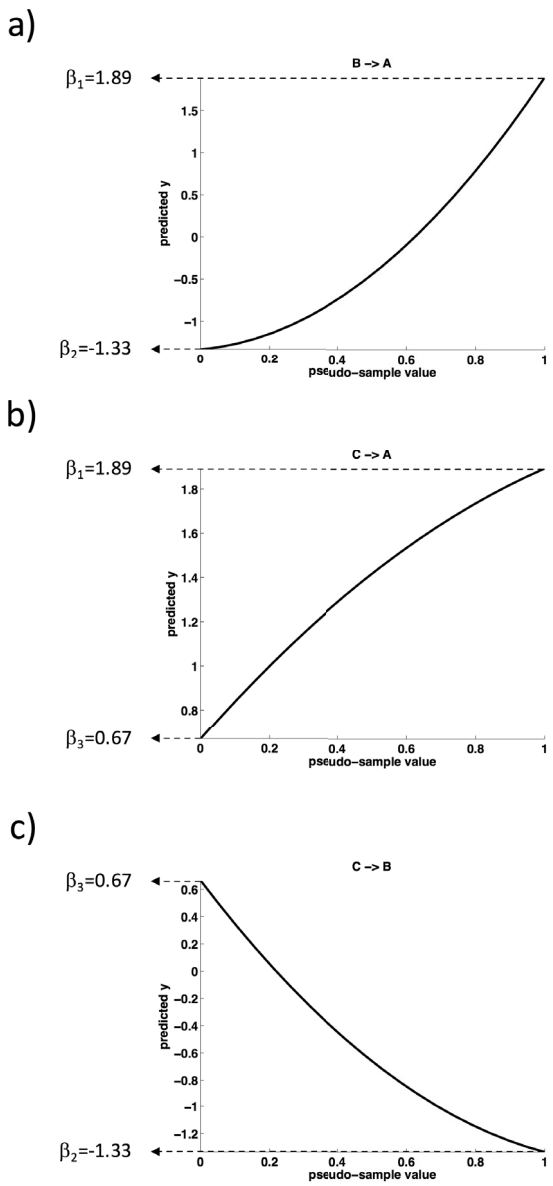


Figure 6.5. Data simulated according to a second-order polynomial model (generation scheme in Equations 6.5-6.6: pseudo-sample trajectories representing the evolution of the predicted response while moving from a) the pure mixture composed solely by constituent B to the pure mixture composed solely by constituent A; b) the pure mixture composed solely by constituent C to the pure mixture composed solely by constituent A; and c) the pure mixture composed solely by constituent C to the pure mixture composed solely by constituent B

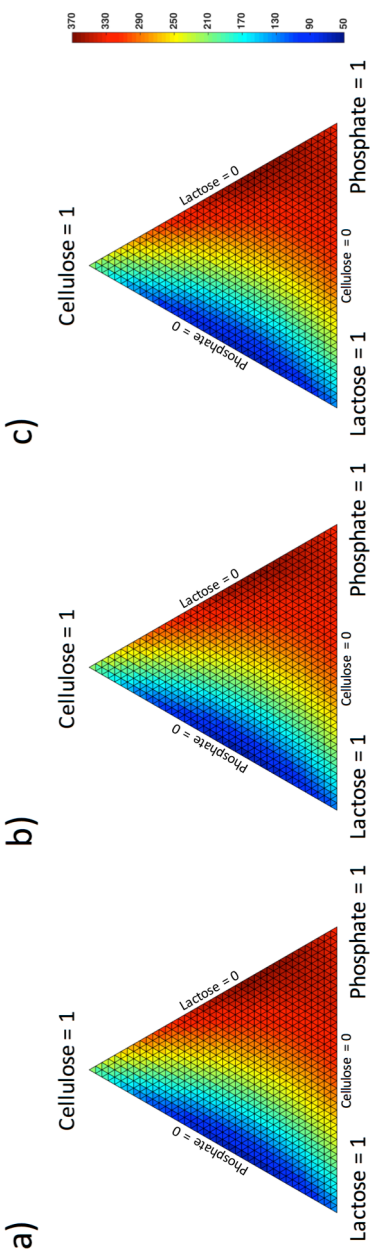


Figure 6.6. Tablet data: response surface plots resulting from a) second-order Scheffé model fitting by means of OLS; b) second-order Cox model fitting by means of PLS; and c) the combination of second-order polynomial K-PLS and pseudo-sample trajectories

6.2.3.4 Colorant data

When the colorant dataset was dealt with, second-, third- and fourth-order Scheffé, Cox and polynomial K-PLS models and RBF K-PLS models were fitted (separately for every response variable) in order to additionally assess the effect of their complexity on the final outcomes. Table 6.19 lists their main performance indices. It can be said that different approaches usually required a different complexity to achieve the minimum RMSECV, but, overall, their performance was found to be rather similar also in this case. For the sake of interpretation, as an illustration, the trace plots resulting from the best Scheffé, Cox and polynomial K-PLS models built for the prediction of the yellow-blue tone (b^*) are displayed in Figure 6.9. They are almost in perfect agreement, and only negligible variations with respect to the outcomes obtained by Alman and Pfeifer in [80] were observed (the same goes for those derived for both L^* and a^* and for all those returned by RBF K-PLS, which are not shown). Concretely, all the constituents exhibited a positive effect on b^*.

Table 6.17. Bubbles data: R^2, Q^2 and Root Mean Square Error in Cross-Validation (RMSECV) values resulting from second-order Scheffé model fitting by means of OLS, second- order Cox model fitting by means of PLS, and second-order K-PLS

	# LV	R^2	Q^2	RMSECV
Scheffé second-order model (OLS)	-	0.94	0.81	0.042
Cox second-order model (PLS)	9	0.94	0.81	0.042
K-PLS second-order model	9	0.94	0.81	0.042

6.2.3.5 Gasoline data

Second-order cross-validated Scheffé and Cox models were tentatively adjusted for the gasoline dataset. Due to the low number of samples, however, not enough degrees of freedom were available for the estimation of the coefficients of more complex polynomials. As highlighted in Table 6.20, both approaches returned negative Q^2 and poor RMSECV values. On the other hand, their performance was clearly outmatched by that of RBF K-PLS. In addition, RBF K-PLS was found to yield figures of merit just slightly worse than those obtained by coding the constituent proportions in terms of so-called pseudo-components [81] and fitting a first-order Scheffé or Cox model including inverse terms, a more standard and common procedure for handling non-linear data coming from a mixture design of experiments [57]. Still, the main advantage of K-PLS over it is that there is no need of performing such a domain transformation and defining these inverse terms prior to the analysis: the optimisation of the kernel transformation function and, in this case, of the σ parameter guarantees a certain flexibility when modelling different types of non-linearities without requiring any further operation to be carried out. It is important to note that this domain transformation would cause all

pseudo-components proportions in the blend to potentially vary from zero to one. Therefore, a small number must be artificially added up to each pseudo-component's proportion in the denominator of the inverse terms in order to avoid infinite values at the edges (i.e. when at least one of the pseudo-component's proportion in the blend is zero). However, the magnitude of this number will affect the estimation of the model coefficients to a greater or lesser extent depending on the severity of the non-linearities of the real effects, thus increasing the relative effort required to properly fit a regression model via this approach when compared to K-PLS.

The first-order pseudo-component Scheffé and Cox models encompassing inverse terms led to identical surface plots (see Figure 6.10.a and 6.10.b). Certain dissimilarities from the one rendered by RBF K-PLS (see Figure 6.10.c) are instead observable, which was expected considering the intrinsic differences among the compared algorithmic methodologies and especially the fact that the first-order pseudo-component Scheffé and Cox models encompassing inverse terms are able to explain strong nonlinearities mainly at the borders of the design space but not in its central area. Nevertheless, a common explanation of the effect of the distinct ingredients on the response variable can be given: the maximum octane rating can be achieved by blending a high quantity of catalytically cracked and low quantities of C5-isomer and reformate.

Table 6.18. Bubbles data: Scheffé model coefficients estimated by Scheffé polynomial fitting by means of OLS, Cox polynomial fitting by means of PLS and K-PLS

	Scheffé model fitting by OLS	Cox model fitting by PLS	K-PLS
β_1	-1.49	-1.49	-1.49
β_2	2.35	2.35	2.35
β_3	-1.35	-1.35	-1.35
β_4	2.11	2.11	2.11
$\beta_{1,2}$	2.08	2.08	2.08
$\beta_{1,3}$	14.55	14.55	14.55
$\beta_{1,4}$	7.51	7.51	7.51
$\beta_{2,3}$	6.70	6.70	6.70
$\beta_{2,4}$	2.63	2.63	2.63
$\beta_{3,4}$	7.82	7.82	7.82

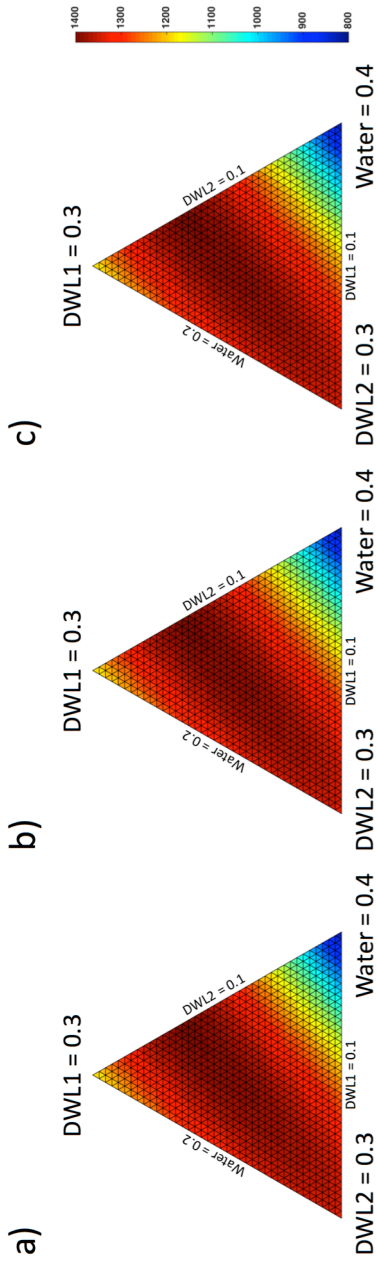


Figure 6.7. Bubbles data: response surface plots resulting from a) second-order Scheffé model fitting by means of OLS; b) second-order Cox model fitting by means of PLS; and c) the combination of second-order polynomial K-PLS and pseudo-sample trajectories

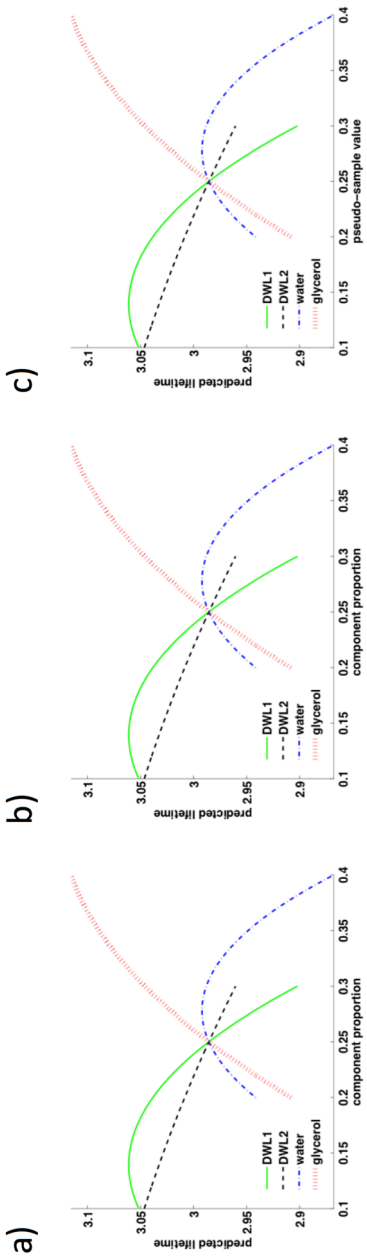


Figure 6.8. Bubbles data: trace plots representing the evolution of the predicted lifetime while varying the proportion of the 4 ingredients of the blend (DWL1, DWL2, water and glycerol) and resulting from a) second-order Scheffé model fitting by means of OLS, b) second-order Cox model fitting by means of PLS, and c) the combination of second-order K-PLS and pseudo-sample trajectories. Here, $s_1 = s_2 = s_3 = s_4 = 1/4$

Table 6.19. Colorant data: R^2, Q^2 and RMSECV values for the three properties of interest (L^*, a^* and b^*) resulting from second/third/fourth-order Scheffé polynomial fitting by means of OLS, second/third/fourth-order Cox polynomial fitting by means of PLS, second/third/fourth-order polynomial K-PLS, and RBF K-PLS. For each response variable the best models in terms of RMSECV are highlighted in bold

	L^*				a^*				b^*			
	# LV	R^2	Q^2	RMSECV	# LV	R^2	Q^2	RMSECV	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.97	0.92	6.29	-	0.96	0.93	3.74	-	0.92	0.87	4.27
Second-order Cox model (PLS)	5	0.97	0.92	6.27	5	0.96	0.93	3.71	5	0.92	0.87	4.24
Second-order polynomial K-PLS model	3	0.97	0.95	4.94	6	0.96	0.93	3.72	6	0.92	0.86	4.27
Third-order Scheffé model (OLS)	-	0.99	0.99	1.35	-	0.98	0.85	5.62	-	0.95	0.50	8.39
Third-order Cox model (PLS)	12	0.99	0.99	1.33	8	0.98	0.88	4.93	3	0.84	0.74	6.12
Third-order polynomial K-PLS model	14	0.99	0.99	1.25	9	0.98	0.94	3.23	10	0.95	0.86	4.41
Fourth-order Scheffé model (OLS)	-	0.99	0.90	7.15	-	0.99	0.76	7.01	-	0.99	0.21	10.56
Fourth-order Cox model (PLS)	12	0.99	0.96	4.61	12	0.99	0.95	3.35	13	0.99	0.92	3.45
Fourth-order polynomial K-PLS model	6	0.98	0.97	3.54	9	0.98	0.93	3.77	10	0.95	0.82	5.01
RBF K-PLS model	14 $\sigma_{L^*}=25$	0.99	0.99	1.30	10 $\sigma_{a^*}=25$	0.98	0.94	3.38	11 $\sigma_{b^*}=25$	0.95	0.89	3.97

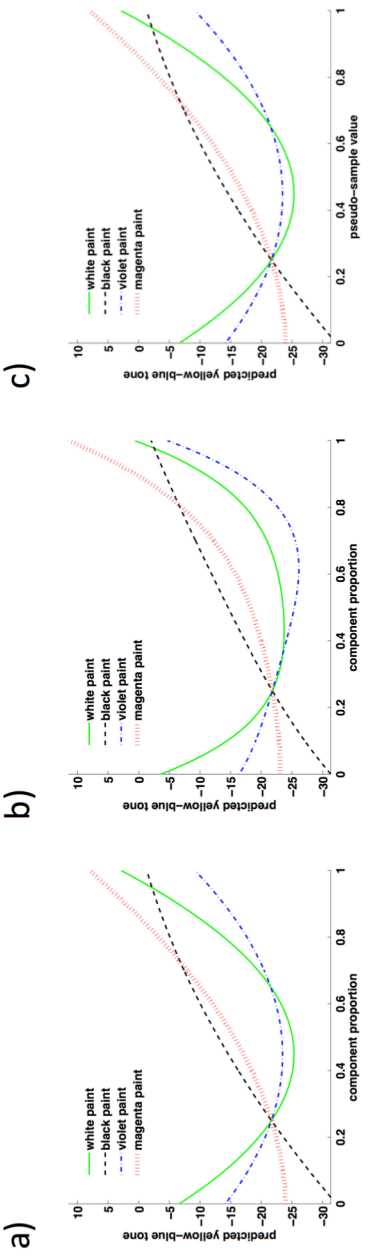


Figure 6.9. Colorant data: trace plots representing the evolution of the predicted yellow-blue tone (b^*) while varying the proportion of the 4 ingredients of the blend (white, black, violet and magenta paints) and resulting from a) second-order Scheffé model fitting by means of OLS, b) fourth-order Cox model fitting by means of PLS, and c) the combination of second-order K-PLS and pseudo-sample trajectories. Here, $s_1 = s_2 = s_3 = s_4 = 1/4$

Table 6.20. Gasoline data: R^2, Q^2 and RMSECV values resulting from second-order Scheffé model fitting by means of OLS, second-order Cox model fitting by means of PLS, first-order (plus inverse terms) pseudo-component Scheffé model fitting by means of OLS, first-order (plus inverse terms) pseudo-component Cox model fitting by means of PLS, and RBF K-PLS. If required and feasible (i.e. a sufficient number of degrees of freedom was available), non-linearity degree (tuned through the value of the σ parameter in RBF K-PLS) and complexity (number of latent variables) were optimised within a leave-one-out cross-validation loop.

	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.85	-1.42	12.40
Second-order Cox model (PLS)	2	0.99	-0.91	11.02
First-order (plus inverse terms) pseudo-component Scheffé model (OLS)	-	0.98	0.88	2.73
First-order (plus inverse terms) pseudo-component Cox model (PLS)	5	0.99	0.82	2.73
RBF K-PLS second-order model ($\sigma = 2$)	8	0.99	0.82	3.46

6.2.3.6 Data simulated according to a highly non-linear model

This section will be focused on further emphasizing the added value of K-PLS with respect to the other methodologies concerned. The outcomes yielded by the application of Scheffé model fitting by means of OLS, Cox model fitting by means of PLS, and K-PLS to the second simulated dataset are reported in Table 6.21.

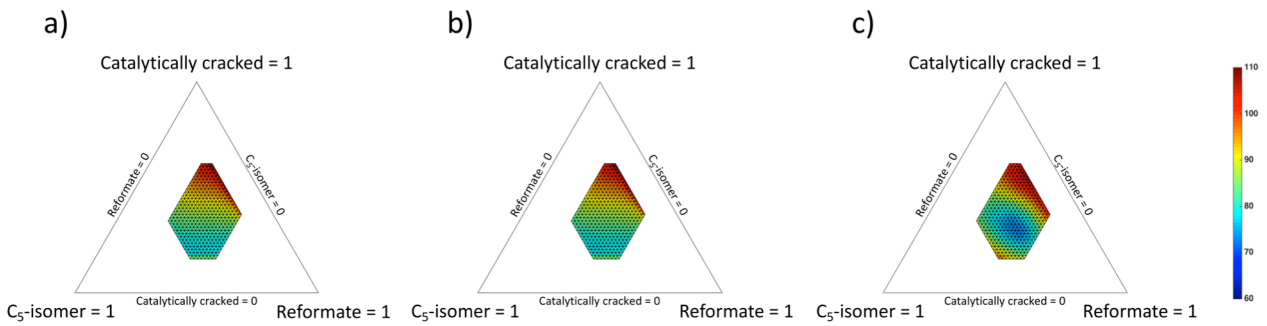


Figure 6.10. Gasoline data: response surface plots resulting from a) first-order (plus inverse terms) pseudo-component Scheffé model fitting by means of OLS, b) first-order (plus inverse terms) pseudo-component Cox model fitting by means of PLS, and c) the combination of RBF K-PLS and pseudo-sample trajectories. As here the original mixture design space is irregular, the graphs are represented as parts of the corresponding simplex with vertices [1,0,0], [0,1,0] and [0, 0, 1] for the sake of an easier visualisation

Table 6.21. Data simulated according to a highly non-linear model (generation scheme in Equation 6.7): R^2, Q^2 and RMSECV values resulting from second-order Scheffé model fitting by means of OLS, second-order Cox model fitting by means of PLS, first-order (plus inverse terms) pseudo-component Scheffé model fitting by means of OLS, first-order (plus inverse terms) pseudo-component Cox model fitting by means of PLS, and RBF K-PLS. If required and feasible (i.e. a sufficient number of degrees of freedom was available), non-linearity degree (tuned through the value of the σ parameter in RBF K-PLS) and complexity (number of latent variables) were optimised within a leave-one-out cross-validation loop

	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.95	0.56	2.14
Second-order Cox model (PLS)	5	0.99	0.56	2.14
First-order (plus inverse terms) pseudo-component Scheffé model (OLS)	-	0.69	-1.11	4.71
First-order (plus inverse terms) pseudo-component Cox model (PLS)	1	0.96	-0.06	3.35
RBF K-PLS second-order model ($\sigma = 0.2$)	7	0.98	0.61	1.83

The R^2, Q^2 and RMSECV displayed values corroborate what was stated before about K-PLS: when strong non-linear relationships (e.g. fourth order, logarithmic, etc.) characterise the data under study, it may outperform in terms of fit and prediction quality both Scheffé model fitting by means of OLS and Cox model fitting by means of PLS. This applies even if first-order Scheffé or Cox models including inverse terms are fitted. Notice that here a RBF kernel transformation and not a polynomial one was found to guarantee the best Q^2 and RMSECV. As already mentioned, RBF K-PLS requires the optimisation of an additional parameter, σ. The variation of such a parameter allows different types of complex trends to be modelled, thus its utilisation might be highly recommended when combinations of unknown non-linearities influence the interdependence between ingredient proportions and properties of interest.

6.2.4. Conclusions

In this section a novel approach for the analysis of data proceeding from mixture designs of experiments and based on the combination of KPLS and pseudo-sample trajectories was proposed. Two interesting points arose from the discussed examples:

- if the considered mixture data were not affected by severe nonlinearities and/or featured a sufficiently high number of observations, K-PLS and pseudo-sample trajectories yielded very similar results to classical Scheffé model fitting by means of OLS and Cox model fitting by means of PLS (see Sections 6.2.3.2, 6.2.3.3 and 6.2.3.4). A way of recovering the parameters of a Scheffé

model (provided that it holds and has the same complexity as the K-PLS one) from the trend of the pseudo-sample trajectories was also derived and validated via a simulated case-study (see Section 6.2.3.1);

- on the contrary, when more non-linear and relatively small data structures had to be analysed, K-PLS proved to be a valid alternative for overcoming the main limitation of both Scheffé model fitting by means of OLS and Cox model fitting by means of PLS (see Sections 6.2.3.5 and 6.2.3.6): it resulted, in fact, in better fit and prediction quality when the nature of such non-linear data was not strictly polynomial. In addition, although the performance of these more classical methodologies can be improved by taking into account inverse terms, often not enough degrees of freedom are available for a stable estimation of the coefficients of these augmented models. K-PLS does not suffer from the same drawback. On top of that, RBF K-PLS through the optimisation of its parameter, σ, may allow different types of complex nonlinear relationships to be modelled. Its use might then be highly recommended when combinations of unknown non-linearities influence the nature of the interdependence between constituent proportions and response variables.

Finally, it was also shown how graphs like the surface plots or the trace plots associated to the mixture design space can be retrieved by the pseudo-sample trajectories enabling a reliable interpretation of the effect of changing the proportion of the different ingredients of the blend on its properties of interest.

Chapter 7
MiDAs: a software for mixture DOE and data analysis

7.1. Introduction

The name of this tool stands for Mixture Design and Analysis. MiDAs is a Matlab-powered software with graphic interface that allows to easily approach mixture design problems for the construction of model-based and space filling Designs of Experiments (DOE), as well as to analyse mixture design data using traditional approaches such as OLS (simply referred to as Multiple Linear Regression, or MLR, inside the software itself) and by Partial Least Squares. Although performing these studies requires some knowledge in statistics in general, and mixture design in particular, the aim of this software is to provide its users with an easy way to apply the tools it offers without requiring extensive knowledge of Matlab or other mathematical programming software, and without requiring having Matlab itself to do so.

MiDAs is structured in two main modules, the first of which is dedicated to the construction of mixture DOEs (including DOEs for mixture problems which also include process or amount variables), and the second dedicated to mixture data analysis. The main menu of the software, shown in Figure 7.1, clearly differentiates both modules.

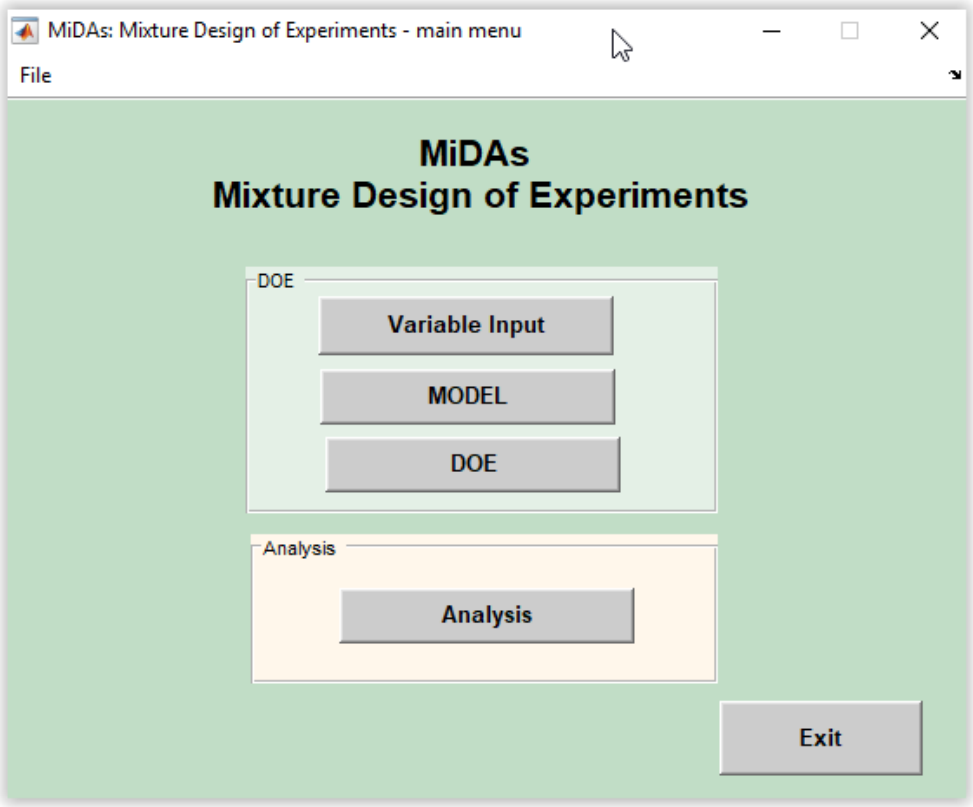


Figure 7.1. MiDAs main menu

The main menu provides 3 options for creating or modifying designs, as well as one option for the analysis of mixture data:

- ***Variable Input:*** this option permits introducing the mixture constituents as well as process variables (no differentiation is made here between 'process variables' and 'amount variables'), and univariate and multivariate restrictions on them, which will determine the shape and size of the experimental space.
- ***MODEL*****:** this option allows selecting a model given the previously defined variables, and including or excluding terms of the model once selected. Although a space filling design may be desired, choosing a model is still needed, even if it will not affect the final DOE.
- ***DOE*****:** given the defined variables, constraints on them and the model in previous options, this option will suggest a Design of Experiments, although the user may want to (and can) specify a different one selecting some optimality cri-

terion, the number of experiments that will constitute the desired DOE, and fixing some other parameters depending on the kind of DOE.

- ***Analysis*:** this option allows fitting different mixture models with MLR and PLS if the appropriate data is provided.

The main menu also presents the possibility of saving the current project into a MAT file or to load data from a MAT file from a previous project.

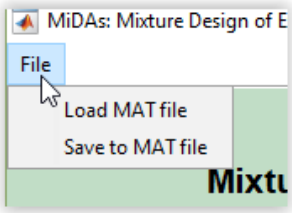


Figure 7.2. Main menu - File menu

The following subsections detail the tools provided by each of these four options.

7.2. Variable input

As seen in Figure 7.3, The window "Variable Input" has three tabs: "Variables", "Univariate Restrictions" and "Relational Constraints", which must be resorted to sequentially. The selected tab will be highlighted in bright green colour.

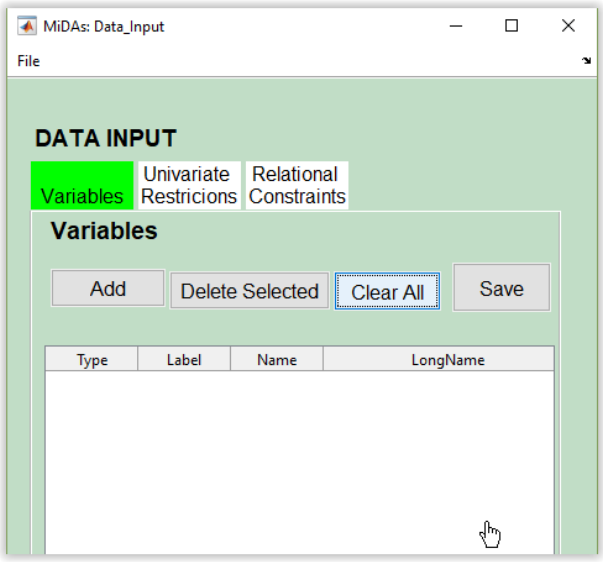


Figure 7.3. Data Input window - Variables tab

To introduce new variables or modify the information of existing ones, one must click on the "Variables" tab, then "Add". This will generate a new register with default values, as seen in Figure 7.4.

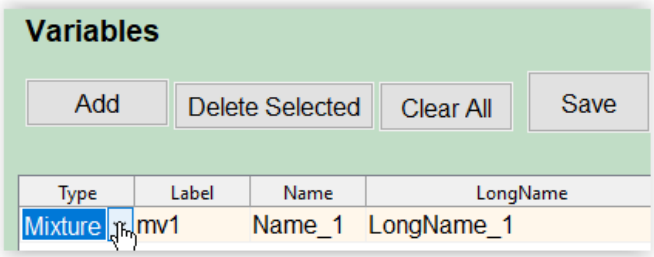


Figure 7.4. Data Input window - Variables tab. Add variable

The values for this register can be modified, and as many new registers as required can also be added before or after modifying the existing ones. The 'properties' of each register are:

- ***Type:*** "Mixture" or "Process". Select one from these in the dropdown menu (see Figure 7.5).
- ***Label:*** short identifier for the variable, to be used later on in other windows. If changed, the new label must be as short as possible and not include blanks nor special characters. Clicking on the cell permits changing it.
- ***Name:*** short name, to be used later on in other windows. It should be short but enough to identify the component or process, and it may be used in some reports of interest for the user. Clicking on the cell permits changing it.
- ***LongName:*** a more elaborate description of the corresponding variable, in case the user requires adding some clarification. It will not be used on any report nor window. Clicking on the cell permits changing it.

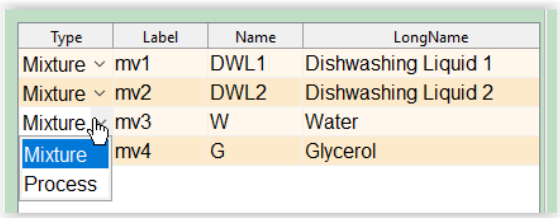


Figure 7.5. Data Input window - Variables tab table. Select variable type

The button “Save” allows saving all changes. Trying to move to a different tab will make a window pop up if changes have not been saved, as this is necessary to continue. This does *not*, however, save the data to a project file. To do so the option “File” must be used

The button “Clear All” clear all inputs the information introduced up until now. It will *not* overwrite any existing data in the project file.

The button “Delete Selected” deletes from the list the last clicked on register, which will have some cell highlighted in blue to indicate it has been selected. It will *not* overwrite any existing data in the project file.

To introduce new univariate restrictions, the “Univariate Restrictions” tab must be clicked if it is not active already. This tab, which can be seen in Figure 7.6, can be used to modify lower and upper bounds imposed on all variables as required.

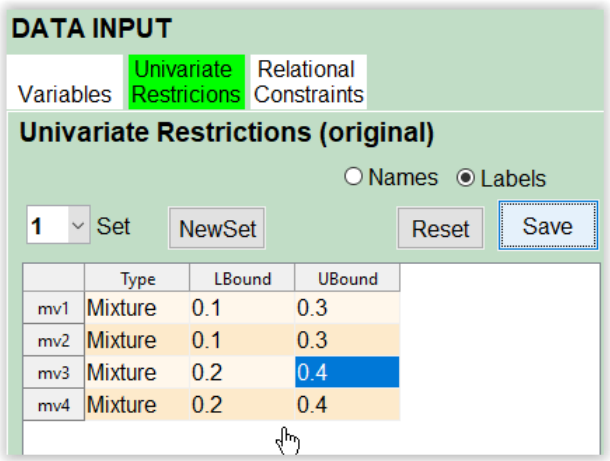


Figure 7.6. Data input window - Univariate Restrictions tab (original).

Row labels can show the short name or the label for the variables. Change this at any time by clicking on radio buttons “Names” or “Labels”. Default values are set to 0 for lower limit and 1 for upper limit.

More than one set of bounds can be defined (see Figure 7.7). Clicking on “NewSet” allows creating a new set of bounds, while the dropdown menu “Set”, with the numbers associated to the different sets, permits selecting which list of bounds (from the ones already defined) will be used later on.

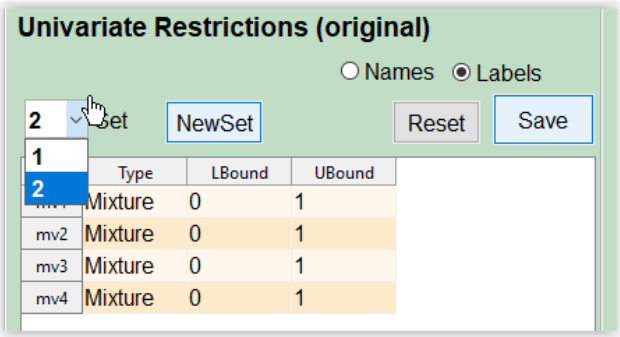


Figure 7.7. Univariate Restrictions. Change set

The button "Reset" sets the lower and upper bounds values to default.

The button "Check Restrictions" (see Figure 7.8) computes the consistent bounds.

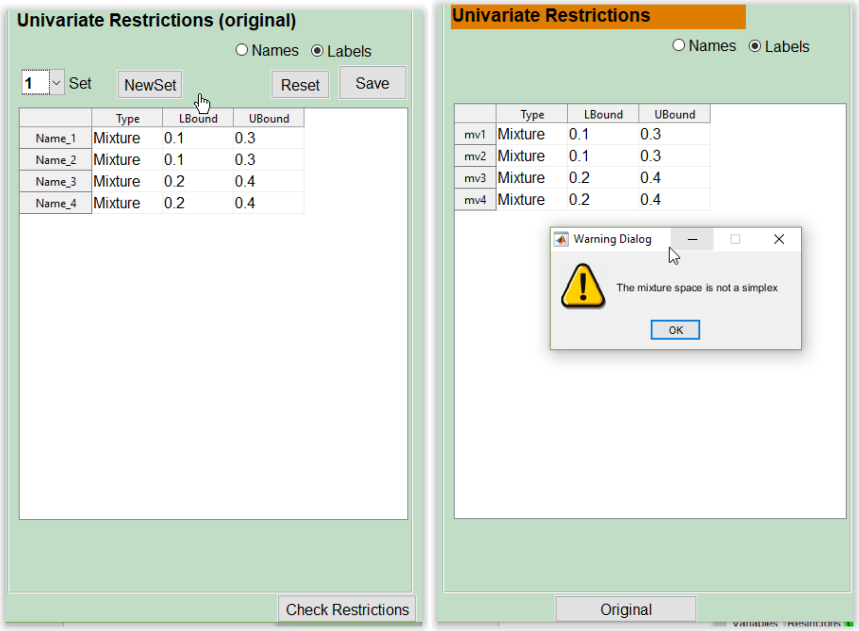


Figure 7.8. Univariate Restrictions. Check Restrictions

MiDAs notifies if any errors in the bounds make it impossible to continue with the creation of the design, for example when the ways the restrictions are defined generate an empty mixture space. It will also inform the user whether the mixture space is an L-simplex, a U-simplex, not a simplex, if it is constituted by a single mixture of if it is void, via a pop up window. It must be noted that this button evaluates the shape of the mixture space by also accounting for any active "Relational Constraints".

Clicking on the button "Original" shows again the original restrictions.

In addition to univariate restrictions, defining multivariate ones (here called 'relational constraints') is possible. To add or modify already existing relational constraints click on "Relational Constraints" tab. By default, the values of the coefficients that accompany each variable in the constraint are initially set to zero, the constraint is assumed to be an equality one ('='), and the 'objective' value (i.e. the one the linear combination of variables must be equal to or lower/higher than) is also set to zero. Additionally, the new relational constraint is set to 'inactive' (see Figure 7.9).

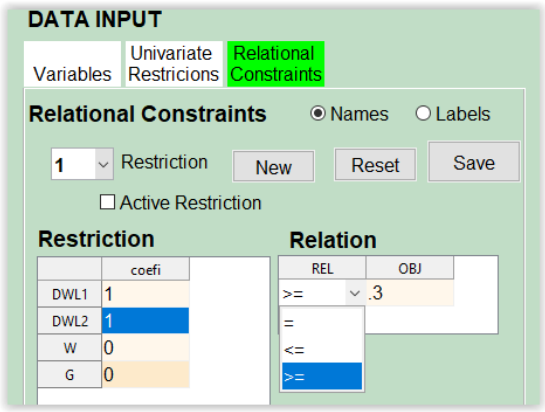


Figure 7.9. Relational Constraints

By clicking on a cell, the value within it can be changed. To change the type of constraint, the dropdown menu below 'REL' must be clicked, and one of the options selected ('=', '<=', '>='). Clicking on 'Save' will save the relational constraint. It is also possible to add more restrictions by clicking on "New" (see Figure 7.10).

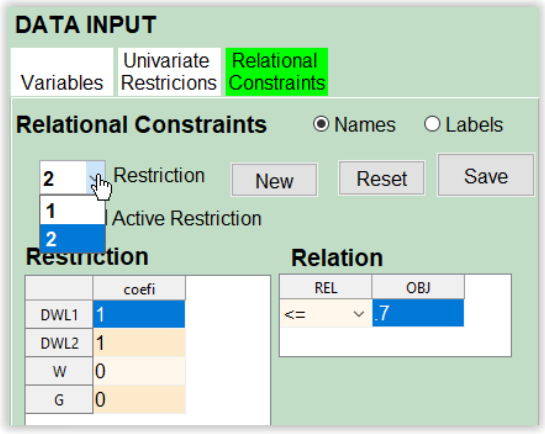


Figure 7.10. Relational Constraints

Multiple restrictions can be defined this way, but the model will consider only those flagged as "Active Restriction". If after defining the restrictions (univariate or relational) new variables are added or existing ones are deleted, or a new data file is loaded, all the defined restrictions will disappear and will have to be re-entered again if necessary. The perfect collinearity restrictions in Equation 5.1 among mixture variables is defined internally and therefore it is not necessarily for the user to include it among the relational constraints they may want to impose.

7.3. Model selection

Back at the main menu, and once information on variables (with at least 3 mixture variables) and restrictions has been introduced, the option "Model selection" can be accessed. MiDAs will propose the most simple model that can be defined while including all variables introduced in the "Data Input" menu by default, but a different model can be chosen nonetheless (see Figure 7.11).

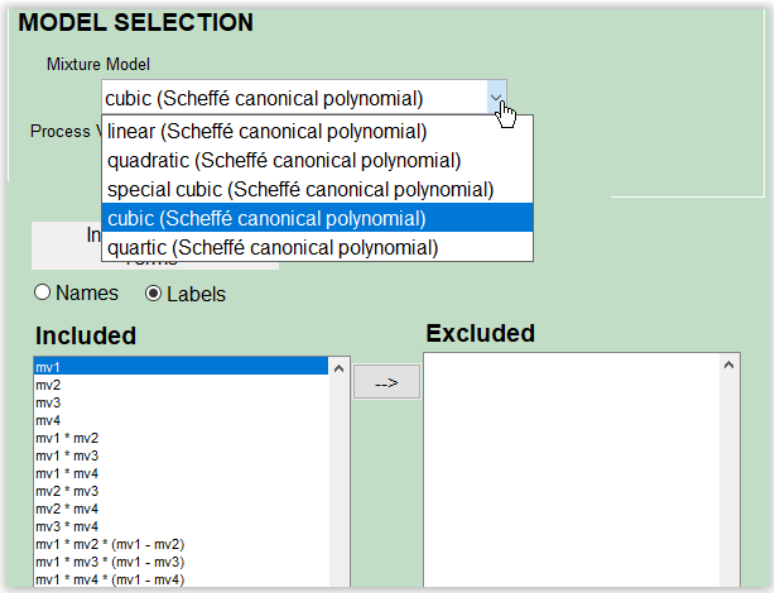


Figure 7.11. Model Selection.

In case there are process variables a list of proposed models for them is also shown in the "Process Variable Model" popup (see Figure 7.12). MiDAs will propose the mixture-process variables model that combines both models in a multiplicative way.

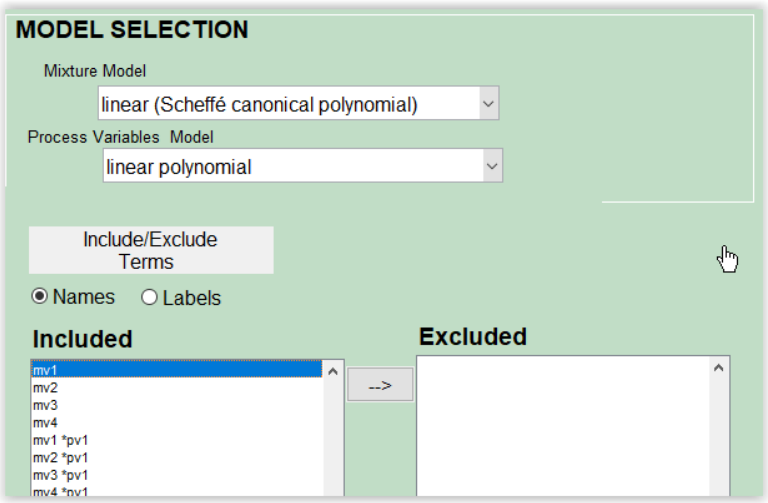


Figure 7.12. Model Selection with process variables

MiDAs will fill the dropdown menu for Model Selection with the corresponding model terms, based on the number and type of variables introduced before. Nonetheless, the user may be interested in excluding some of the terms of the complete model (see Figure 7.13). To do so, the term to be excluded must be clicked on, and clicking on the arrow pointing to the right will exclude such term and all other terms that depend on it according to the principle of hierarchy. The list of included terms is stored in a workspace variable to be used in the DOE.

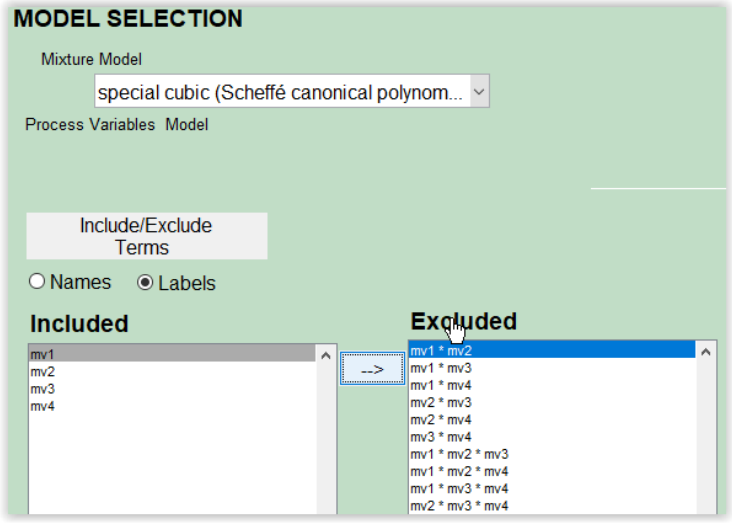


Figure 7.13. Model Selection. Include/exclude terms

7.4. DOE construction

Once the model has been selected, the DOE option in the main menu can be accessed. By clicking on it, MiDAs will make use of the data introduced via the Data Input and Model Selection windows and suggest a DOE, referred to in the software as the "original" one (see Figure 7.14), which can be modified by the user later on. Depending on the mixture space being a simplex or not, the presence of process variables or not, and the model selected, MiDAs may propose one of the simplex-based DOEs presented in Section 5.4.1, or will propose a list of optimality criteria to choose from (D-optimal, I-optimal or distance-based, for a space filling DOE) to construct a DOE via the "Optimize" tab.

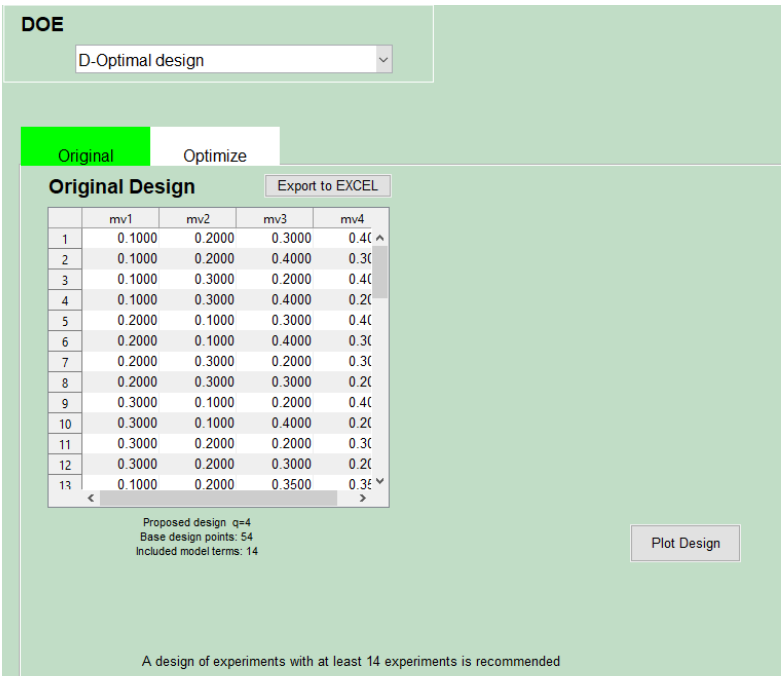


Figure 7.14. Design of Experiments. Original proposed design

In all cases, a table with the points of the proposed design will be made visible, and a count of the number of points will appear next to the number of terms included in the model. If there are less than two or more than four mixture variables, a warning is displayed informing that the graphical representation of the design is not possible. Otherwise, the option "Plot Design" can be clicked on to visualize the DOE in the mixture space (see Figure 7.15).

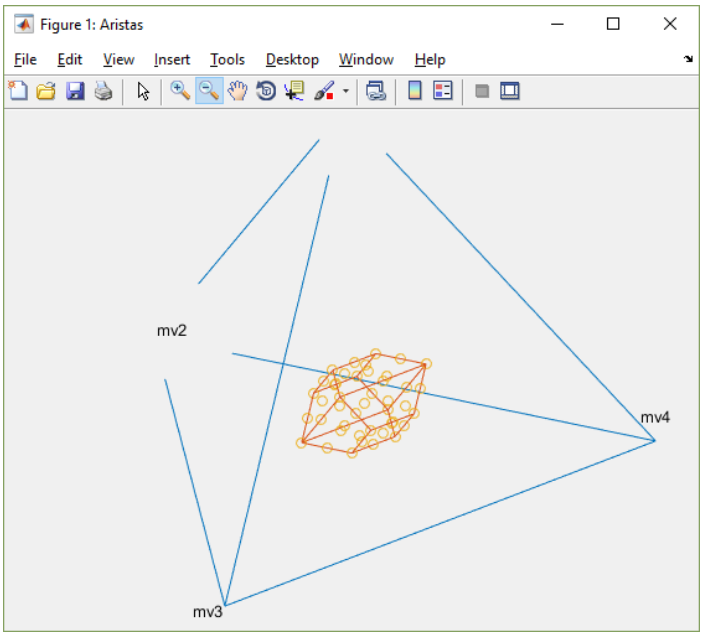


Figure 7.15. Original proposed design plot

By clicking in "Optimize" tab, a new proposed design will be obtained based on the number of points of the desired design, and the number of iterations and max. length of excursion when the point-exchange algorithm is used to construct it. After setting new values for points, iterations and max.length of excursion, clicking on "Optimize" will provide the optimized DOE as well as its D-efficiency and G-efficiency (see Figure 7.16). If the graphical representation is possible, the optimized DOE can be viewed by clicking on "Plot Design" in this same tab (see Figure 7.17).

To save all the work done up until now, the option "Save to MAT file" in the menu bar option "File" must be selected. The DOE can also be exported to an excel file by clicking on the button "Export to EXCEL", and all the data related to this design may be accessed later by loading a saved mat file from any other Matlab program or by importing the data from the EXCEL file.

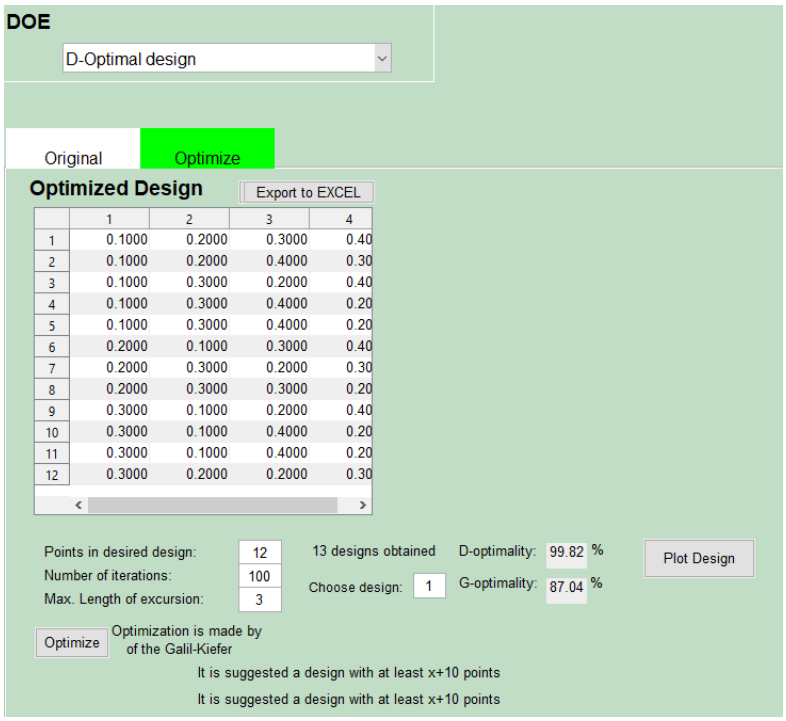


Figure 7.16. Optimized design

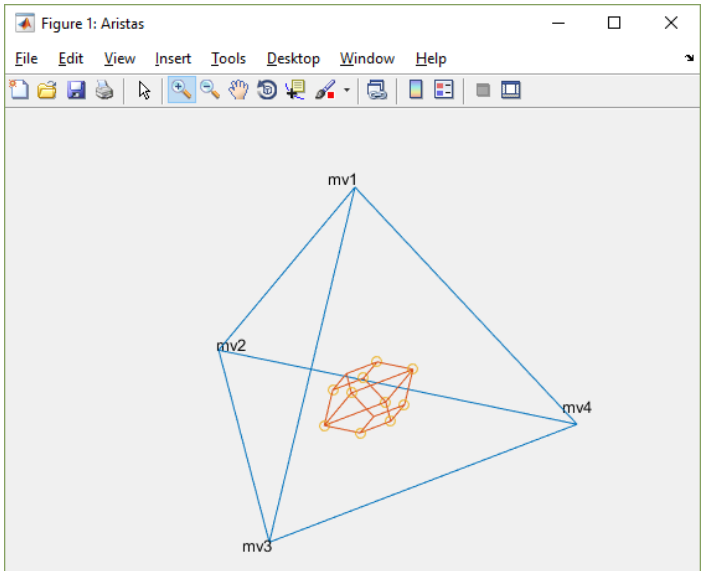


Figure 7.17. Plot optimized design

7.5. Data analysis

The option "Analysis" allows fitting different models using MLR or PLS if the appropriate data is provided. The main menu is shown in Figure 7.18.

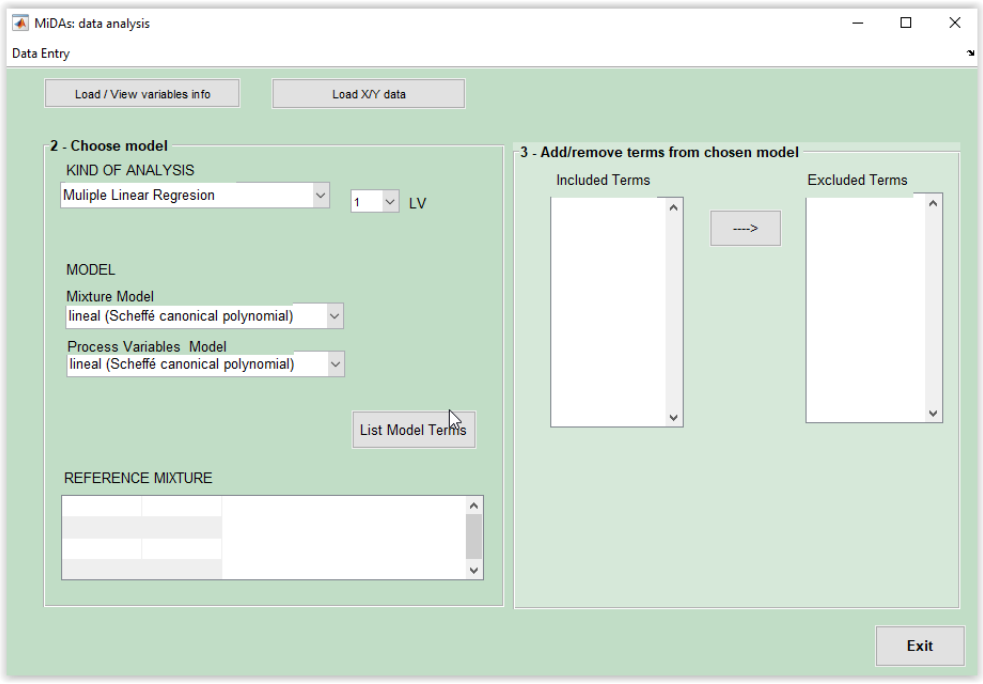


Figure 7.18. Data Analysis

From the main menu experimental data from a previously constructed DOE of from an Excel file can be retrieved. If such data is loaded from an Excel file then the upper and lower bounds and multivariate/relational constraints must be defined. Two options exist the load the data:

- *Load / View variables info*

Pressing on "*Load / View variables info"* opens the screen "Regress Variable Info", which can be seen in Figure 7.19.

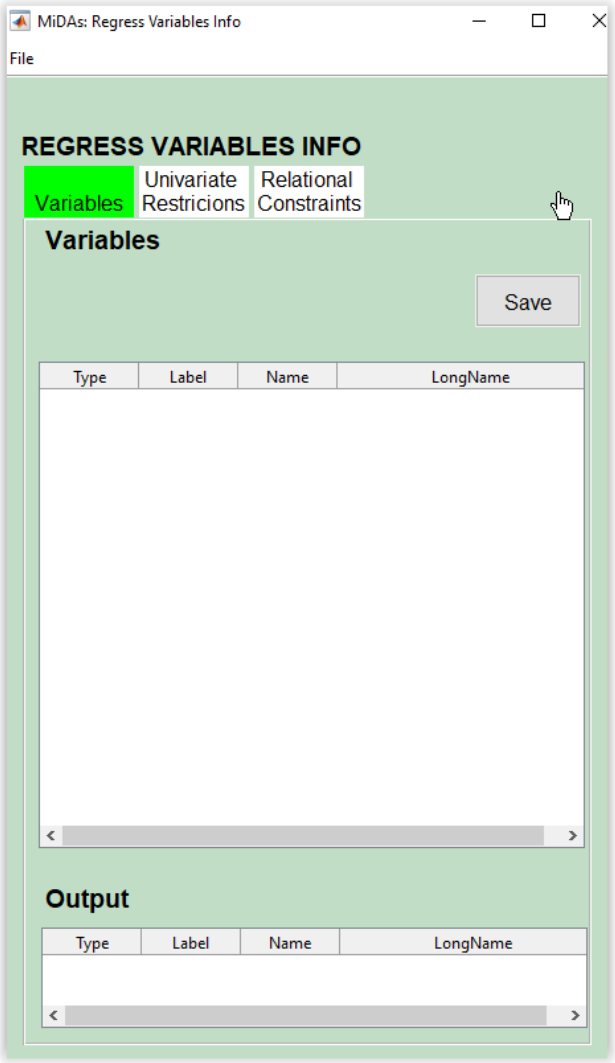


Figure 7.19. Regress Variable Info screen

By clicking "File" on the menu bar and then "Load mat file" (see Figure 7.20), data from previous analysis can be retrieved. "Save as" can be used to save the data of the current analysis in a .mat file. "Load DOE mat file" permits getting the data generated in a previous DOE project in MiDAs.

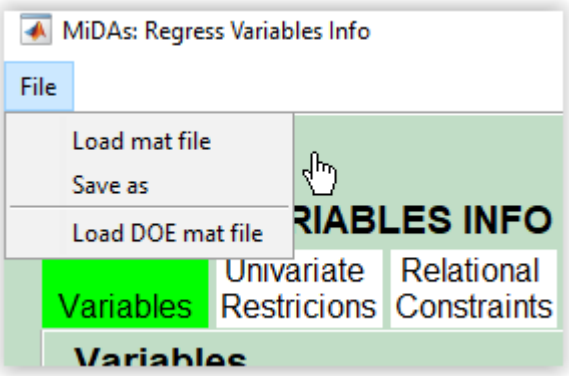


Figure 7.20. "File" options

When loading data from a DOE previously generated using MiDAs it will be necessary to add the output variables and check, modify or add univariate and multivariate restrictions additional to the ones in the corresponding DOE project file.

- *Load X/Y data*

If this option is selected, experimental data will be loaded from an EXCEL file (see Figure 7.21)

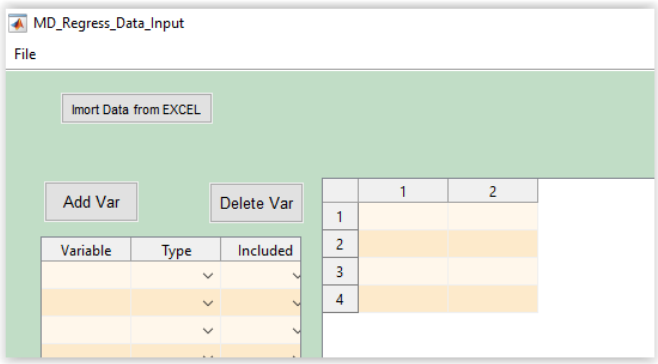


Figure 7.21. Load X/Y data screen

Clicking on "Import Data From EXCEL" will show the menu in Figure 7.22, where it is possible to select an Excel file through the option "Select EXCEL", and then click on the Excel sheet from which the data will be taken by pressing "Load Sheet". If the first line of the Excel sheet contains variable labels, ticking the "With headers" checkbox will allow saving these labels. The loaded experimental data can be modified by clicking on the cells containing the values to edit them.

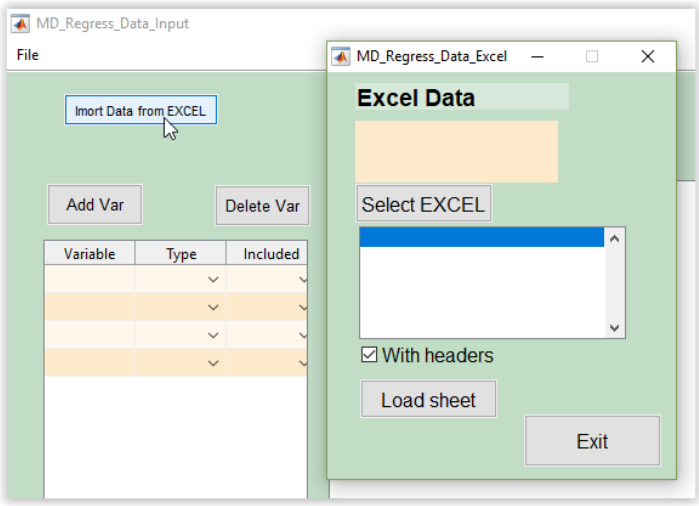


Figure 7.22. Excel file selection

By default, all variables will be set as mixture type (see Figure 7.23), and therefore variables of other types must be changed to "Process" or "Output" accordingly. The variable labels can also be edited, and more variables added. Variables can also be excluded from the following analysis.

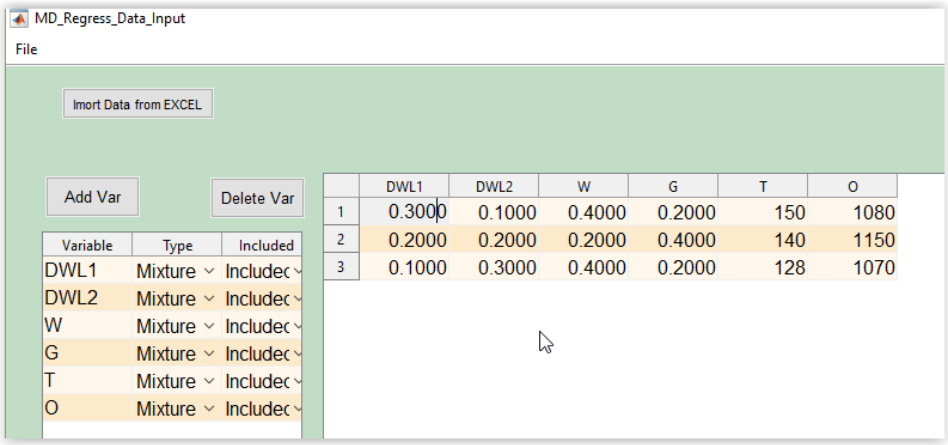


Figure 7.23. Analysis data input screen

To modify labels, clicking on the corresponding cell is enough to be able to change it. To change type, the "Type" popup should be opened for the corresponding variable and the correct one chosen from the list. To exclude a variable, the "Included" popup can be pressed to select "Excluded", and vice versa (see Figure 7.24).

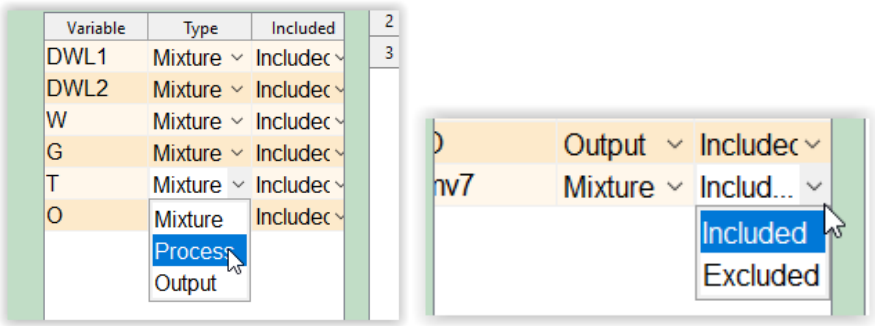


Figure 7.24. Change type and include

Data of variables marked as "Excluded" cannot be edited, and the corresponding columns will be highlighted in red (see Figure 7.25).

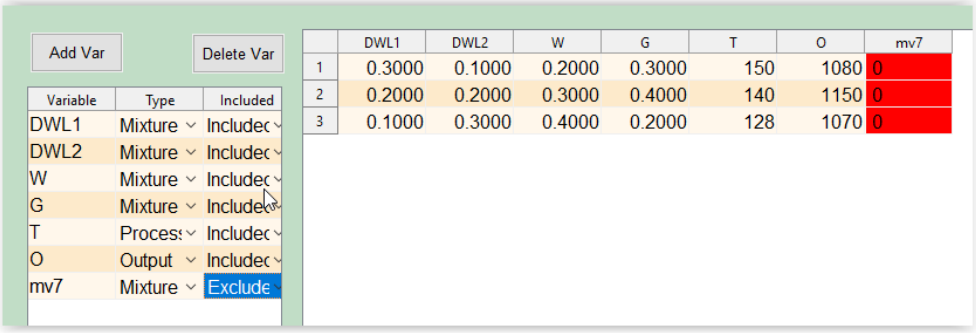


Figure 7.25. Sample input data screen

Once finished, simply closing this screen will save the loaded data and make it available for posterior analysis.

If any changes are required regarding the constraints that define the experimental space, they can be done by clicking on "Load/View variable info" in the main analysis screen. Doing this will open the window shown in Figure 7.26, which now contains the information regarding the variables from the loaded experimental data (either from a DOE MiDAs project or from an Excel file). If the data comes from an Excel file, MiDAs will automatically set lower and upper bounds equal to the minimum and maximum values from the experimental data matrix, respectively.

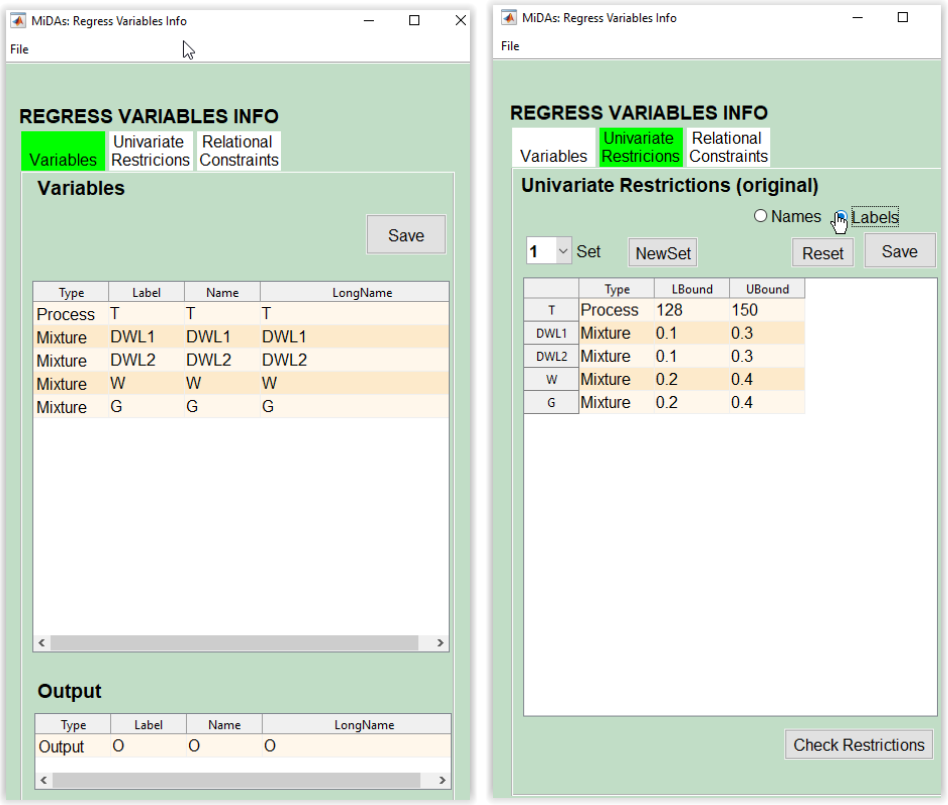


Figure 7.26. Review and change univariate restrictions

It must be noted that the information in this screen regarding the type and label of each variable cannot be modified here. If any of them needs changing, this will have to be done in the screens seen in Figures 7.23 to 7.25.

Multivariate/relational constraints are never considered if the data comes from an Excel file (except for the perfect collinearity restrictions in Equation 5.1 among mixture variables), and must therefore be added, if needed, or modified if required when the data has been retrieved from a pre-existing analysis or DOE project in MiDAs. This is done in the corresponding tab, which has the same aspect as the one in Figure 7.10.

Finally, the button “Check Restrictions” ought to be pressed so that the consistent limits are calculated, the simplex condition is checked, and a reference mixture (coincident by default with the centroid of the mixture space) and list of terms are proposed. By default the simplest model structure that can be considered with all variables involved, to be fit via MLR, is proposed. This will result in the main analysis window looking similar to the one in Figure 7.27.

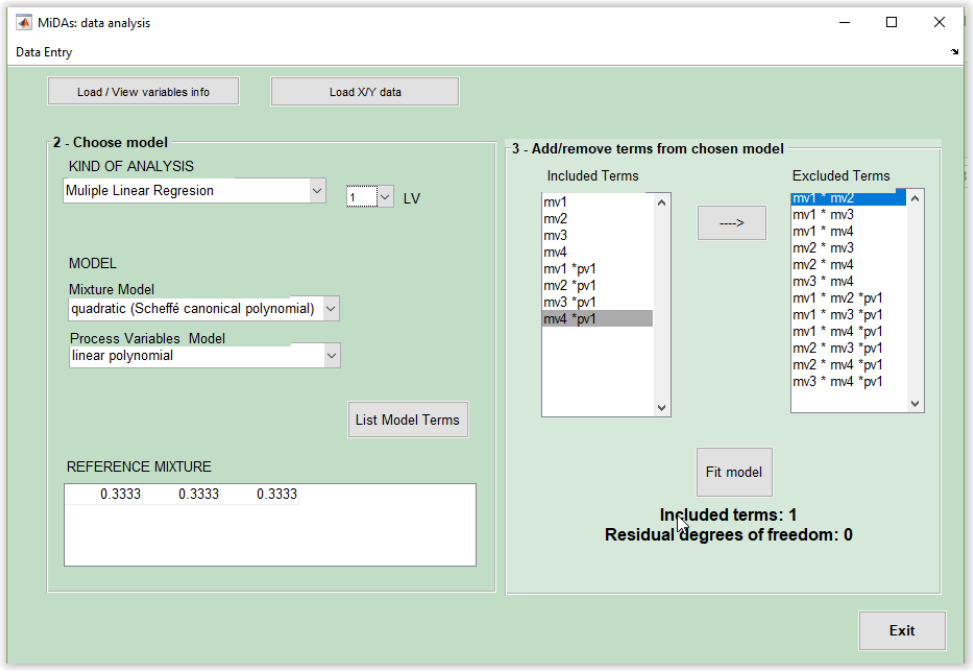


Figure 7.27. Start Data Analysis

- *Choose the analysis method (MLR or PLS)*

By default, MiDAs assumes that the model will be fitted via OLS, and the appropriate model structure is proposed. However, the user can define a base model (for mixture and process variable separately if the method “Multiple Linear Regression” is selected, or for both types of variables without distinctions if PLS is chosen).

Once the base model to be adjusted has been chosen, clicking on "List Model Terms" will generate the corresponding list of terms, some of which can be discarded if the user desires to do so (always respecting the hierarchical principle among terms concerning mixture variables). If MLR is resorted to and there are more terms in the model than observations in the experimental dataset, a warning will appear informing the user that the analysis cannot continue until enough terms have been eliminated from the model or more observations are included into the data table. If there are enough data to fit the model, the "Fit Model" button is displayed. Pressing it will open another screen, which will be different depending on whether the model is fitted using MLR or PLS.

7.5.1. Data analysis with MLR

The MLR Model window shows the results from fitting the selected model with the available data via OLS and, as seen in Figure 7.28, is separated into 4 parts. In the upper left corner, the model summary is available with the model parameters estimates, as well as the values for R^2 and R^2-adjusted. The results of the ANalysis Of VAriance (ANOVA) appear in the lower left side of the screen. A table of residuals is shown on the lower right side. In the upper right area, the traces plot can be seen. To the right of this graph, when there are process variables, sliding cursors (one for each process variable) are displayed to choose different values of the process variables, and a button to repeat the trace graph once the values of the process variables have been changed to the desired ones.

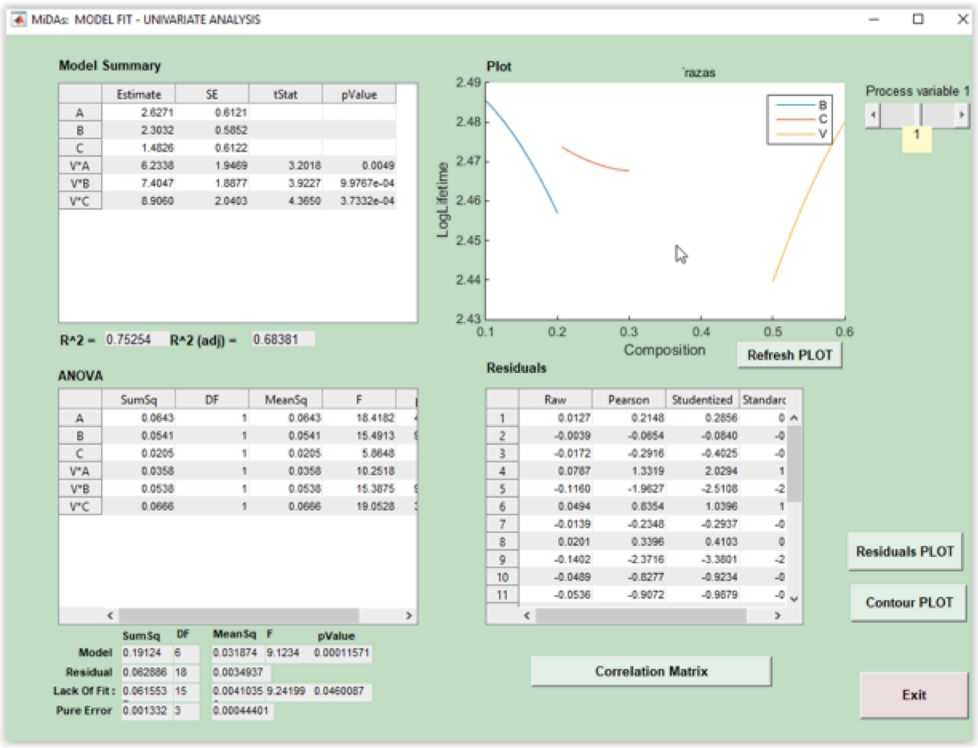


Figure 7.28. MLR

The "Contour" button to the right of the residuals table allows plotting the representation of the contours plot of the response surface for the model obtained (see Figure 7.29), after the proportions for all but 3 mixture components, and the values for all process variables, are fixed.

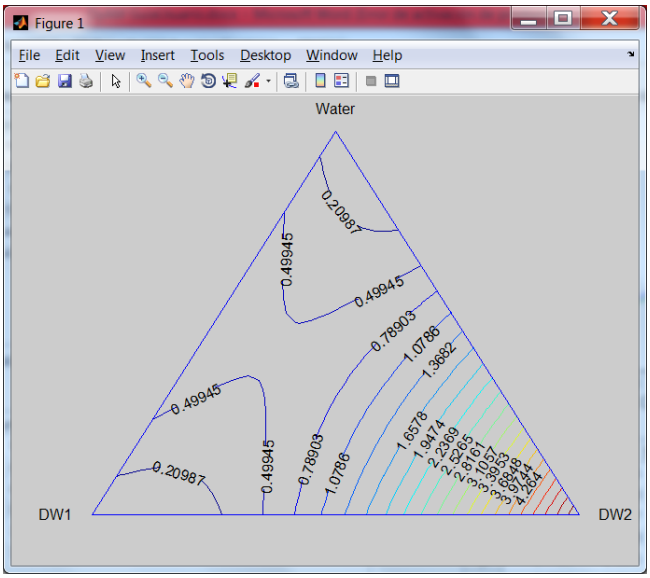


Figure 7.29. Contour plot

The "Residuals Plot" button, also to the right of the residuals table, displays when pressed a normal probability plot of the studentized residuals (see Figure 7.30).

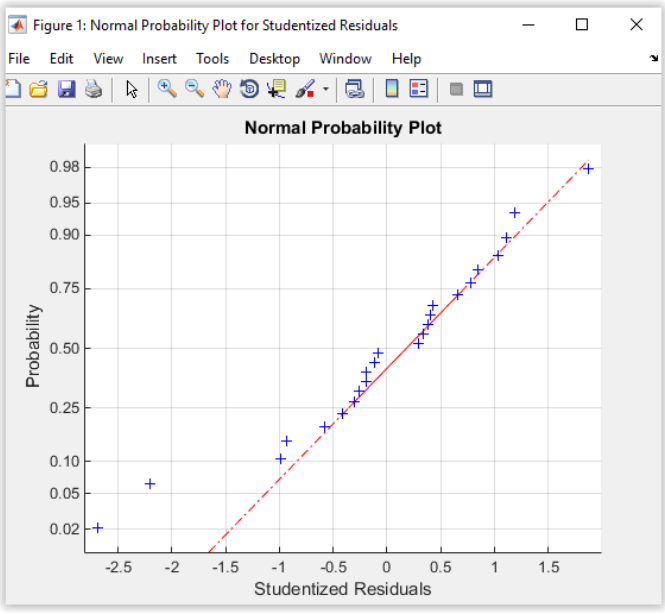
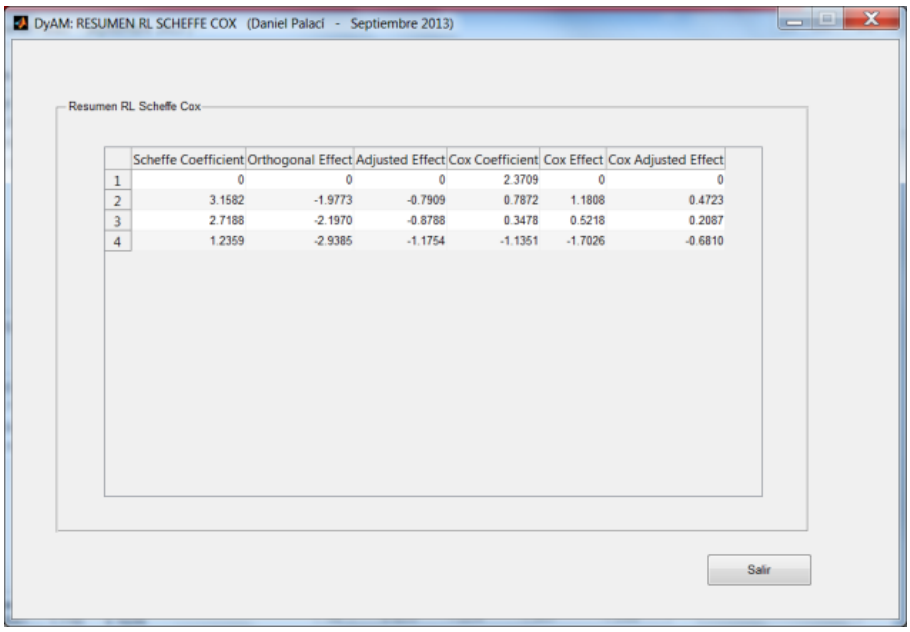


Figure 7.30. Normal probability chart of studentized residuals

Below the residuals table appears the "correlation matrix" view button, which represents the matrix of correlations between the parameters of the adjusted model (in the same order in which they are listed in the model summary table). If the adjusted model is linear, without process variables, and no terms have been removed from the model, the "summary RL (Scheffe-Cox)" view button also appears, which provides the table shown in Figure 7.31, that contains the Scheffé coefficients and the orthogonal and adjusted effects for the linear model, as well as the coefficients of the Cox model and the corresponding total and adjusted effects, with respect to the reference mixture introduced in the main analysis window.



	Scheffe Coefficient	Orthogonal Effect	Adjusted Effect	Cox Coefficient	Cox Effect	Cox Adjusted Effect
1	0	0	0	2.3709	0	0
2	3.1582	-1.9773	-0.7909	0.7872	1.1808	0.4723
3	2.7188	-2.1970	-0.8788	0.3478	0.5218	0.2087
4	1.2359	-2.9385	-1.1754	-1.1351	-1.7026	-0.6810

Figure 7.31. Sample of Scheffé and Cox summary table of linear regression

7.5.2. Data analysis with PLS

If the model is fitted via PLS, a window similar to that of the MLR is shown instead, where the model summary also includes the value of the Q^2, and the possibility to assess the adequacy of selecting more or less latent variables (considering, at most, as many as indicated in the main analysis window when fitting the model) via cross-validation.

PART III

Design Space definition and optimization through the latent space

Chapter 8
Preliminary considerations

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

1. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Sequential experimental approach to improve the design space estimation using latent-variable model inversion. Part I. Defining the Experimental Region. *Submitted*.

8.1. Quality by Design and the Design Space

The term Quality-by-Design (QbD) [83] has its origin in the pharmaceutical industry, and refers to an initiative that promotes the implementation of science-based methodologies to deliver products with the desired specifications by increasing the process flexibility and robustness, understood as the capability to tolerate changes in the materials involved and processing conditions (i.e. the inputs) without negatively affecting the quality of the outputs. In this context, the Design Space (DS) is a key concept defined as "the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality" [84]. The philosophy of the QbD, however, is not limited to the pharmaceutical field, but can be extended to any process where it may become relevant to find the most convenient combinations of inputs that guarantee the desired outputs (i.e. almost any manufacturing process).

An important point must be highlighted with respect to this, because applying the QbD approach requires proper understanding of the process it will be applied to. This approach is meant to be used in an active environment, where a model is used for control, optimization or the design of new products, when the desired properties are known, and the way to obtain has to be defined, as opposed to passive environments in which the model would be used for, e.g., calibration or monitoring purposes. Therefore, any already available model ought to be inverted, since at this point it will not be used to

predict a process outputs from some specific inputs, but to define the necessary inputs to achieve some specific outputs.

On the other hand, because of the way the DS is defined, this subspace is expected to comprise a subregion of the domain defined by the set of historical products that have already been developed [85,86], also referred to as knowledge space (KS)[x]. This, as well as the fact that in most cases a product is required to meet several specifications simultaneously, requires three scenarios to be considered, among which only one of them will guarantee the existence of the DS:

1. If at least one of the specifications on the outputs for a given hypothetical product cannot be met, independently of the inputs, the DS cannot be said to exist. Only relaxing or modifying entirely the restrictions imposed by such specifications may lead to the existence of a DS (although not for the product as initially defined).

2. If the specifications on the outputs for a given hypothetical product can only be met if operating outside of the KS, the DS cannot be said to exist, since it is assumed to be a subregion within the KS, not outside of it. However, whenever possible, expanding the KS by modifying the operating conditions such that the subregion of operating conditions that guarantee the desired product falls within it may, at least theoretically, enable the definition of the DS.

3. If all of the specifications on the outputs can be simultaneously met within the KS, the DS exists and may be comprised by a single combination of inputs (i.e. the dimensionality of the DS is equal to zero, or a single point) or different combinations of inputs (i.e. the dimensionality of the DS is equal or greater than one, a line, and at most that of the KS)

The DS, when it exists and its dimension is higher than zero, may be understood as the subspace of the KS within which any set of inputs leads to equally valid outputs in terms of meeting the defined specifications, therefore providing flexibility until additional constraints or some optimization criteria are considered. Thus, in some way, the DS may be interpreted as the set of all equally optimal solutions of an optimization problem.

Consider, as an example, the following polynomial model defining a hypothetical input-output causal relationship, used in one of the examples in [87]:

[x] This definition of the DS assumes that the desired quality attributes are defined such that they can be obtained from products similar to the already produced ones in the past. In practice, the goal may be finding the most appropriate operating conditions to achieve new, significantly different products from past ones. While this is not an scenario accounted for in the present thesis, this must be considered outside of the context of this work.

$$y = -21 + 4.3 \cdot x_1 + 0.022 \cdot x_2 - 0.0064 \cdot x_3 + 1.1 \cdot x_4 - 0.12 \cdot x_5$$
$$\text{s.t. } x_3 = x_1^2 \;;\; x_4 = x_2^2 \;;\; x_5 = x_1 \cdot x_2 \tag{8.1}$$

where y represents the 'quality attribute of interest' expressed as a function of the input variables x_1, x_2, x_3, x_4 and x_5. Notice that, although non-linear terms (interaction and quadratic terms) are included in Equation 8.1, this model is still linear in its parameters, and that only two of the five inputs (x_1 and x_2) are independent. Figure 8.1a illustrates the results from plotting the response surface plot corresponding to the quadratic model including the terms x_1, x_2, x_1^2, x_2^2 and $x_1 \cdot x_2$, by using noiseless data following Equation 8.1 without uncertainty in its parameters. Figure 8.1b shows the DS for a desired output $y_{\text{DES}} = 204.86$, as well as the DS for the more realistic case $184.86 \leq y_{\text{DES}} \leq 224.86$ (i.e. in practice, the DS usually encompasses ranges of acceptable values for the different quality attributes of interest, instead of specific ones), obtained after inverting Equation 8.1.

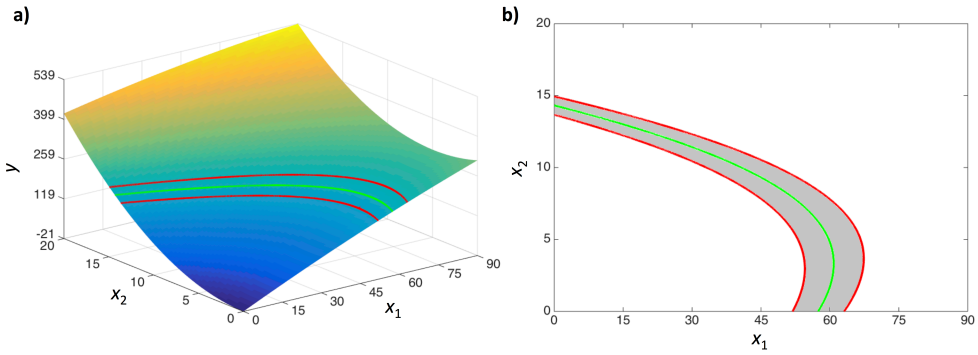


Figure 8.1 a) Response surface plot for the model in Equation 8.1 in the experimental domain corresponding to x_1 ranging from 0 to 90 and x_2 from 0 to 20, and its intersection with the planes corresponding to *y*=204.86 (green curve), and *y*=184.86 and *y*=224.86 (lower and upper red curves respectively), and b) design space, which contains all combinations of $\{x_1, x_2\}$ leading to *y*=204.86 (green line), *y*=184.86 and *y*=224.86 (lower and upper red curves respectively) and any combinations of $\{x_1, x_2\}$ leading to values for *y* between *y*=184.86 and *y*=224.86 (grey area)

As can be seen, if a range of values for the outputs within specified limits were is deemed acceptable instead of a single value, the DS is constituted by the union of the regions defined by the contour lines corresponding to the whole range of acceptable values within the specified limits (e.g. in Figure 1b, the grey area delimited by the two red curves).

Consider a more general example to illustrate the differences between the knowledge space, the design space and the so-called Normal Operating Conditions (NOC), assuming a process with only two input variables, x_1 and x_2, and two output variables y_1 and y_2 (see Figure 8.2).

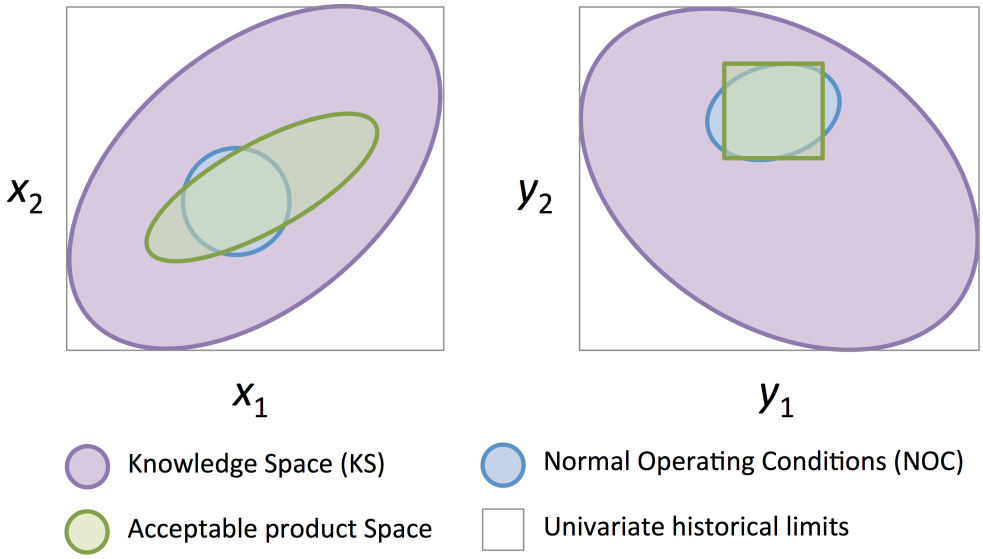


Figure 8.2 Hypothetical Knowledge Space (KS), Normal Operating Conditions (NOC), 'acceptable product space' (which is equivalent to the Design Space (DS) in the **X** space), and univariate historical limits in both the inputs (**X**) and outputs (**Y**) spaces for a generic production process with only two input and two output variables involved.

In Figure 8.2, the NOC are differentiated from the KS in that the KS is constituted by all possible combinations of operating conditions the process has ever been run under, and the characteristics of the product (y_1 and y_2) achieved, while the NOC represents only the subset of different combinations of inputs under which the process "currently" operates. Note that, as is usually the case in practice, a small portion of the production in NOC is outside of specifications (i.e. outside of the 'acceptable product space'), which are usually univariate. Furthermore, there is a larger region in the inputs subspace that would also guarantee the desired quality of the product, but is outside of the subspace of the NOC. This means that a proper definition of the DS would, in this example, increase the flexibility in the way the process is operating, while guaranteeing the desired outputs.

8.2. Limitations of the optimization and DOE in the original space

As commented in Section 2.1, independent variation in the input variables is required to guarantee causality when using data-driven approaches [3] and, although large amounts of 'happenstance' data are available in most production processes, the variation in the inputs is commonly not independent (i.e., data are not obtained from a DOE that guarantees this independent variation in the inputs). In these contexts, classical linear regression (LR) or even machine learning (ML) methods cannot be used for process optimization because none of many good prediction models that can be fitted is

unique or causal. This is because the process variables are usually highly correlated and the number of independent variations in the process is much smaller than the number of measured variables. Furthermore, when multiple quality attributes of interest (i.e. output variables) are involved, methods such as OLS present the added disadvantage of having to build one model for each of the outputs.

Therefore, since causal models are required for optimization, but building these models by methods such as OLS requires resorting to DOEs in the original space (i.e. the space of the original variables), the optimization problem in the original space will indirectly suffer from the same drawbacks as the DOE in the original space (e.g. the selection of a representative model structure for the data, the requirement of extensive experimentation to fit it, collinearity among factors, etc.). Furthermore, and even if first principle models can be resorted to, most (if not all) optimization algorithms will suffer from significant computational cost and convergence issues due to the high dimensionality of the space within which the optimization takes place in most productions processes, and the presumably also large number of restrictions imposed.

8.3. Optimization in the latent space

Methods based on projection onto latent structures, such as Partial Least Squares (PLS) regression-based techniques, allow the analysis of large datasets with highly correlated data by relating the inputs and outputs through few, uncorrelated Latent Variables (LV) that identify the underlying causal relationship between process inputs and outputs, and provide models for both the X (inputs) and Y (outputs) spaces. Since causal relationships can be inferred in the latent space [27], this permits the use of historical datasets for optimization purposes, which reduces the amount of experimentation required, or even prevents it altogether [28].

Furthermore, since the initial number of variables involved has been reduced to a smaller number of uncorrelated LV, the computational cost of any optimization problem in the latent space will be reduced when compared to the equivalent problem in the original space, even more so as the number of variables in the original space increases and the number of LV remain relatively few.

One of the main drawbacks of using PLS regression-based techniques is, as with any data-driven approach, the uncertainty associated to the data and the model, which must be accounted for in the definition of the DS in the latent space, the transferral of restrictions from the space of the original variables to the LV, and the definition of the optimization problem.

Assuming that at least one feasible combination of inputs exists that leads to the desired outputs, some approaches based on Latent Variable Regression Model Inversion (LVRMI) have been proposed in the past to obtain a single combination of inputs via the direct inversion, or to define the so-called Null Space (NS) which contains the projection onto the latent space of all the combinations of inputs theoretically guarantying

the desired outputs. Jaeckle and MacGregor [88] proposed the guidelines to obtain points on the NS when as many outputs as the rank of the matrix of outputs used to fit the model are considered, through the definition of what they identify as the null components. The subspace in which points obtained are located by using this approach is the so-called "combined pseudo-NS" by García-Muñoz et al [89]. These authors presented an alternative method to differentiate the NS associated to each output for which a desired value is specified, and propose a different approach to obtain points on each of these NS. Alternatively, an optimization problem may be solved to obtain such combination of inputs, depending on the restrictions imposed and the weight given to the different terms of the formulated objective function [90].

These data-driven methods, however, present several drawbacks:

1. They suffer from the uncertainty associated to such data, which is not explicitly taken into account neither when defining the NS, nor when solving the optimization problem, although this is an issue that has been addressed in the literature [91],[92]. Regarding this, DS bracketing has been suggested [87],[93] in order to define a restricted subspace within the KS, inside which the DS is expected to lay. Although experiments inside this region do not guarantee to provide the desired outputs, taking into account the uncertainty for this bracketing allows compensating for the lack of representativeness due to (relatively small) nonlinearities.

2. They do not provide an analytical expression for the NS, which is necessary to obtain the analytical expression for the limits of the confidence region associated to it, therefore making it impossible to use them as constraints in e.g. the optimization approach as formulated in [90].

3. They are limited to the case in which the quality attributes of interest coincide with some of/all the output variables considered when fitting the PLS-regression model. There are, however, some scenarios in which this may not suffice:

    a) If a quality attribute of interest (either an output or a linear combination of outputs) cannot be accurately explained by the PLS-regression model when included as an output variable, but its relationship with other output variables that are well explained by the model allows a linear combination of such outputs to be considered instead. As an example, let y_1 be the mass flow of a mixture resulting from a process, y_2 the proportion of the ingredient of interest in the mixture, and $y_3 = y_1 \cdot y_2$ the absolute mass flow of such ingredient. If the quality attribute of interest is y_2, but only y_1 and y_3 can be well predicted by the fitted PLS-regression model, then the estimation of the DS for a desired purity $y_{2,DES}$ will suffer from much

more uncertainty that would the definition of the NS for the linear combination $y_{2,DES} \cdot y_1 - y_3 = 0$.

b) If a quality attribute of interest can be expressed as a linear combination of outputs, but the weightings of the outputs involved in this linear combination suffer short-term changes (e.g. within a few hours, or even minutes). In this case, even if the quality attribute can be accurately explained by the PLS-regression model if included as another output before fitting such model, given a set of weightings, a different PLS-regression model would have to be fitted each time the weightings change. An example of this is the quality attribute being the expected monetary income from selling the products resulting from a manufacturing process, and the outputs being the generated mass or volume of each product. The monetary income can be expressed as a linear combination of these outputs weighted by their corresponding price per mass/volume unit, but frequent fluctuations in these prices would render a PLS-regression model including the incomes as an output useless every time one of these changes happens.

Chapter 9
Defining the design space in the latent space

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

1. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Sequential experimental approach to improve the design space estimation using latent-variable model inversion. Part I. Defining the Experimental Region. *Submitted*.

9.1. Partial Least Squares model fitting and prediction uncertainty

Partial Least Squares (PLS) regression [26] is a latent variable-based approach used to model the inner relationships between a matrix of inputs $\mathbf{X}$ [N×M] and a matrix of output variables $\mathbf{Y}$ [N×L] which was already presented in Section 3.2.2. In order to evaluate the model performance when projecting the n-th observation onto it, the Hotelling T_n^2 and the squared prediction error $SPE_{\mathbf{x}_n}$ are used, where T_n^2 estimates the squared Mahalanobis statistical distance in the latent subspace between the projected observation and the average one for the historical dataset $\mathbf{X}$, and $SPE_{\mathbf{x}_n}$ measures the squared Euclidean distance between the centred and scaled observation and its projection onto the latent subspace defined by the A LVs.

T_n^2 can be calculated as:

$$T_n^2 = \boldsymbol{\tau}_n^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot \boldsymbol{\tau}_n \tag{9.1}$$

with $\boldsymbol{\Lambda}^{-1}$ defined as the [A×A] diagonal matrix containing the inverse of the A variances of the scores associated to the LVs, and $\boldsymbol{\tau}_n = \mathbf{W}^{*\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot (\mathbf{x}_n - \mathbf{m}_{\mathbf{X}})$ the [A×1] vector of scores corresponding to the projection of the n-th observation $\mathbf{x}_n$ onto the latent subspace of the PLS model. $\mathbf{m}_{\mathbf{X}}$ and $\mathbf{D}_{\mathbf{s_X}}$ are, as in Section 3.2.2, the [M×1] column vector of centring factors and the [M×M] diagonal matrix of scaling factors applied to the M input variables before fitting the PLS-regression model, respectively.

On the other hand, given $\hat{\mathbf{x}}_n = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{X}}$:

$$SPE_{\mathbf{x}_n} = (\mathbf{x}_n - \hat{\mathbf{x}}_n)^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-2} \cdot (\mathbf{x}_n - \hat{\mathbf{x}}_n) = \mathbf{e}_n^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-2} \cdot \mathbf{e}_n \tag{9.2}$$

Assuming multinormally distributed scores and residuals for the projected observations, the T_n^2 and $SPE_{\mathbf{x}_n}$ values for a given observation that conforms to those in the calibration dataset are expected to be below their $100(1-\alpha)\%$ upper limits [94], T_{lim}^2 and SPE_{lim}, respectively:

$$\begin{aligned} T_{\mathrm{lim}}^2 &= \frac{(N-1)\cdot A}{N-A} \mathrm{F}_{M,N-M,\alpha} \\ SPE_{\mathrm{lim}} &= \frac{s_{\mathrm{SPE}}}{2\cdot \bar{x}_{\mathrm{SPE}}} \chi^2_{2\cdot \bar{x}_{\mathrm{SPE}}^2/s_{\mathrm{SPE}},\alpha} \end{aligned} \tag{9.3}$$

where $\mathrm{F}_{M,N-M,\alpha}$ is the $100(1-\alpha)$ percentile of a Fisher distribution with M and $(N–M)$ degrees of freedom, $\bar{x}_{\mathrm{SPE}}$ and s_{SPE} are the sample mean and standard deviation of the *SPE* for the observations from the calibration data set, and $\chi^2_{2\cdot \bar{x}_{\mathrm{SPE}}^2/s_{\mathrm{SPE}},\alpha}$ is the $100(1-\alpha)$ percentile of a χ^2 distribution with $2\cdot \bar{x}_{\mathrm{SPE}}^2/s_{\mathrm{SPE}}$ degrees of freedom.

Once the PLS-regression model has been fitted, it can be used directly to obtain the $[L\times 1]$ vector of average predictions of the output variables, $\hat{\mathbf{y}}_{\mathrm{obs}}$, given the $[M\times 1]$ vector corresponding to a particular observation, $\mathbf{x}_{\mathrm{obs}}$, fulfilling that $T_{\mathbf{x}_{\mathrm{obs}}}^2 \leq T_{\mathrm{lim}}^2$ and $SPE_{\mathbf{x}_{\mathrm{obs}}} \leq SPE_{\mathrm{lim}}$, as

$$\begin{aligned} \hat{\mathbf{y}}_{\mathrm{obs}} = \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau}_{\mathrm{obs}} + \mathbf{m}_{\mathbf{Y}} &= \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \mathbf{W}^{*\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot (\mathbf{x}_{\mathrm{obs}} - \mathbf{m}_{\mathbf{X}}) + \mathbf{m}_{\mathbf{Y}} \\ \hat{\mathbf{y}}_{\mathrm{obs}} &= \mathbf{B}^{\mathrm{T}} \cdot \mathbf{x}_{\mathrm{obs}} + \mathbf{b}_0 \\ \mathbf{B} &= \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{W}^* \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \\ \mathbf{b}_0 &= \mathbf{m}_{\mathbf{Y}} - \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \mathbf{W}^{*\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot \mathbf{m}_{\mathbf{X}} \end{aligned} \tag{9.4}$$

This prediction is not free from uncertainty. The estimation of prediction uncertainty is done by using Ordinary Least Squares (OLS) type expressions, as suggested in the approach proposed by Faber and Kowalski [95]. This approach assumes that the matrix of scores **T** obtained from fitting the PLS model is independent from **Y**, which is clearly not true [96]. However, this approximation was observed to lead to similarly good results when compared with those achieved by means of other approaches such as linearization or re-sampling based methods, or the so-called U-deviation method used in the software *the Unscrambler*, while being easier to implement [91].

The $100\cdot(1-\alpha)\%$ confidence interval (CI) of the prediction of the l-th output variable $\hat{y}_{\mathrm{obs},l}$ given an observation $\mathbf{x}_{\mathrm{obs}}$ is calculated as:

$$CI_{y_{\mathrm{obs},l}} = \hat{y}_{\mathrm{obs},l} \pm t_{N-df,\alpha/2} \cdot s_{\hat{y}_{\mathrm{obs},l}} \tag{9.5}$$

where $t_{N-df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of a Student's t-distribution with $(N-df)$ degrees of freedom, df being the degrees of freedom consumed by the PLS-

regression model, which are suggested in [91] to be calculated as indicated in [97], and $s_{\hat{y}_{\mathrm{obs},l}}$ the estimated standard deviation of the prediction of the *l*-th output variable for a particular observation $\mathbf{y}_{\mathrm{obs}}$:

$$s_{\hat{y}_{\mathrm{obs},l}} = s_{f_l} \cdot \sqrt{1 + h_{\mathrm{obs}} + \frac{1}{N}} \tag{9.6}$$

where h_{obs} is the observation leverage

$$h_{\mathrm{obs}} = \boldsymbol{\tau}_{\mathrm{obs}}^{\mathrm{T}} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}_{\mathrm{obs}} \tag{9.7}$$

and s_{f_l} the standard error of calibration corresponding to the *l*-th output, obtained as suggested in [91]

$$s_{f_l} = \sqrt{\frac{\sum_{n=1}^{N} \left(y_{n,l} - \hat{y}_{n,l}\right)^2}{N - df}} \tag{9.8}$$

$y_{n,l}$ and $\hat{y}_{n,l}$ being, respectively, the *n*-th measured and predicted value for the *l*-th output of the model calibration dataset.

It must be noted that, because of the way $s_{\hat{y}_{\mathrm{obs},l}}$ is computed by following this procedure, $CI_{y_{\mathrm{obs},l}}$ depends on the leverage of the observation, and therefore the amplitude of the confidence interval is expected to be lower for observations close to the centre of projection (smaller leverage) than for those far away from it (higher leverage). Alternatively, Romera [98] has more recently been proposed a local linearization-based approach to explicitly express the (asymptotic) variance for the estimators of the PLS-regression model parameters relating **X** to **Y**, **B**, which can be used to obtain a less biased estimate of $s_{\hat{y}_{\mathrm{obs},l}}$ than through the aforementioned approximation. In the following sections, however, the expression in Equations 9.5 to 9.8 will be resorted to due to them being easily applied also to input variables, and not just to the outputs.

9.2. Transferring restrictions to the latent space

Since the DS has been defined as a subregion of the KS, the very first step to take is the definition of such KS. Because a Latent Variable Regression Model (LVRM), and more specifically a PLS-regression model, is used, once univariate or multivariate restrictions on the inputs and outputs have been defined, these constraints should be transferred to the latent space (i.e. as restrictions on the scores). While the projection of these restrictions onto the LV subspace has been a standard part of e.g. the ProMV software, the analytical explicit formulation of such projections has not been provided in past literature, and are presented in this section. Having these expressions makes it possible to compare restrictions in the latent space among each other and with restrictions such as the ones imposed by the Hotelling T^2 hyper-ellipsoid.

Consider a set of inequality constraints imposed on the input and output variables to be expressed as:

$$\begin{aligned}\mathbf{A_x}\cdot\hat{\mathbf{x}}_{\mathrm{NEW}} &= \mathbf{A_x}\cdot\left(\mathbf{D}_{\mathbf{s_X}}\cdot\mathbf{P}\cdot\boldsymbol{\tau}_{\mathrm{NEW}}+\mathbf{m_X}\right)\le\mathbf{d}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}\\ \mathbf{A_y}\cdot\hat{\mathbf{y}}_{\mathrm{NEW}} &= \mathbf{A_y}\cdot\left(\mathbf{D}_{\mathbf{s_Y}}\cdot\mathbf{Q}\cdot\boldsymbol{\tau}_{\mathrm{NEW}}+\mathbf{m_Y}\right)\le\mathbf{d}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}\end{aligned} \tag{9.9}$$

$\mathbf{A_x}$ and $\mathbf{A_y}$ being $[I_1\times M]$ and $[J_1\times L]$ matrices whose *i*-th and *j*-th rows, respectively, contain the coefficients of the *i*-th linear combination of inputs and *j*-th linear combination of outputs; $\mathbf{d}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}$ and $\mathbf{d}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}$ being $[I_1\times 1]$ and $[J_1\times 1]$ column vectors whose *i*-th and *j*-th elements are, respectively, the values that must not be surpassed by the corresponding linear combination of inputs and outputs; and $\mathbf{m_X}$, $\mathbf{m_Y}$, $\mathbf{D}_{\mathbf{s_X}}$ and $\mathbf{D}_{\mathbf{s_Y}}$ being as defined in Section 3.2.2. I_1 and J_1 are, respectively, the number of inequality constraints imposed on the inputs and on the outputs.

After reorganizing terms in Equation 9.9, it yields:

$$\begin{aligned}\mathbf{A_x}\cdot\mathbf{D}_{\mathbf{s_X}}\cdot\mathbf{P}\cdot\boldsymbol{\tau}_{\mathrm{NEW}} &\le \mathbf{d}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}-\mathbf{A_x}\cdot\mathbf{m_X}\\ \mathbf{A_y}\cdot\mathbf{D}_{\mathbf{s_Y}}\cdot\mathbf{Q}\cdot\boldsymbol{\tau}_{\mathrm{NEW}} &\le \mathbf{d}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}-\mathbf{A_y}\cdot\mathbf{m_Y}\end{aligned} \tag{9.10}$$

Then, a set of inequality constraints in the latent space can be defined:

$$\begin{gathered}\mathbf{A}_{\boldsymbol{\tau}}\cdot\boldsymbol{\tau}_{\mathrm{NEW}}\le\mathbf{d}_{\boldsymbol{\tau}_{\mathrm{NEW}}}\\ \mathbf{A}_{\boldsymbol{\tau}}=\begin{bmatrix}\mathbf{A_x}\cdot\mathbf{D}_{\mathbf{s_X}}\cdot\mathbf{P}\\ \mathbf{A_y}\cdot\mathbf{D}_{\mathbf{s_Y}}\cdot\mathbf{Q}\end{bmatrix}\quad;\quad\mathbf{d}_{\boldsymbol{\tau}_{\mathrm{NEW}}}=\begin{bmatrix}\mathbf{d}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}-\mathbf{A_x}\cdot\mathbf{m_X}\\ \mathbf{d}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}-\mathbf{A_y}\cdot\mathbf{m_Y}\end{bmatrix}\end{gathered} \tag{9.11}$$

Similarly, if a set of equality constraints on some of the inputs and outputs is defined:

$$\begin{aligned}\mathbf{F_x}\cdot\hat{\mathbf{x}}_{\mathrm{NEW}} &= \mathbf{F_x}\cdot\left(\mathbf{D}_{\mathbf{s_X}}\cdot\mathbf{P}\cdot\boldsymbol{\tau}_{\mathrm{NEW}}+\mathbf{m_X}\right)=\mathbf{f}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}\\ \mathbf{F_y}\cdot\hat{\mathbf{y}}_{\mathrm{NEW}} &= \mathbf{F_y}\cdot\left(\mathbf{D}_{\mathbf{s_Y}}\cdot\mathbf{Q}\cdot\boldsymbol{\tau}_{\mathrm{NEW}}+\mathbf{m_Y}\right)=\mathbf{f}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}\end{aligned} \tag{9.12}$$

$\mathbf{F_x}$ and $\mathbf{F_y}$ being $[I_2\times M]$ and $[J_2\times L]$ matrices whose *i*-th and *j*-th rows, respectively, contain the coefficients of the *i*-th linear combination of inputs and *j*-th linear combination of outputs, defined in a similar way to $\mathbf{A_x}$ and $\mathbf{A_y}$ in Equations 9.9 to 9.11; and $\mathbf{f}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}$ and $\mathbf{f}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}$ being $[I_2\times 1]$ and $[J_2\times 1]$ column vectors whose *i*-th and *j*-th elements are, respectively, the values that the corresponding linear combination of inputs and outputs are required to have.

Following the same procedure as with the inequality constraints:

$$\begin{gathered}\mathbf{F}_{\boldsymbol{\tau}}\cdot\boldsymbol{\tau}_{\mathrm{NEW}}=\mathbf{f}_{\boldsymbol{\tau}_{\mathrm{NEW}}}\\ \mathbf{F}_{\boldsymbol{\tau}}=\begin{bmatrix}\mathbf{F_x}\cdot\mathbf{D}_{\mathbf{s_X}}\cdot\mathbf{P}\\ \mathbf{F_y}\cdot\mathbf{D}_{\mathbf{s_Y}}\cdot\mathbf{Q}\end{bmatrix}\quad;\quad\mathbf{f}_{\boldsymbol{\tau}_{\mathrm{NEW}}}=\begin{bmatrix}\mathbf{f}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}-\mathbf{A_x}\cdot\mathbf{m_X}\\ \mathbf{f}_{\hat{\mathbf{y}}_{\mathrm{NEW}}}-\mathbf{A_y}\cdot\mathbf{m_Y}\end{bmatrix}\end{gathered} \tag{9.13}$$

It should be noted that each equality restriction imposed implies the definition of a hyper-plane of dimensionality (A-1) inside the latent space of the model fitted with A latent variables, within which any $\boldsymbol{\tau}_{\mathrm{NEW}}$ must be located. Therefore imposing two

equality constraints requires $\boldsymbol{\tau}_{\mathrm{NEW}}$ to fall within the intersection of two (A-1)-dimensional hyper-planes, that is, within an (A-1)-dimensional subspace, if both restrictions are equivalent, or an (A-2)-dimensional subspace otherwise, if such intersection exists. This means that, at most, A compatible (i.e. an intersection exists), non-redundant equality constraints can be imposed, in which case a 0-dimensional (i.e. a point) subspace would result. Even then, there is no guarantee that this subspace falls within the subspace defined by any other inequality constraints that may have been imposed beforehand.

Furthermore, it may be argued that hard equality constraints are too restrictive, considering that the uncertainty associated to the data may make it the case that no single point within the latent space meets all of them, but a set of inputs slightly outside of it and with $SPE_{\mathbf{x}_{\mathrm{NEW}}} \leq SPE_{\mathrm{lim}}$ may be found. Therefore, while imposing some hard equality constraints such as the perfect linear correlation existing among the ingredients of a mixture (i.e. the proportions of all of the components of the mixture must sum up to 1, or 100%) may make sense, relaxing other equality constraints proportionally to the uncertainty in the prediction of the corresponding inputs, outputs or linear combinations of them could be advisable as a first step in any optimization problem, whenever not doing so makes finding any acceptable set of inputs impossible.

Consider, as an example, the dataset shown in Table 9.1, corresponding to five observations simulated following the input-output causal relationship in Equation 8.1, and adding noise to it so that it does not follow exactly such relationship.

Table 9.1. Simulated dataset via Equation 8.1 with added random normal noise

Observation (n)	x_1	x_2	$x_3(x_1^2)$	$x_4(x_2^2)$	$x_5(x_1 \cdot x_2)$	y
1	5.43	7.54	125.64	58.51	50.49	61.85
2	5.43	15.97	126.20	258.48	74.44	278.99
3	99.23	7.54	9893.38	59.29	737.15	307.89
4	99.23	15.97	9765.16	254.11	1576.28	436.40
5	52.33	11.76	2787.64	139.21	583.76	266.08
6	52.33	11.76	2849.95	135.67	630.73	260.52

Consider also the following inequality restrictions imposed on the inputs:

Table 9.2. Lower and upper limits imposed as restrictions on the inputs involved in the input-output causal relationship in Equation 8.1

Variable	Lower limit	Upper limit
x_1	5.43	99.23
x_2	7.54	15.97
$x_3(x_1^2)$	125.64	9893.38
$x_4(x_2^2)$	58.51	258.48
$x_5(x_1 \cdot x_2)$	50.49	1576.28

It must be noted that the limits imposed as univariate restrictions on the inputs correspond the lowest and highest observed values for the corresponding variables in the dataset shown in Table 9.1. Therefore, all of such observations fall within these limits when the restrictions are considered in the space of the original variables. Furthermore, the univariate restrictions on x_3, x_4 and x_5 are imposed since, in practice, it is assumed that the true relationship among the variables and x_1 and x_2 would not be known.

Consider now a PLS-regression model fitted with 2 LV, the univariate restrictions in Table 9.2 to be transferred to the latent space, and the observations in Table 9.1 to be projected onto the model subspace. Figure 9.1 illustrates the result from the projection of these restrictions (blue dashed line), the six observations (black crosses accompanied by the number indicating which observation they correspond to) and the Hotelling T^2 95% confidence ellipse (magenta dashed line).

As seen in Figure 9.1, only five of the ten imposed univariate inequality constraints remain when projected onto the latent subspace of the PLS-regression model, with the remaining five being redundant. This shows that the reduction in dimensionality also affects which restrictions remain relevant. Furthermore, only the projection of half of the six observations (3, 5 and 6) fall within the region delimited by the projection of the univariate inequality constraints, which would represent an estimation the KS, with the other three (1, 2 and 4) being outside of it.

It must be noted that all observations present $SPE_{\mathbf{x}_n}$ values below its 95% confidence limit, $SPE_{\lim} = 7.63$, as calculated in Equation 9.3, except for the fourth observation ($SPE_{x_4} = 8.29$).

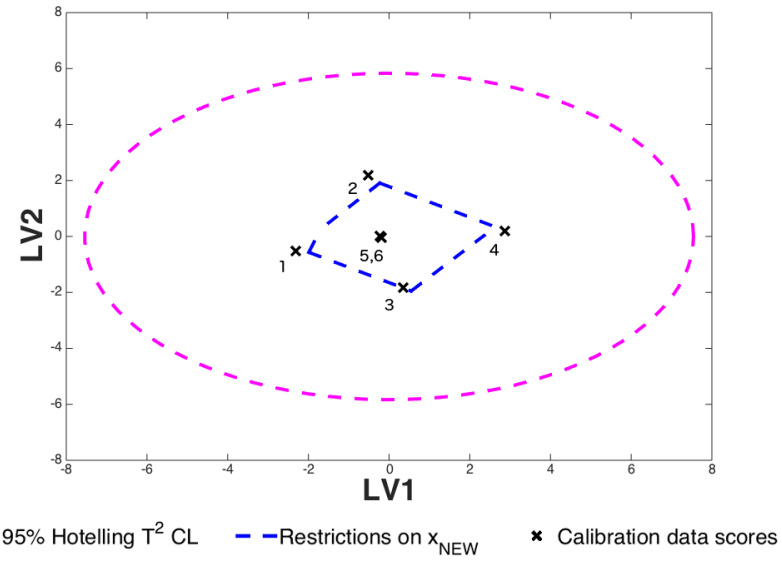


Figure 9.1 Representation of the restrictions in Table 9.2 transferred to the latent space (dashed blue lines) corresponding to a PLS-regression model fitted with 2 LV and the data in Table 9.1, and projection of these data onto the same subspace (black 'x') with one/two numbers besides close to each identifying the number of the observation in Table 9.1.

Therefore, the fact that the projection of these observations falls outside of the region delimited by the projection of the restrictions that all observations actually meet in the original space can only be explained by the propagation of the uncertainty in the prediction of the inputs to the projection of the restrictions imposed on them. In fact, although the PLS-regression model fitted with 2 LV explains more than 97% of the variability in the X space ($R^2(\mathbf{X}) \approx 0.9714$), the unexplained variability is enough for this to happen.

If a PLS-regression model is fitted using the same data, but with 4 LV instead of 2 LV, all of the projected observations will meet all of the projected restrictions. In practice, when a larger number of variables are involved and equality restrictions may also be imposed, increasing the number of LV would present several disadvantages, such as i) introducing additional, undesired noise into the model as 'information', and ii) increasing the dimensionality of the model subspace, which will also increase the computational cost in any optimization problem and in the definition of the DS.

An alternative approach would include relaxing these constraints to account for the uncertainty. To do so in a sensible way, however, a procedure would be required to back-propagate this uncertainty such that analytical expressions for the new, relaxed restrictions are obtained, but such procedure has not yet been made available in the literature. An approach to do this is proposed in Section 9.3.

9.3. Partial Least Squares model inversion

9.3.1. The direct inversion

The most common use of a PLS model is in its direct form, that is, as a predictive tool for the output variables given some input data. However, the same model can be used in its inverse form to suggest a combination of inputs $\mathbf{x}_{\mathrm{NEW}}$ [M×1], with projection $\boldsymbol{\tau}_{\mathrm{NEW}}$, needed to achieve a set of desired values $\mathbf{y}_{\mathrm{DES}}$ for the outputs, $\mathbf{y}_{\mathrm{NEW}} = \mathbf{y}_{\mathrm{DES}}$. In theory, this will be valid as long as $\mathbf{y}_{\mathrm{DES}}$ conforms to the correlation structure from the calibration dataset used to fit the PLS-regression model.

As pointed out in [86], however, since a reduction in the dimensionality of the space has taken place (from operating in the space of the original variables to doing so in the latent space), three possible scenarios must be taken into account:

I. *L>A*: in this case, since the number of outputs (L) for which desired values are defined is greater than the dimensionality of the latent space (A), the most likely case is that no solution exists that provides the desired values for all of the outputs.
II. *L=A:* in this situation, a single solution exists that provides all $\mathbf{y}_{\mathrm{DES}}$
III. *L<A:* this is the most common scenario, and corresponds to a case in which multiple sets of inputs $\mathbf{x}_{\mathrm{NEW}}$ and their projections $\boldsymbol{\tau}_{\mathrm{NEW}}$ will theoretically lead to the same desired outputs $\mathbf{y}_{\mathrm{DES}}$.

For the first scenario, the least squares solution can be obtained as indicated by Equation 9.14:

$$\boldsymbol{\tau}_{\mathrm{NEW}} = (\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{Q})^{-1} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}}{}^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{9.14}$$

When applied to the second and third scenarios, this solution is sometimes also referred to as the direct inversion of the PLS model, and provides a single set of scores $\boldsymbol{\tau}_{\mathrm{NEW}}$ corresponding to a set of inputs (i.e., the combination of initial settings, process conditions, raw material properties, etc.) theoretically leading to $\mathbf{y}_{\mathrm{DES}}$, as well as the inputs themselves, $\hat{\mathbf{x}}_{\mathrm{NEW}}$.

This solution is unique in the second scenario. In the third scenario, a subspace usually referred to as Null Space (NS) can be found in which, according to the model, the same set of desired values for the outputs, $\mathbf{y}_{\mathrm{DES}}$, will be theoretically obtained for different inputs combinations. In such case, the solution in Equation 9.14 it corresponds, among all of those input combinations, to the one whose projection onto the latent space is closest to the centre of projection of the model subspace (not the one with lowest leverage), as demonstrated in Section 14.3.1.

In the second scenario a more correct expression for the solution of the direct inversion is shown in Equation 9.15.

$$\boldsymbol{\tau}_{\mathrm{NEW}} = \mathbf{Q}^{-1} \cdot \mathbf{D}_{\mathbf{s_Y}}^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{9.15}$$

Finally, for the third scenario:

$$\boldsymbol{\tau}_{\mathrm{NEW}} = \mathbf{Q}^{\mathrm{T}} \cdot (\mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}})^{-1} \cdot \mathbf{D}_{\mathbf{s_Y}}^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{9.16}$$

And for all three different scenarios:

$$\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau}_{\mathrm{NEW}} + \mathbf{m}_{\mathbf{X}} \tag{9.17}$$

9.3.2. Direct inversion-dependant definitions of the Null Space

According to [88], the solution to the LVRMI is unique if the effective rank of $\mathbf{Y}$, $r_{\mathbf{Y}}$, is the same as the number of LV used to fit the PLS regression model (i.e. $r_{\mathbf{Y}} = A$). Otherwise, as mentioned in [88] the $(A - r_{\mathbf{Y}})$-dimensional NS can be found. To guarantee this, and as mentioned in [88], first a selective PCA [99] is performed in order to select the $r_{\mathbf{Y}}$ outputs most representative of $\mathbf{Y}$. The resulting subspace is usually referred to as Null Space (NS) because, according to the PLS model, no change in the outputs is expected when moving from one point to another on it, and any set of inputs $\mathbf{x}_{\mathrm{NS}}$ associated to it can be obtained as:

$$\mathbf{x}_{\mathrm{NS}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \mathbf{G} \cdot \mathbf{r} + \mathbf{m}_{\mathbf{X}} \tag{9.18}$$

where $\mathbf{G}$ is an $[A \times (A - r_{\mathbf{Y}})]$ matrix whose columns are the left singular vectors of $\mathbf{Q}$ associated with its $(A - r_{\mathbf{Y}})$ zero singular values, and $\mathbf{r}$ is an $[(\mathrm{A} - r_{\mathbf{Y}}) \times 1]$ vector of random real values used to obtain arbitrary points on the NS. Which of these combinations of inputs are feasible or practical will ultimately depend on the constraints imposed on them [88].

It must be noted that this approach implicitly assumes that $r_{\mathbf{Y}} = L$. Otherwise, a selection of the most representative output variables must be performed beforehand to reduce the dimensionality of the $\mathbf{Y}$ matrix until $r_{\mathbf{Y}} = L$. Since the 'discarded' output variables will be highly correlated with some/all of the ones left when fitting the model, the values they take them can be computed afterwards.

A slightly different approach has also been used in the literature [89], where the term 'combined pseudo null-space' is used to refer to the subspace in which points meeting Equations 9.18 are obtained, and different NS are defined for each output variable. This alternative definition is based on the fact that, if a displacement $\Delta\boldsymbol{\tau}_{\mathrm{NEW}}$ from the direct inversion solution to any other point on the latent space is produced, such that $\Delta\mathbf{y}_{\mathrm{NEW}} = \mathbf{0}$, then:

$$\mathbf{Q} \cdot \Delta\boldsymbol{\tau}_{\mathrm{NEW}} = \mathbf{0} \tag{9.19}$$

Which can be reformulated as

$$\begin{aligned} q_{1,1} \cdot \Delta\tau_{\mathrm{NEW},1} + q_{1,2} \cdot \Delta\tau_{\mathrm{NEW},2} + \cdots + q_{1,A} \cdot \Delta\tau_{\mathrm{NEW},A} = 0 \\ q_{2,1} \cdot \Delta\tau_{\mathrm{NEW},1} + q_{2,2} \cdot \Delta\tau_{\mathrm{NEW},2} + \cdots + q_{2,A} \cdot \Delta\tau_{\mathrm{NEW},A} = 0 \\ \vdots \\ q_{L,1} \cdot \Delta\tau_{\mathrm{NEW},1} + q_{L,2} \cdot \Delta\tau_{\mathrm{NEW},2} + \cdots + q_{L,A} \cdot \Delta\tau_{\mathrm{NEW},A} = 0 \end{aligned} \tag{9.20}$$

where the 'combined pseudo null-space' corresponds to the hyper-plane that best approximates all of the L equalities in Equation 9.20 and contains the result of the direct inversion obtained through Equations 9.16 and 9.17, while each equality provides the directions that, together with the direct inversion, define the NS of one of the outputs.

It is important to note that these approaches are implicitly limited by the fact that they require the solution of the direct inversion to provide the desired values for all of the outputs. Otherwise neither the 'combined pseudo null-space' nor some/any of the NS associated to each of the outputs, as defined in [89] will contain any point satisfying its own definition.

9.3.3. Analytical definition of the Null Space

An alternative approach for the definition of the NS is presented in this section, which does not suffer from the limitations, associated to the formulation of the NS as done in Section 9.3.2. Namely, the presented method does not depend on the existence of at least one combination of inputs/scores leading to the desired values for all of the outputs. Additionally, it provides a more informative, analytical definition of the NS associated to each output variable, as will be later discussed.

To this end, consider the l-th NS to be defined as the subspace constituted by all combinations of inputs that theoretically guarantee the desired value for the l-th (with l=1, 2, …, L) output variable, but not necessarily for the rest.

Then, given a desired value $y_{\mathrm{DES},l}$ for the l-th output variable, a [L×1] vector $\hat{\mathbf{y}}_{\mathrm{NS}_l}$ corresponding to any set of values for the outputs associated to a set of scores $\boldsymbol{\tau}_{\mathrm{NS}_l}$ on the l-th NS can be found such that

$$\mathbf{o}_l^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NS}_l} = \mathbf{o}_l^{\mathrm{T}} \cdot \left(\mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau}_{\mathrm{NS}_l} + \mathbf{m}_{\mathbf{Y}}\right) = \mathbf{o}_l^{\mathrm{T}} \cdot \mathbf{y}_{\mathrm{DES}} = y_{\mathrm{DES},l} \tag{9.21}$$

with $\mathbf{o}_l$ being the [L×1] vector whose l-th element is equal to one, and the rest are zeros.

Equation 9.21 can be reorganized as:

$$\mathbf{o}_l^{\mathrm{T}} \cdot (\mathbf{m}_{\mathbf{Y}} - \mathbf{y}_{\mathrm{DES}}) + \mathbf{o}_l^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau}_{\mathrm{NS}_l} = 0 \tag{9.22}$$

which is the equation of a hyper-plane in the latent space associated to the l-th NS of the form:

$$\begin{gathered} v_{l,0} + \sum_{a=1}^{\mathrm{A}} v_{l,a} \cdot \tau_{\mathrm{NS}_l,a} = 0 \; ; \mathbf{v}_l = [v_{l,1}, v_{l,2}, \ldots, v_{l,A}]^{\mathrm{T}} \\ v_{l,0} = \mathbf{o}_l^{\mathrm{T}} \cdot (\mathbf{m}_{\mathbf{Y}} - \mathbf{y}_{\mathrm{DES}}) \\ \mathbf{v}_l = \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}} \cdot \mathbf{o}_l \end{gathered} \tag{9.23}$$

When applied to all L output variables:

$$\begin{gathered} \mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}_{\mathrm{NS}} = \mathbf{0} \\ \mathbf{v}_0 = \begin{bmatrix} v_{1,0} \\ v_{2,0} \\ \vdots \\ v_{L,0} \end{bmatrix} = \mathbf{m}_{\mathbf{Y}} - \mathbf{y}_{\mathrm{DES}} \quad ; \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_L^{\mathrm{T}} \end{bmatrix} = \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}} \cdot \mathbf{Q} \end{gathered} \tag{9.24}$$

$\mathbf{y}_{\mathrm{DES}}$ being the [L×1] column vector with $y_{\mathrm{DES},l}$ as its l-th element, and $\boldsymbol{\tau}_{\mathrm{NS}}$ a set of scores in the intersection of the L NS (if it exists).

It must be noted that the [L×(A+1)] matrix $[\mathbf{v}_0\ \mathbf{V}]$ contains in each row the (A+1) parameters of the equation of the NS associated to the desired valued for the l-th quality attribute, with the l-th element in the first column being the expected difference between the centring factor for the l-th output and its desired value, and the A remaining corresponding to $\mathbf{v}_l$, the vector signalling the direction of maximum variability in the latent space for the expected value of the l-th output, such that a displacement of module one in the same direction, in the latent space, will produce an increase in the expected value of the l-th output equal to the module of $\mathbf{v}_l$.

Furthermore, as long as the rank of $\mathbf{X}$, $r_{\mathbf{X}}$, is higher than one (the rank corresponding to one output variable) it will be possible to define a NS corresponding to the desired value of one of the outputs. The intersection of all these NS, if it exists, will be (A-L)-dimensional subspace and that may be referred to as 'combined NS', and coincident with the 'combined pseudo-NS' presented in [88] only if A>L. This subspace can be expressed as function of the inputs instead of the scores by simply considering the relationship among them expressed in the third equality in Equation 3.4.

It is also important to note that this definition of the NS only requires its position ($v_{l,0}$) and a single vector orthogonal to the NS ($\mathbf{v}_l$), unlike the definition in [89], which requires a point in the NS and (A-1) linearly independent vectors parallel to it.

An added benefit of this formulation is the possibility to extend the definition of the NS not just to the desired values for the outputs, but also for a linear combination of them.

Consider R quality attributes to be defined as a linear combination of the L outputs, and the r-th quality attribute of interest to be expressed as $d_r = \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{y}$, with $\mathbf{a}_r$ being the [L×1] vector that contains the coefficients that relate the L output variables with the r-th quality attribute. Let the r-th NS be the subspace constituted by all combinations of

inputs that theoretically guarantee the desired value for the *r*-th quality attribute of interest, but not necessarily for the rest. Then, given a desired value $d_{\mathrm{DES},r}$ for the *r*-th quality attribute, a [L×1] vector $\hat{\mathbf{y}}_{\mathrm{NS}_r}$ corresponding to any set of values for the outputs associated to a set of scores $\boldsymbol{\tau}_{\mathrm{NS}_r}$ on the *r*-th NS can be found such that

$$\mathbf{a}_r^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NS}_r} = \mathbf{a}_r^{\mathrm{T}} \cdot \left(\mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau}_{\mathrm{NS}_r} + \mathbf{m}_{\mathbf{Y}}\right) = d_{\mathrm{DES},r} \tag{9.25}$$

Equation 9.25 is simply a more general expression of Equation 9.21, where $\mathbf{o}_l$ has been substituted by $\mathbf{a}_r$, and therefore the same steps can be followed so that, in the end:

$$\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}_{\mathrm{NS}} = \mathbf{0}$$
$$\mathbf{v}_0 = \begin{bmatrix} v_{1,0} \\ v_{2,0} \\ \vdots \\ v_{R,0} \end{bmatrix} = \mathbf{A}_{\mathrm{r}} \cdot \mathbf{m}_{\mathbf{Y}} - \mathbf{d}_{\mathrm{DES}} \quad ; \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_R^{\mathrm{T}} \end{bmatrix} = \mathbf{A}_{\mathrm{r}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \tag{9.26}$$

$\mathbf{A}_{\mathrm{r}}$ being a [R×L] matrix with $\mathbf{a}_r^{\mathrm{T}}$ as its *r*-th row, $\mathbf{d}_{\mathrm{DES}}$ a [R×1] column vector with $d_{\mathrm{DES},r}$ as its *r*-th element, and $\boldsymbol{\tau}_{\mathrm{NS}}$ a set of scores in the intersection of the R NS (if it exists). Again, as long as the rank of $\mathbf{X}$, $r_{\mathbf{X}}$, is higher than one (the rank corresponding to one quality attribute) it will be possible to define a NS corresponding to the desired value of one of the quality attributes of interest.

It should be noted that Equation 9.24 is a particular case of Equation 9.26 where $\mathbf{A}_{\mathrm{r}} = \mathbf{I}_L$, $\mathbf{I}_L$ being the [L×L] identity matrix.

9.3.4. Confidence region of the Null Space

As previously stated, neither prediction is free from uncertainty, and neither is any result achieved by model inversion. Several studies have been carried out and different proposals exist that aim at providing an estimation of the DS. In particular, Tomba et al. [90] propose a jackknife approach to determine the confidence regions for both the direct inversion and the NS estimation. On the other hand, a Bayesian-based procedure is suggested by Bano et al. [100] for the probabilistic definition of the DS in the latent space. Both of these approaches, however, present the drawback of being computationally costly as dataset size and/or the dimensionality of the latent space increases. Alternatively, Facco et al. [87] propose a methodology to account for the back-propagation of the uncertainty in model inversion to bracket the design space, which can be used whenever a single output is considered. The steps to apply it can be summarized as follows:

1. Fit a PLS-regression model relating the inputs with the output of interest
2. Define the desired value for the output, y_{DES}.
3. Obtain the direct inversion $\hat{\mathbf{x}}_{\mathrm{NEW}}$ and $\boldsymbol{\tau}_{\mathrm{NEW}}$ as in Equations 9.16 and 9.17.
4. Determine the $100(1-\alpha)\%$ CI of the prediction of the output variable given $\boldsymbol{\tau}_{\mathrm{NEW}}$

5. Draw the two NS corresponding to the limits of the CI defined in step 4, by using Equation 9.18.
6. Define the "experimental space" as the subspace of the latent space inside the Hotelling T^2 confidence hyper-ellipsoid and delimited by the two NS obtained in step 5.

The DS estimate obtained this way can be understood as a subspace within which experimentation can be performed to more easily identify the true DS.

However, since an analytical expression for the NS can be obtained, as illustrated in Section 9.3.3, defining analytically the limits of the confidence region associated to it is also possible. Consider Equations 9.5 to 9.8 for the $100 \cdot (1-\alpha)\%$ CI of the prediction of the *l*-th output variable $\hat{y}_{\text{obs,l}}$ given an observation $\mathbf{x}_{\text{obs}}$. Then, the confidence region associated to the NS for the *l*-th output, whose equation can be obtained through Equation 9.23, can be defined by back-propagating this uncertainty when inverting the model. It can then be demonstrated (see Section 14.3.2) that any new sample $\mathbf{x}_{\text{NEW}}$ with projection $\boldsymbol{\tau}_{\text{NEW}}$ onto the latent space within the confidence limits for the estimation of the *l*-th NS must meet the following condition:

$$-t_{N-A,\alpha/2} \leq \frac{v_{l,0} + \mathbf{v}_l^{\text{T}} \cdot \boldsymbol{\tau}_{\text{NEW}}}{s_{\tilde{y}_{\text{NEW},l}}} \leq t_{N-A,\alpha/2}$$

$$s_{\tilde{y}_{\text{NEW},l}} = SE_{\hat{y}_l} \cdot \sqrt{1 + h_{\boldsymbol{\tau}_{\text{NEW}}} - \left(\frac{v_{l,0} + \mathbf{v}_l^{\text{T}} \cdot \boldsymbol{\tau}_{\text{NEW}}}{\mathbf{v}_l^{\text{T}} \cdot \mathbf{v}_l}\right)^2 \cdot \mathbf{v}_l^{\text{T}} \cdot (\mathbf{T}^{\text{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v}_l + \frac{1}{N}} \quad (9.27)$$

$s_{\tilde{y}_{\text{NEW},l}}$ being the estimated standard deviation of the *l*-th output for a hypothetical observation $\tilde{\mathbf{x}}_{\text{NEW}}$ with projection onto the latent space $\tilde{\boldsymbol{\tau}}_{\text{NEW}}$ such that $\tilde{\boldsymbol{\tau}}_{\text{NEW}}$ is the set of scores located in the *l*-th NS (corresponding to the desired value for the *l*-th output) closest to $\boldsymbol{\tau}_{\text{NEW}}$.

This suggested formulation of the confidence interval can be used to define an experimental region as in [87] but slightly differs from it in that this one takes into account the leverage of the new sample, leading to non-linear confidence limits that are tangent in the narrowest point to the limits as calculated in [87] when only one output is considered and using standardized scores.

However, since in practice the DS will usually encompass a range of acceptable values for the outputs/quality attributes of interest, and not a specific single value, a more realistic definition of the experimental region will account for this. To do so, consider the Lower Specification Limit (LSL), $y_{\text{LSL},l}$, and the Upper Specification Limit, $y_{\text{USL},l}$, such that a new product will be considered to be within specifications regarding the *l*-th output if $y_{\text{LSL},l} \leq y_{\text{NEW},l} \leq y_{\text{USL},l}$. Two NS may be defined associated to the LSL and USL, to which hyperplanes of the form shown in Equation 9.23 correspond with intercepts $v_{\text{LSL},l,0} = \mathbf{o}_l^{\text{T}} \cdot \mathbf{m}_{\mathbf{Y}} - y_{\text{LSL},l}$ and $v_{\text{USL},l,0} = \mathbf{o}_l^{\text{T}} \cdot \mathbf{m}_{\mathbf{Y}} - y_{\text{USL},l}$, respectively. Then, any new sample $\mathbf{x}_{\text{NEW}}$ with projection $\boldsymbol{\tau}_{\text{NEW}}$ onto the latent space within this experi-

mental region must meet:

$$-t_{N-A,\alpha/2} \leq \frac{v_{\text{LSL},l,0} + \mathbf{v}_l^{\text{T}} \cdot \boldsymbol{\tau}_{\text{NEW}}}{s_{\tilde{y}_{\text{NEW,LSL},l}}} \;;\; \frac{v_{\text{USL},l,0} + \mathbf{v}_l^{\text{T}} \cdot \boldsymbol{\tau}_{\text{NEW}}}{s_{\tilde{y}_{\text{NEW,USL},l}}} \leq t_{N-A,\alpha/2}$$

$$s_{\tilde{y}_{\text{NEW,LSL},l}} = SE_{\hat{y}_l} \cdot \sqrt{1 + h_{\boldsymbol{\tau}_{\text{NEW}}} - \left(\frac{v_{\text{LSL},l,0} + \mathbf{v}_l^{\text{T}} \cdot \boldsymbol{\tau}_{\text{NEW}}}{\mathbf{v}_l^{\text{T}} \cdot \mathbf{v}_l}\right)^2 \cdot \mathbf{v}_l^{\text{T}} \cdot (\mathbf{T}^{\text{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v}_l + \frac{1}{N}} \quad (9.28)$$

$$s_{\tilde{y}_{\text{NEW,USL},l}} = SE_{\hat{y}_l} \cdot \sqrt{1 + h_{\boldsymbol{\tau}_{\text{NEW}}} - \left(\frac{v_{\text{USL},l,0} + \mathbf{v}_l^{\text{T}} \cdot \boldsymbol{\tau}_{\text{NEW}}}{\mathbf{v}_l^{\text{T}} \cdot \mathbf{v}_l}\right)^2 \cdot \mathbf{v}_l^{\text{T}} \cdot (\mathbf{T}^{\text{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v}_l + \frac{1}{N}}$$

The adaptation of the expression in Equations 9.27 and 9.28 for the NS corresponding to the *r*-th output variable requires substituting the expression for the NS associated to the desired value (in Equation 9.27) or the LSL and USL (in Equation 9.28) for the *l*-th output with that for the corresponding values for the *r*-th quality attribute of interest. Estimating the standard deviation of the *r*-th quality attribute of interest, expressed as a linear combination of outputs, is also needed. To do this, consider *R* quality attributes of interest to be defined, each being a linear combination of the *L* outputs, such that the *r*-th quality attribute d_r for a given observation x_{obs} with outputs y_{obs} is expressed as:

$$d_{\text{obs},r} = \mathbf{a}_r^{\text{T}} \cdot \mathbf{y}_{\text{obs}} \quad (9.29)$$

$\mathbf{a}_r$ being the [L×1] vector that contains the coefficients relating the L output variables with the *r*-th quality attribute of interest.

The $100 \cdot (1 - \alpha)\%$ CI of the prediction of $d_{\text{obs},r}$ given an observation $\mathbf{x}_{\text{obs}}$ is calculated as:

$$\text{CI}_{d_{\text{obs},r}} = \mathbf{a}_r^{\text{T}} \cdot \hat{\mathbf{y}}_{\text{obs}} \pm t_{N-A,\alpha/2} \cdot s_{\hat{d}_{\text{obs},r}} \quad (9.30)$$

$s_{\hat{d}_{\text{obs},r}}$ being the estimated standard deviation of the *r*-th linear combination of outputs for an observation $\mathbf{x}_{\text{obs}}$, which can be obtained as:

$$s_{\hat{d}_{\text{obs},r}} = \sqrt{\mathbf{a}_r^{\text{T}} \cdot \mathbf{S}_{\mathbf{f}} \cdot \mathbf{a}_r \cdot \left(1 + h_{\text{obs}} + \frac{1}{N}\right)} \quad (9.31)$$

where $\mathbf{S}_{\mathbf{f}}$ is the variance-covariance matrix of prediction errors:

$$\mathbf{S_f} = \begin{pmatrix} s_{f_1}^2 & \widehat{\text{cov}}(f_1, f_2) & \cdots & \widehat{\text{cov}}(f_1, f_L) \\ \widehat{\text{cov}}(f_2, f_1) & s_{f_2}^2 & & \widehat{\text{cov}}(f_2, f_L) \\ \vdots & & \ddots & \vdots \\ \widehat{\text{cov}}(f_L, f_1) & \widehat{\text{cov}}(f_L, f_2) & \cdots & s_{f_L}^2 \end{pmatrix}$$

$$s_{f_l}^2 = \frac{\sum_{n=1}^{N} \left(y_{n,l} - \hat{y}_{n,l}\right)^2}{N - df} \tag{9.32}$$

$$\widehat{\text{cov}}(f_l, f_{l'}) = \frac{\sum_{n=1}^{N} \left(y_{n,l} - \hat{y}_{n,l}\right) \cdot \left(y_{n,l'} - \hat{y}_{n,l'}\right)}{N - df}$$

as illustrated in Section 14.3.3, $y_{n,l}$ and $\hat{y}_{n,l}$ being, respectively, the *n*-th measured and predicted value for the *l*-th output of the model calibration dataset.

It can be easily seen that Equations 9.5 to 9.8 constitute a particular case of Equations 9.30 to 9.32, when $\mathbf{a}_r$ is substituted by $\mathbf{o}_l$, a vector of zeros except for a 'one' in the *l*-th position (i.e. for the *l*-th output variable).

The analytical expressions for the confidence region of the NS and the experimental region corresponding to the *r*-th linear combination of outputs are equivalent to Equations 9.27 and 9.28, substituting e.g. $s_{\hat{y}_{\text{NEW},l}}$, the estimated standard deviation of the *l*-th output, by $s_{\hat{d}_{\text{obs,r}}}$, the estimated standard deviation of the *r*-th quality attribute.

It must also be noted that these expressions can also be extended to confidence regions for NS and experimental regions associated to inputs and linear combinations of inputs, not just outputs. This may be of interest e.g. whenever the uncertainty is to be considered when restrictions are imposed on inputs whose values can be observed but not freely manipulated, or even if this is possible. Consider the example illustrated in Figure 9.1. Figure 9.2 shows the same example, but now the 95% confidence limits associated to the restrictions in Table 9.2 in the latent space, instead of the projection of such restrictions, are represented. In this case all the projected observations fall inside the region delimited by the projection of the 95% confidence limits associated to the univariate inequality constraints. Note that the *lower* confidence limits for 'greater than or equal to'-type restrictions, and *upper* confidence limits for the 'lower than or equal to'-type restrictions are represented, but not their respective upper and lower limits.

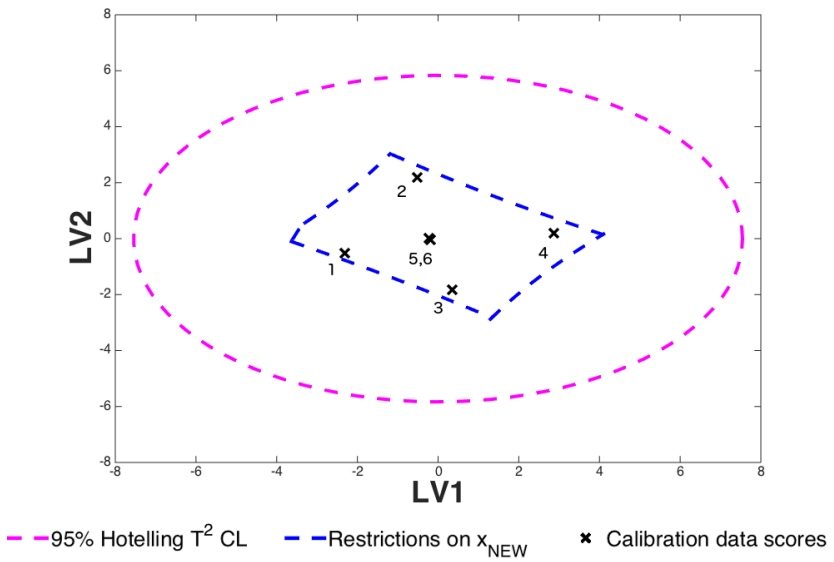


Figure 9.2 Representation of the 95% confidence limits associated to the restrictions in Table 9.2 transferred to the latent space corresponding to a PLS-regression model fitted with 2 LV and the data in Table 9.1, and projection of these data onto the same subspace**.**

9.4. Subspace most likely to contain the True Design Space

Once the restrictions on the original variables have been projected onto the latent space, and considering the confidence limits of the definition of the NS associated to the desired values for the quality attributes of interest, $\mathbf{d}_{\mathrm{DES}}$, the subspace most likely to contain the True Design Space (TDS) will be the portion of the latent space within these limits and inside the Hotelling T^2 confidence hyperellipsoid for a given confidence level. The estimation of this subspace is important because, as the experimental region defined in [87], it can be understood as a subspace within which experimentation can be performed to more easily identify the TDS.

In this section three examples will be illustrated in order to compare this proposed experimental region with that defined in [87] as well as the subspace of acceptable solutions for an optimization problem as defined in [90].

For the first one the hypothetical input-output causal relationship in Equation 8.1 is used to allow a simple graphical visualization of the procedure. This case study will serve as an example to illustrate the application of the proposed approach to define the experimental region, and to compare the results obtained with those from the literature.

A second case study addresses another example with four moderately correlated inputs and two moderately correlated outputs, in which specific values are desired for both

outputs. This helps to illustrate a more complicated scenario in which more than one quality attribute is considered, and that can not be addressed by means of the available methodologies proposed in the literature.

Up until now, in all the case studies discussed, both in this Section and in the literature, the quality attributes of interest coincide with one or more of the outputs used to fit the PLS-regression model, and it is assumed that such model is able to predict them with enough accuracy. As mentioned in Section 8.3, however, this may not be enough in some situations. The third case study serves to illustrate this scenario, and corresponds to the synthesis of 1,2-dichloroethane (EDC) in a vinyl-chloride monomer (VCM) production process plant, simulated by using the software PRO/II and in accordance with the guidelines provided in [101]. In this example, the quality attribute of interest cannot be well explained or predicted by the PLS-regression model, but a linear combination of other process outputs can be used to bypass such limitation. This is, however, a problem that could no be addressed with the methodologies proposed in the literature.

9.4.1. Datasets

9.4.1.1 Case study 1: mathematical model

For the first case study, data were simulated according to the model in Equation 8.1. In this example, x_1 and x_2 are manipulated variables (i.e., those an operator would be able to freely adjust), while x_3, x_4 and x_5 are three measured variables (i.e. those the operator would not be able to freely manipulate, but can measure), and y is the quality attribute of interest. An initial dataset with 6 observations from a D-optimal DOE with 5% of the variability of each variable added as noise to the measured values of x_3, x_4, x_5 and y is used. Table 9.3 shows the mean and standard deviation for such dataset, for all x_i variables, as well as for y. The lower and upper limit imposed on the different x_i are also indicated, which are, in this case, both the levels of factors x_1 and x_2 used to build the D-optimal DOE (for the limits corresponding to these two inputs), and the hard constraints that would be imposed on all variables for optimization. The true relationship of x_3, x_4 and x_5 with x_1 and x_2 is assumed not to be known, as would be the case in most real situations. This is the reason why the lower and upper limits on x_3, x_4 and x_5 are imposed (if the relationship among them was known, these restrictions would be known to be redundant).

This first case study corresponds to an example already addressed in [87], where an estimation of the DS was considered for a given y_{DES} by accounting for the uncertainty in definition of the NS and delimiting the corresponding subspace to the fraction of it within the Hotelling T^2 ellipsoid.

Table 9.3. Case Study 1: Characterization of the input and output calibration datasets

Variable	Mean	Std. dev.	Lower limit	Upper limit
x_1	52.33	41.95	5.43	99.23
x_2	11.76	3.77	7.54	15.97
$x_3(x_1^2)$	4257.99	4480.60	125.64	9893.38
$x_4(x_2^2)$	150.88	88.91	58.51	258.48
$x_5(x_1 \cdot x_2)$	608.81	556.93	50.49	1576.28
y	52.33	41.95	-	-

9.4.1.2 Case study 2: simulated data with two correlated outputs

For the second case study, a multi-normally distributed dataset with 50 observations, with 4 input variables and 2 output variables following a specific correlation structure, was simulated following the procedure explained in [102] and [103] to generate multivariate normal data. The correlation matrix used to generate the dataset can be seen in Table 9.4

Table 9.4. Case Study 2: Correlation matrix used so construct the dataset

	x_1	x_2	x_3	x_4	y_1
x_2	0.45				
x_3	0.54	0.50			
x_4	0.30	0.55	0.70		
y_1	0.50	-0.10	-0.15	0.20	
y_2	-0.30	0.34	0.50	-0.15	-0.70

Table 9.5 shows the mean and standard deviation for the generated dataset, for all x_i and y_i variables.

Table 9.5. Case Study 2: Characterization of the input and output calibration datasets

Variable	Mean	Std. dev.	Lower limit	Upper limit
x_1	0	1	-2.39	2.92
x_2	0	1	-1.91	1.79
x_3	0	1	-1.87	1.87
x_4	0	1	-2.22	1.87
y_1	0	1	-	-
y_2	0	1	-	-

9.4.1.3 Case study 3: simulated Vinyl-Chloride Monomer manufacturing

Data corresponding to the production of ethylene dichloride (EDC) as an intermediate product to VCM were simulated according to the information detailed in [101] in Pro/II. In order to introduce some variability in this process without loss of validity of any assumption made during the simulation, manipulated variables were allowed to vary a maximum of 5% around the assigned values in [101].

Figure 9.3 illustrates a simplified block diagram corresponding to such production process. In this case study, focus is directed towards the purity of EDC obtained after saturation during the EDC purification step (stream EDC4). The top section (blue section in Figure 9.3) corresponds to the direct chlorination section of the plant, to which two stream with pure chlorine and ethylene, respectively, are fed. The bottom section (yellow section in Figure 9.3) corresponds to the oxy-chlorination section, to which air and pure ethylene are fed together with a recycled stream constituted mostly by hydrochloric acid. Both streams are mixed and subjected to saturation (green section in Figure 9.3) as the first step during the EDC purification.

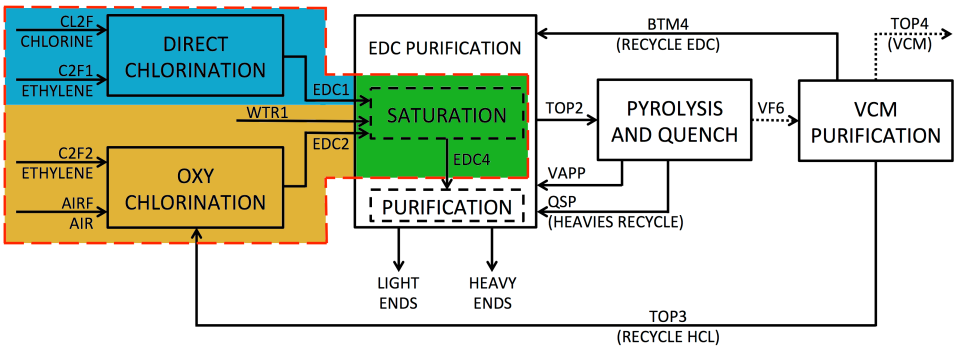


Figure 9.3 Simulated EDC production process as an intermediate product to obtain VCM

Table 9.6 shows the name and the meaning of the variables involved in this case study. Process variables x_1 to x_5 are manipulated input variables, while variables x_6 to x_8 are measured ones; $z_1$is a molar fraction of the predominant reagent in stream TOP3, and y_1 and y_2 are the relevant outputs in this case study. A distinction has been explicitly established between x_i variables and z_1, since the last one is a molar fraction of a stream and must take values between 0 (or 0%) and 1 (or 100%), while no similar constraints tie x_i variables.

Table 9.6. Case Study 3: Definition of the variables involved

Name	Meaning
x_1	Flow (lb-mol/h) of CL2F (Cl2)
x_2	Flow (lb-mol/h) of C2F1 (Ethylene)
x_3	Flow (lb-mol/h) of AIR (O2+N2)
x_4	Flow (lb-mol/h) of C2F2 (Ethylene)
x_5	Flow (lb-mol/h) of WTR1 (Water)
x_6	Temperature (ºC) of TOP3
x_7	Pressure (Psig) of TOP3
x_8	Molar flow (lb-mol/h) of TOP3
z_1	Molar fraction of Hydrochloric Acid (HCl) in TOP3
y_1	Total molar flow (lb-mol/h) of stream EDC4
y_2	Molar rate of EDC in stream EDC4

An initial dataset with 20 simulated samples uniformly spanning the hyperrectangle defined by the upper and lower limits on variables x_1 to x_8 and z_1 is used for this example. Table 9.7 shows the mean and standard deviation for each variable in this dataset. The lower and upper limit constraints imposed on the x_i variables and z_1, which bound the experimental domain, are also indicated. These restrictions are defined in order to simulate the univariate limits usually imposed on the input variables in practice to avoid extrapolation outside of the historical, normal operating conditions.

Table 9.7. Case Study 3: Characterization of the input and output calibration datasets

Variable	Mean	Std. dev.	Lower limit	Upper limit
x_1	142.19	13.36	121.22	161.47
x_2	157.21	16.70	132.45	179.71
x_3	312.24	32.17	278.21	344.08
x_4	118.02	13.99	105.70	136.83
x_5	54.53	4.59	45.73	67.28
x_6	-30.93	0.53	-31.87	-29.91
x_7	135.23	2.55	130.51	141.02
x_8	241.76	24.98	228.50	252.56
z_1	0.99	$4.443 \cdot 10^{-5}$	0.99	0.99
y_1	240.52	19.43	-	-
y_2	0.98	$7.6 \cdot 10^{-4}$	-	-

Additionally, the constraint $x_2 \geq 0.995 \cdot x_1$ is imposed to guarantee that at least 99.5% of chlorine can be converted during the direct chlorination step of the process, so that it meets the requirements for its proper simulation.

9.4.2. Results

9.4.2.1 Case study 1: mathematical model

Figure 9.4 shows the procedure followed in order to define the experimental space given $y_{\mathrm{DES}} = 204.86$ in the first example, by considering all of the aforementioned issues in Sections 9.2 and 9.3. Two LV were chosen to fit the PLS-regression model in order to allow easier graphical representation of the results, although four LV would be a much better option according to the leave-one-out cross validation method [104]

(Q^2(2LV)=0.336; Q^2(4LV)=0.955). The proposed methodology, however, allows the definition of the experimental region independently of the dimensionality of the latent space.

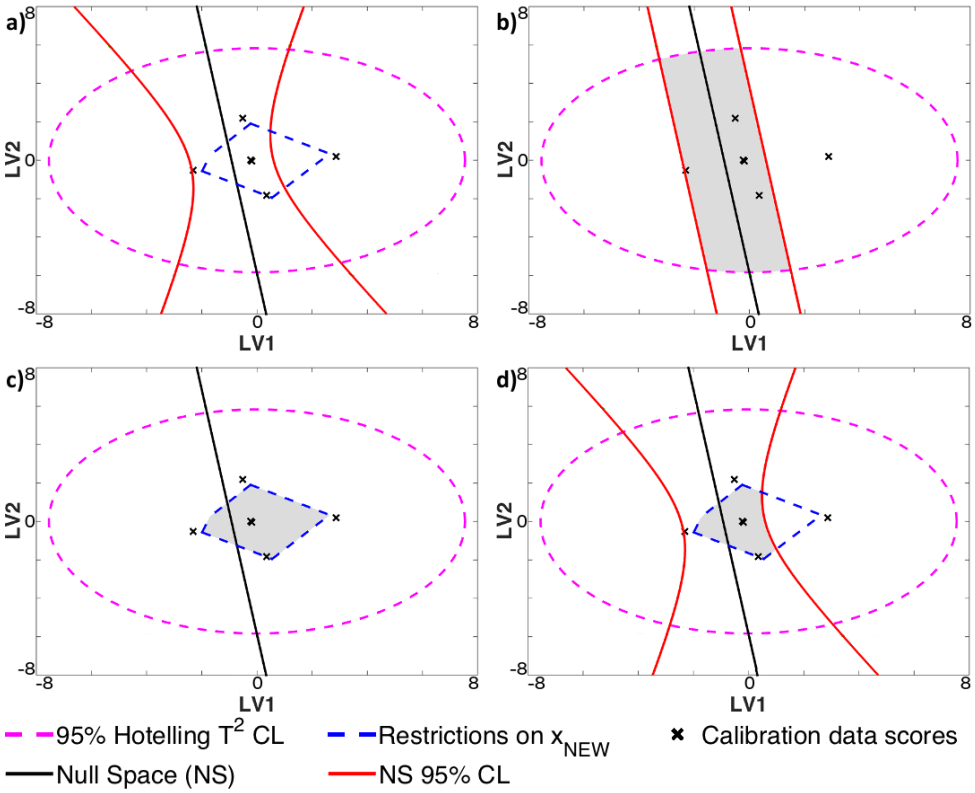


Figure 9.4 For the first case study and $\boldsymbol{y_{DES} = 204.86}$: a) Representation of all restrictions in the latent space; and delimitation of the experimental space (grey area) b) by bracketing the DS as in [87], c) as in the optimization approach in [90], and d) with the proposed approach.

In Figure 9.4a, the Hotelling T^2 95% confidence ellipses when a PLS-regression model with 2 LV was fitted, the initial dataset being as described in Section 9.4.1.1. To do this, Equation 9.3 is used to calculate the Hotelling T^2. The projections of the 6 samples (two of them superimposed close to the centre) and the NS (black line) for $y_{\mathrm{DES}} = 204.86$ on the model subspace and its confidence limits (red lines) defined as in Equation 9.27 are also represented, as well as the restrictions imposed on the inputs (lower and upper limits as defined in Table 1) when transferred onto the latent space through Equation 9.11 (dashed blue lines). These restrictions are included in order to avoid solutions outside of the subspace defined by the range of variability of the inputs in the past, which if accepted as an estimation of the knowledge space does not coin-

cide, in this case, with the interior of the Hotelling T^2 confidence ellipse, but only a fraction of it. Note that, although ten restrictions are originally applied to the inputs (five lower bounds and five upper bounds) only half of them remain in the latent space (five dashed blue lines) because, due to the existing correlation among the inputs, the other half are redundant in the 2-dimensional latent space.

It should also be noted that, because of the uncertainty associated to the PLS regression model, the projection of some of the samples in the initial dataset fall outside of the area within the constraints on the inputs (dashed blue lines), as opposed to what would have been intuitively expected. More specifically, the two samples up and left fall outside of the boundaries defined by the restrictions on the inputs because of the lack of representativeness of the PLS model fitted with 2 LV, which is unable to explain the non-linear relationship among the two independent inputs (x_1 and x_2) and the observable inputs (x_3, x_4 and x_5). If the restrictions x_3, x_4 and x_5 are not applied, or are defined by taking into account their relationship with x_1 and x_2, the projection of both observations would be within boundaries. The same would occur if 4 LV were used to fit the model, but the graphical representation would become more complex and less intuitive. On the other hand, the sample on the right side is outside of the boundaries as a consequence of being the worst predicted sample because of the amount of noise introduced.

The grey area in Figure 9.4b corresponds to the delimitation of the DS as proposed in [87] (although no analytical expressions for the red lines were provided nor were they used as hard constraints), while the shadowed area in Figure 9.4c is the experimental space as would be defined in the optimization approach described in [90] (although in this optimization approach no analytical expressions where provided for the restrictions in the latent space, either, they would be implicitly used as hard constraints). Lastly, Figure 9.4d shows the experimental space as would be obtained by applying the approach proposed in this section.

It can be concluded that, in this example, the delimitation proposed leads to a smaller experimental space than with the approaches in [87] or [90]. Although a small subset of input combinations not inside the bracketed DS according to [87] could be accepted now, due to the curvature of the confidence limits as defined here, a larger subset of input combinations inside the experimental space according to [87] are excluded for not actually being a part of the KS (i.e. outside of the boundaries defined by the lower and upper limits in Table 9.3). On the other hand, a subset of input combinations acceptable according to [90] are excluded because they fall outside of the 95% confidence region of the NS for $y_{\mathrm{DES}} = 204.86$.

9.4.2.2 Case study 2: simulated data with two correlated outputs

This second case study is peculiar in that two quality attributes are considered simultaneously, as opposed to the approach in [87]. Regarding the optimization, the only dif-

ference with the approach in [90] is in the definition of the confidence regions for the two NS and the use of their limits as hard constraints.

Figure 9.5 shows the same procedure followed in the previous case study, but now applied to this example, with $\mathbf{y}_{\mathrm{DES}} = [1.2375; -0.4024]$. To differentiate the two output variables considered, the NS and corresponding confidence limits for $y_{1,\mathrm{DES}}$ are represented with continuous lines, and dashed lines are used for $y_{2,\mathrm{DES}}$. Two LV were chosen to fit the PLS-regression model in order to allow easier graphical representation of the results, although four LV would have been better option to explain more of the variability of $\mathbf{X}$ ($R^2(\mathbf{X},2\mathrm{LV})=0.642$; $R^2(\mathbf{X},4\mathrm{LV})=0.87$), and to better predict Y according to the leave-one-out cross validation method ($Q^2(2\mathrm{LV})=0.858$; $Q^2(4\mathrm{LV})=0.997$). Again, the proposed methodology could be applied with a higher-dimensional latent space, although visualizing the experimental space becomes more difficult in such case.

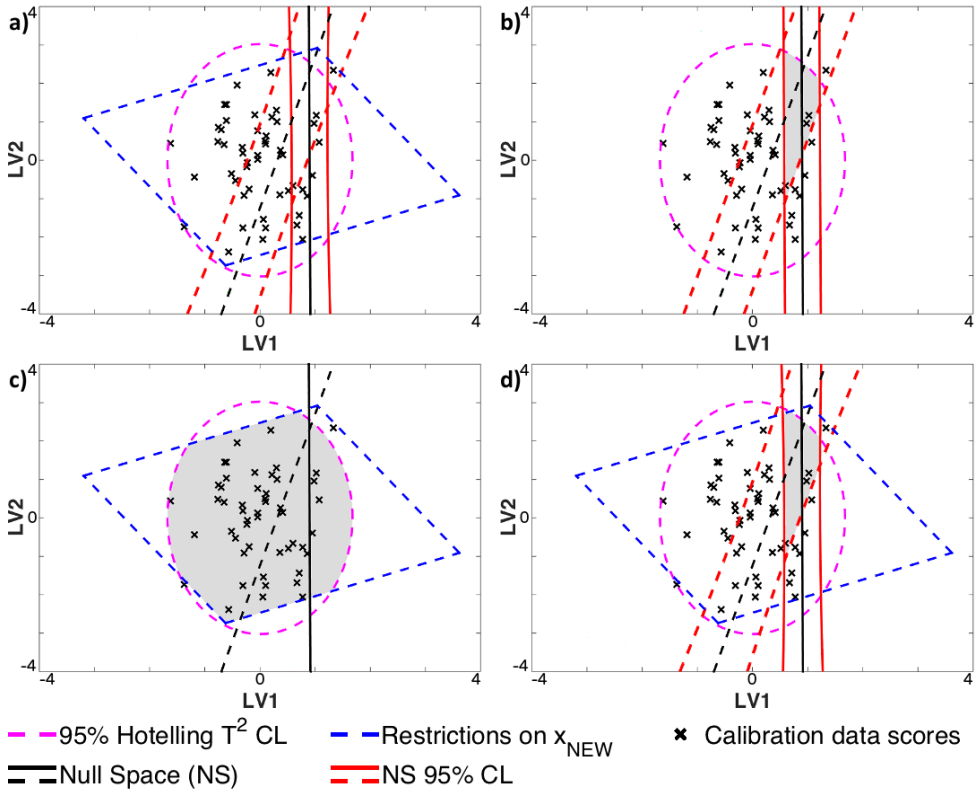


Figure 9.5 For the second case study and $\mathbf{y}_{\mathbf{DES}} = [1.2375, -0.4024]$: a) Representation of all restrictions in the latent space; and delimitation of the experimental space b) by bracketing the DS as in [87], c) as in the optimization approach in [90], and d) with the proposed approach.

Figure 9.5a is the equivalent to Figure 9.4a for this second case study. In this case, only the projection of one of the observations falls outside of the area defined by the restrictions on the inputs. Furthermore, assuming the Hotelling T^2 ellipse to be a good estimation of the knowledge space would have been a much better approximation than in the first case study, although a small fraction of it would still be left out by the restrictions on the inputs. It must be noted that, again, only half of the eight restrictions imposed on the inputs remain in the latent space, with the other four ones being redundant in the latent space due to the existing correlation structure among them.

The shadowed area in Figure 9.5b corresponds to the intersection between the corresponding delimitations of the DS for $y_{1,\mathrm{DES}}$ and $y_{2,\mathrm{DES}}$ that would have been obtained by applying a slight variation of the methodology in [87], using as leverages for the corresponding NS those of the scores with minimum leverage on each NS instead of the leverage of the direct inversion (i.e. the intersection of both NS) to determine the amplitude of the CI in the fourth step of that methodology (see Section 9.3.4). The intersection of both subspaces is considered to define the experimental space, and not the union, because only combinations of inputs leading to the desired values for both outputs simultaneously would be accepted. It must be noted that in this case such intersection exists and at least some part of it is located within the subspace defined by the Hotelling T^2 ellipse and all the other restrictions on the inputs, but this may not be the case in more complex examples or if different values for $y_{1,\mathrm{DES}}$ and $y_{2,\mathrm{DES}}$ are specified. Furthermore, by defining the intersection this way, the correlation between y_1 and y_2 is disregarded, which implicitly overestimates the size of the subspace most likely to provide the desired values for both outputs simultaneously.

The shadowed area in Figure 9.5c is the experimental space as would be defined in the optimization approach described in [90]. Lastly, Figure 9.5d shows the experimental space as would be obtained by applying the approach proposed in this thesis, which happens to be almost identical to what would have been obtained if the slight variation of the methodology in [87] (only applied when a single output variable was considered) shown in Figure 9.5b had been applied. This similarity decreases if fewer samples are used to fit the PLS-regression model (resulting in a more pronounced curvature of the limits of the NS confidence regions when using the proposed approach) and the leverage of the direct inversion solution is used for the estimation of the width of the NS confidence region with the methodology in [87].

9.4.2.3 Case study 3: simulated Vinyl-Chloride Monomer manufacturing

This third case study is interesting because neither the methodology to bracket the DS in [87] nor the optimization formulation in [90] allow approaching a problem like this.

In this example, previous to the application of the proposed methodology, the initial fitting of the corresponding models relating input variables x_1 to x_8 and z_1 with the output variables y_1 and y_2 via PLS led to models with relatively poor explanatory and

predictive capability with regards to y_2 (R^2(**X**)<0.95, $R^2(y_2)$<0.46 and $Q^2(y_2)$<<0.01 even with 5 LV), which is the most relevant quality attribute of interest in this example.

However, if z_1, which represent the relative amount of a chemical compounds in a stream, is transformed into a new variable $x_9 = z_1 \cdot x_8$, and if y_2 is substituted by $y_3 = y_1 \cdot y_2$, the PLS regression models fitted with these 'new' variables provide much better explanatory and predictive capabilities (for example, R^2(**Y**)>0.953 and R^2(**X**)>0.563 when using two LVs, and R^2(**Y**)>0.965 and R^2(**X**)>0.863 when using five LVs).

It should be noted that, although the quality attribute of interest in this case is y_2, the model fitted does not include it. Instead, y_1 and y_3 are used. However, since $y_3 = y_1 \cdot y_2$, then it can also be said that $y_2 = y_3/y_1$, and given that $y_1 > 0$, if a desired value $y_{2,\mathrm{DES}}$ for y_2 is specified, then any combination of values for the input variables in the DS shall satisfy the next equality:

$$y_{2,\mathrm{DES}} \cdot y_1 - y_3 = 0 \tag{9.33}$$

As in previous example, although five LV would provide a better fit, and the methodology proposed in this paper could be applied nonetheless, a PLS-regression model with only two LV is fitted here in order to better illustrate the results achieved. Figure 9.6 shows the NS associated to Equation 9.33 for $y_{2,\mathrm{DES}} = 0.9792$, and the model fitted with data as described in Section 9.4.1.3, together with the corresponding confidence interval.

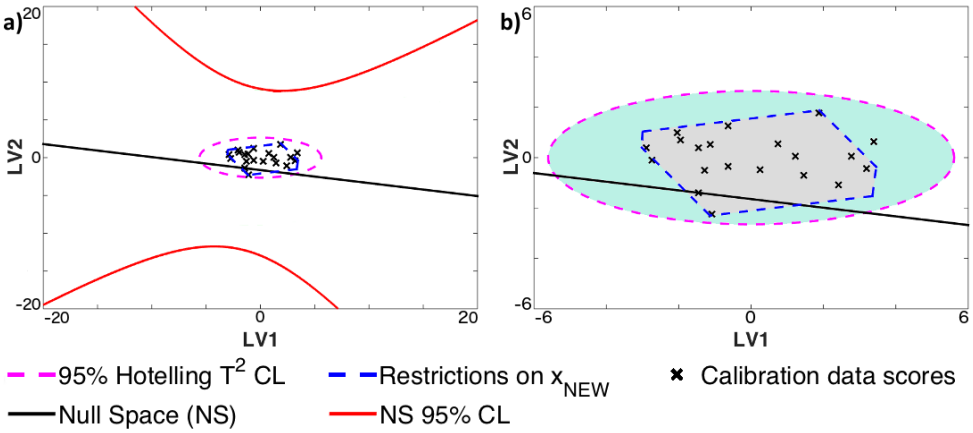


Figure 9.6 For the third case study and $y_{2,\mathrm{DES}} = 0.9792$: a) Representation of all restrictions in the latent space; and b) delimitation of the experimental space by bracketing the DS as in [87] (green + grey areas), and as in the optimization approach in [90] and with the proposed approach (grey area).

In Figure 9.6a the Hotelling T^2 confidence ellipse when a PLS-regression model with 2 LV was fitted, the initial dataset being as described in Section 9.4.1.3. The 95% confidence limits for the NS and the restrictions on the inputs transferred onto the latent

space are also represented. It can be seen that in this case, due to the uncertainty in the DS estimate, the hard restrictions associated to the NS confidence limits are redundant, as is the Hotelling T^2 hard constraint.

Figure 9.6b represents a zoomed in version of Figure 9.6a to better visualize the experimental space. In this case, the experimental space resulting from applying the methodology in [87] would include the whole interior of the Hotelling T^2 (green and grey area), while both the optimization approach in [90] and the proposed approach would further restrict the experimental space (grey area). Therefore, the procedure presented in this section again guarantees the more robust delimitation of the experimental space, by simultaneously accounting for the restrictions on the inputs and a variation of the bracketing of the DS as suggested in [87].

9.4.3. Conclusions

The proposed methodology for the definition of the subspace of the latent space most likely to contain the TDS combines the already proposed constraints in the literature for optimization [90] with the idea of bracketing the DS [87]. This procedure has been shown to provide robust results in the delimitation of the experimental region, in the sense that this subspace is shown to be at least as reduced as the one resulting from the most restrictive approach between those in [90] and [87], as illustrated in the second and third case studies, if not more restricted, as seen in the first case study. This is achieved by having developed a new way to analytically define the NS, as well as the limits of its confidence region. This new definition presents three main advantages:

1. It allows the use of the limits of the confidence region for each NS as hard constraints in the definition of the experimental region, which was not possible in the optimization problem defined in [90].
2. The non-linearity of the limits of the confidence region of each NS, due to the effect of the leverage on the prediction uncertainty, can also be accounted for in their analytical expression, which could not be explicitly considered in the bracketing of the DS proposed in[87] .
3. An extension of the definition of the NS for quality attributes defined as linear combinations of outputs is made possible. This is useful to approach problems that could not be addressed otherwise, as illustrated in the third case study presented in this paper.

In a similar way, the procedure to transfer restrictions on the inputs and outputs to the latent space has been proposed. This allows comparing restrictions that are not directly comparable when defined as affecting the inputs or the outputs, both among themselves as well as with constraints defined in the latent space, such as those regarding the Hotelling T^2. As an example, the hard constraint on the Hotelling T^2 in the first and third case studies is seen to be redundant (i.e. less restrictive) with respect to the constraints on the inputs. This assessment is not possible without transferring the restrictions on the inputs to the latent space.

It must be noted that, although not illustrated in this section, the extension of this methodology for the DS associated to a range of desired values, instead of specific ones, for one or more quality attributes of interest, is straightforward (as detailed in Section 9.3.4).

9.5. Subspace least likely to fall outside of the True Design Space

In Section 9.4 the procedure to define the subspace most likely to contain the TDS is outlined. Such strategy may be useful for experimental purposes, when the interest is mainly not discarding any set of processing conditions that may be a part of the TDS. A more common desire is to guarantee that the specifications demanded by the customer are met, or at least that they are not met as few times as possible. This means, in practice, that defining the subspace least likely to fall outside of the TDS is at least as important, if not more, as defining the subspace most likely to contain it.

In order to define this subspace, and for the more general case of a quality attribute being a linear combination of the output variables included in the fitted PLS-regression model, consider again Equations 9.30 to 9.32 for the $100 \cdot (1-\alpha)\%$ CI of the prediction of the *r*-th quality attribute $d_{\text{obs},r}$ given an observation $\mathbf{x}_{\text{obs}}$. Consider also a range of acceptable values for the *r*-th quality attribute of interest, such that two specification limits can be defined, the Lower Specification Limit (LSL), $d_{\text{LSL},r}$, and the Upper Specification Limit, $d_{\text{USL},r}$, such that a product will be considered to be within specifications regarding this quality attribute if its observed value, $d_{\text{obs},r}$, meets that $d_{\text{LSL},r} \leq d_{\text{obs},r} \leq d_{\text{USL},r}$.

Let $\tilde{\boldsymbol{\tau}}_{\text{LSL},r}$ be a vector of scores such that the lower limit for the $100 \cdot (1-\alpha)\%$ confidence interval (CI) of the prediction of a linear combination $\tilde{d}_{\text{LSL},r}$ for an observation $\tilde{\mathbf{x}}_{\text{LSL},r}$, with projection $\tilde{\boldsymbol{\tau}}_{\text{LSL},r}$, is coincident with the LSL for the *r*-th quality attribute; and let $\tilde{\boldsymbol{\tau}}_{\text{USL},r}$ be a vector of scores such that the upper limit for the $100 \cdot (1-\alpha)\%$ confidence interval (CI) of the prediction of a linear combination $\tilde{d}_{\text{USL},r}$ for an observation $\tilde{\mathbf{x}}_{\text{USL},r}$, with projection $\tilde{\boldsymbol{\tau}}_{\text{USL},r}$, is coincident with the USL for the *r*-th quality attribute.

To put it another way, consider $\mathbf{x}_{\text{LSL},r}$ to be an observation such that the prediction of the linear combination of outputs d_r, $\hat{d}_r = d_{\text{LSL},r}$, and $\boldsymbol{\tau}_{\text{LSL},r}$ to be the projection of $\mathbf{x}_{\text{LSL},r}$ onto the latent subspace. $\tilde{\boldsymbol{\tau}}_{\text{LSL},r}$ is the set of scores closest to $\boldsymbol{\tau}_{\text{LSL},r}$ such that the lower limit for the $100 \cdot (1-\alpha)\%$ CI of the prediction of d_r is equal to $d_{\text{LSL},r}$. Equivalently, if $\boldsymbol{\tau}_{\text{USL},r}$ is the projection of a vector $\mathbf{x}_{\text{USL},r}$ such that $\hat{d}_r = d_{\text{USL},r}$, then $\tilde{\boldsymbol{\tau}}_{\text{USL},r}$ is the set of scores closest to $\boldsymbol{\tau}_{\text{USL},r}$ such that the upper limit for the $100 \cdot (1-\alpha)\%$ CI of the prediction of d_r is equal to $d_{\text{USL},r}$.

Using OLS type expressions:

$$\begin{aligned} d_{\mathrm{LSL},r} &= \tilde{d}_{\mathrm{LSL},r} - t_{N-A,\alpha/2} \cdot s_{\tilde{d}_{\mathrm{LSL},r}} \\ d_{\mathrm{USL},r} &= \tilde{d}_{\mathrm{USL},r} + t_{N-A,\alpha/2} \cdot s_{\tilde{d}_{\mathrm{USL},r}} \end{aligned} \tag{9.34}$$

By taking into account Equations 9.30 to 9.32, Equation 9.34 can also be expressed as:

$$\begin{aligned} \mathbf{a}_r^{\mathrm{T}}\left(\mathbf{D}_{\mathbf{S_Y}}\mathbf{Q}\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r} + \mathbf{m}_{\mathbf{Y}}\right) - t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}\sqrt{1 + \tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r} + \frac{1}{N}} - d_{\mathrm{LSL},r} = 0 \\ \mathbf{a}_r^{\mathrm{T}}\left(\mathbf{D}_{\mathbf{S_Y}}\mathbf{Q}\tilde{\boldsymbol{\tau}}_{\mathrm{USL},r} + \mathbf{m}_{\mathbf{Y}}\right) + t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}\sqrt{1 + \tilde{\boldsymbol{\tau}}_{\mathrm{USL},r}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\tilde{\boldsymbol{\tau}}_{\mathrm{USL},r} + \frac{1}{N}} - d_{\mathrm{USL},r} = 0 \end{aligned} \tag{9.35}$$

Which provides the expressions corresponding to the limits of the region within which any set of scores is associated to a set of inputs theoretically guaranteeing that at least $100 \cdot (1-\alpha)\%$ of the time the value for the corresponding quality attribute of interest will be within the specifications limits. Consequently, any set of scores $\boldsymbol{\tau}_{\mathrm{ws},r}$ for a product within specifications with respect to the *r*-th quality attribute will meet that:

$$\begin{aligned} \mathbf{a}_r^{\mathrm{T}}\left(\mathbf{D}_{\mathbf{S_Y}}\mathbf{Q}\boldsymbol{\tau}_{\mathrm{ws},r} + \mathbf{m}_{\mathbf{Y}}\right) - t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}\sqrt{1 + \boldsymbol{\tau}_{\mathrm{ws},r}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\boldsymbol{\tau}_{\mathrm{ws},r} + \frac{1}{N}} - d_{\mathrm{LSL},r} \geq 0 \\ \mathbf{a}_r^{\mathrm{T}}\left(\mathbf{D}_{\mathbf{S_Y}}\mathbf{Q}\boldsymbol{\tau}_{\mathrm{ws},r} + \mathbf{m}_{\mathbf{Y}}\right) + t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}\sqrt{1 + \boldsymbol{\tau}_{\mathrm{ws},r}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\boldsymbol{\tau}_{\mathrm{ws},r} + \frac{1}{N}} - d_{\mathrm{USL},r} \leq 0 \end{aligned} \tag{9.36}$$

For, e.g., graphical representation purposes, $\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}$ and $\tilde{\boldsymbol{\tau}}_{\mathrm{USL},r}$ can be also obtained given the closest points $\boldsymbol{\tau}_{\mathrm{LSL},r}$ and $\boldsymbol{\tau}_{\mathrm{USL},r}$ located in the corresponding NS for the *r*-th LSL and USL, respectively. Consider, for $\boldsymbol{\tau}_{\mathrm{LSL},r}$**,** the equation for the NS associated to the *r*-th LSL, $\mathrm{NS}_{\mathrm{LSL},r}$which can be expressed as:

$$\begin{aligned} v_{\mathrm{LSL,r},0} + \sum_{a=1}^{\mathrm{A}} v_{r,a} \cdot \tau_{\mathrm{LSL,r},a} = 0 \,;\, \mathbf{v}_r = [v_{\mathrm{r},1}, v_{\mathrm{r},2}, \ldots, v_{r,A}]^{\mathrm{T}} \\ v_{\mathrm{LSL,r},0} = \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{m}_{\mathbf{Y}} - d_{\mathrm{LSL,r}} \\ \mathbf{v}_r = \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}} \cdot \mathbf{a}_r \end{aligned} \tag{9.37}$$

The relationship between $\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}$ and $\boldsymbol{\tau}_{\mathrm{LSL},r}$ can be expressed as:

$$\begin{aligned} \tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r} = \boldsymbol{\tau}_{\mathrm{LSL},r} + dd\left(\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right) \cdot \mathbf{u}_{\mathrm{NS}_r} \\ dd\left(\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right) = \frac{v_{\mathrm{LSL,r},0}}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} + \mathbf{u}_{\mathrm{NS}_r}^{\mathrm{T}} \cdot \tilde{\boldsymbol{\tau}}_{\mathrm{LSL,r}} \quad ; \quad \mathbf{u}_{\mathrm{NS}_r} = \frac{\mathbf{v}_r}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} \end{aligned} \tag{9.38}$$

where $\mathbf{u}_{\mathrm{NS}_r}$ is the unitary vector orthogonal to $\mathrm{NS}_{\mathrm{LSL},r}$ (or $\mathrm{NS}_{\mathrm{USL},r}$) and $dd\left(\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right)$ is the directed distance, which has the same sign and magnitude as

the Euclidean distance when the direction of the vector from $\mathrm{NS}_{\mathrm{LSL},r}$ to $\tilde{\boldsymbol{\tau}}_{\mathrm{LSL,r}}$ is the same as that of $\mathbf{u}_{\mathrm{NS}_r}$, but is negative if they have opposing directions.

The first equality in Equation 9.35 can then be written:

$$dd\left(\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right) = \frac{t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}}{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{u}_{\mathrm{NS}_r}} \cdot \sqrt{1 + h_{\tilde{\boldsymbol{\tau}}_{\mathrm{LSL,r}}} + \frac{1}{N}} \qquad (9.39)$$
$$h_{\tilde{\boldsymbol{\tau}}_{\mathrm{LSL,r}}} = \tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}{}^{\mathrm{T}} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}$$

This equation can be solved recursively, such that $dd\left(\tilde{\boldsymbol{\tau}}_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right)$ is obtained given $\boldsymbol{\tau}_{\mathrm{LSL}}$, and from them both $\tilde{\boldsymbol{\tau}}_{\mathrm{LSL}}$ is computed directly following Equation 9.38. The procedure to obtain $\tilde{\boldsymbol{\tau}}_{\mathrm{USL}}$ is almost identical.

It is important to note that the existence of the subspace defined this way (i.e. at least one $\boldsymbol{\tau}_{\mathrm{ws},r}$ exists that meets both restrictions in Equation 9.36) will depend on the difference between the *r*-th USL and LSL, and the uncertainty of the model. To assess this, consider the points with lowest leverages from the curves defined by Equation 9.35, denoted as $\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}$ and $\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}$, respectively. These sets of scores will be related to the corresponding sets of scores in the NS for the *r*-th LSL and USL with lowest leverages, $\boldsymbol{\tau}^*_{\mathrm{LSL},r}$ and $\boldsymbol{\tau}^*_{\mathrm{USL},r}$. To obtain $\boldsymbol{\tau}^*_{\mathrm{LSL},r}$ and $\boldsymbol{\tau}^*_{\mathrm{USL},r}$, the equations for the corresponding NS as a function of the standardized scores is required.

For $\boldsymbol{\tau}^*_{\mathrm{LSL}}$, Equation 9.37 can be reformulated as:

$$\begin{aligned} v_{\mathrm{LSL},r,0} + \mathbf{v}_r^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{1/2} \cdot \boldsymbol{\tau}^{\mathrm{o}}_{\mathrm{LSL,r}} &= 0 \\ v_{\mathrm{LSL},r,0} &= \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{m}_{\mathbf{Y}} - d_{\mathrm{LSL},r} \\ \mathbf{v}_r &= \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}} \cdot \mathbf{a}_r \end{aligned} \qquad (9.40)$$

where $\boldsymbol{\tau}_{\mathrm{LSL},r} = \boldsymbol{\Lambda}^{1/2} \cdot \boldsymbol{\tau}^{\mathrm{o}}_{\mathrm{LSL},r}$, $\boldsymbol{\Lambda}^{1/2}$ being the [A×A] diagonal matrix containing the A standard deviations of the scores associated to the LVs in the calibration dataset. $\boldsymbol{\tau}^{\mathrm{o},*}_{\mathrm{LSL},r}$, the set of standardized scores closest to the centre of projection (and with lowest leverages), can be demonstrated to be:

$$\boldsymbol{\tau}^{\mathrm{o},*}_{\mathrm{LSL},r} = -\boldsymbol{\Lambda}^{1/2} \cdot \mathbf{v}_r \cdot (\mathbf{v}_r^{\mathrm{T}} \cdot \boldsymbol{\Lambda} \cdot \mathbf{v}_r)^{-1} \cdot v_{\mathrm{LSL},r,0} \qquad (9.41)$$

Therefore:

$$\begin{aligned} \boldsymbol{\tau}^*_{\mathrm{LSL},r} &= -\boldsymbol{\Lambda} \cdot \mathbf{v}_r \cdot (\mathbf{v}_r^{\mathrm{T}} \cdot \boldsymbol{\Lambda} \cdot \mathbf{v}_r)^{-1} \cdot v_{\mathrm{LSL},r,0} \\ \boldsymbol{\tau}^*_{\mathrm{USL},r} &= -\boldsymbol{\Lambda} \cdot \mathbf{v}_r \cdot (\mathbf{v}_r^{\mathrm{T}} \cdot \boldsymbol{\Lambda} \cdot \mathbf{v}_r)^{-1} \cdot v_{\mathrm{USL},r,0} \\ v_{\mathrm{LSL},r,0} &= \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{m}_{\mathbf{Y}} - d_{\mathrm{LSL},r} \\ v_{\mathrm{USL},r,0} &= \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{m}_{\mathbf{Y}} - d_{\mathrm{USL},r} \\ \mathbf{v}_r &= \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}} \cdot \mathbf{a}_r \end{aligned} \qquad (9.42)$$

$\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}$ and $\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}$ can then be obtained through the aforementioned recursive approach, via Equations 9.38 and 9.39. For the subspace to exist, the following equality must be met:

$$\mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}} \cdot \mathbf{Q} \cdot (\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL}} - \tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL}}) \geq 0 \tag{9.43}$$

If Equation 9.38 and 9.39 are taken into account, Equation 9.43 can be reformulated as:

$$\begin{gathered} d_{\mathrm{USL},r} - d_{\mathrm{LSL},r} \geq \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}} \cdot \mathbf{Q} \cdot \left(dd\left(\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right) - dd\left(\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}, \mathrm{NS}_{\mathrm{USL},r}\right) \right) \cdot \mathbf{u}_{\mathrm{NS}_r} \\ dd\left(\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}, \mathrm{NS}_{\mathrm{LSL},r}\right) = \frac{t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}}{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{u}_{\mathrm{NS}_r}} \cdot \sqrt{1 + h_{\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}} + \frac{1}{N}} = 0 \\ dd\left(\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}, \mathrm{NS}_{\mathrm{USL},r}\right) = \frac{t_{N-A,\frac{\alpha}{2}} \cdot SE_{d_r}}{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{u}_{\mathrm{NS}_r}} \cdot \sqrt{1 + h_{\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}} + \frac{1}{N}} = 0 \end{gathered} \tag{9.44}$$

Therefore:

$$d_{\mathrm{USL},r} - d_{\mathrm{LSL},r} \geq SE_{d_r} \cdot t_{N-A,\frac{\alpha}{2}} \cdot \left(\sqrt{1 + h_{\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}} + \frac{1}{N}} - \sqrt{1 + h_{\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}} + \frac{1}{N}} \right) \tag{9.45}$$

where $h_{\tilde{\boldsymbol{\tau}}^*_{\mathrm{LSL},r}}$ and $h_{\tilde{\boldsymbol{\tau}}^*_{\mathrm{USL},r}}$ can be obtained through the aforementioned recursive approach, via Equation 9.38 and 9.39, given $\boldsymbol{\tau}^*_{\mathrm{LSL}}$ and $\boldsymbol{\tau}^*_{\mathrm{USL}}$, which depend exclusively on d_{LSL} and d_{USL}.

The subspace least likely to fall outside of the TDS will exist only if Equation 9.45 can be met. This means that, for example, this subspace does not exist if a single value is specified for a given quality attribute, instead of a range of values, since in such a case $d_{\mathrm{USL},r} - d_{\mathrm{LSL},r} = 0$, while $SE_{d_r} > 0$ already in most cases, due to the uncertainty associated to the model.

9.6. Assessing the adequacy of a PLS-regression model for inversion

While Sections 9.4 and 9.5 provide tools to define subspaces most appropriate for experimentation and for guaranteeing the quality of the products, respectively, no *a priori* assessment of the PLS-regression model as more or less adequate for its inversion was made. In e.g. [90] assessing the performance of the model for the prediction of the input variables from the scores, and extracting the required number of LV to properly predict both **X** and **Y** is suggested, but this still does not actually provide a measure of the performance of the model used in its inverse form. In this section two different approaches to perform this evaluation are discussed, and a modification of the second proposed.

9.6.1. Assessment via direct inversion

Similarly to R^2 and Q^2, commonly used as a measure to quantify the explanatory or predictive capability of a model used in its direct form (either for **X** or **Y**), $\tilde{R}^2_{\mathbf{X}}$ and $\tilde{Q}^2_{\mathbf{X}}$ are defined here to refer to the explanatory or predictive capability of the PLS-regression model used in its inverse form to obtain the sets of inputs theoretically leading to the outputs for observations already used to fit this model and for observations not used to fit it, respectively.

Intuitively, it may seem that a way to estimate $\tilde{R}^2_{\mathbf{X}}$ and $\tilde{Q}^2_{\mathbf{X}}$ could be via cross-validation, as done for R^2 and Q^2, by 'predicting' each of the observations in the input space through direct inversion, given the corresponding values for the outputs for those observations, and comparing the 'predicted' values with the 'real' ones. However, this procedure is only valid as long as each combination of values for the output variables can only be achieved with a given combination of values for the inputs. This is because otherwise, as explained in Section 9.3, the direct inversion provides a single set of inputs among all possible combinations of them theoretically providing certain outputs.

9.6.2. Assessment by comparison with the closest solution in the null-space

Considering the issue mentioned in the previous section, Jaeckle & MacGregor [88] already proposed an approach to obtain $\tilde{R}^2_{\mathbf{X}}$ and $\tilde{Q}^2_{\mathbf{X}}$. This approach considered the existence of the NS by comparing the projection of each observation (in the calibration or validation dataset) not with the result from the direct inversion, but with the set of scores in the corresponding (combined pseudo) NS closest to this projection. However, this approach presents some limitations or drawbacks, because:

i) it is based on a definition of the NS (the so-called combined pseudo NS in Section 9.3.2) that makes it impossible to assess how appropriate a PLS-regression model may be for inversion when information of only one, or some, but not all of the outputs is available (i.e. there are missing data) or relevant (i.e. not all of the outputs available in the dataset are of interest at a certain moment);
ii) since a priori a selective PCA on **Y** has been performed before obtaining the NS, the information of the outputs discarded as a consequence of the selective PCA is missing; and
iii) it cannot be applied to quality attributes defined as a linear combination of outputs. In this section, a new way to compute $\tilde{R}^2$ and $\tilde{Q}^2$ is provided which aims at overcoming these limitations.

The procedure proposed by Jaeckle & MacGregor [88], however, can be slightly modified by accounting for the analytitcal definition of the NS as presented in Section 9.3.3, both for the outputs and for any quality attribute expressed as a linear combination of them. To do this, consider the set of equations for the R NS associated to the values of the R quality attributes of interest for the n-th observation in a dataset:

$$\mathbf{v}_{n,0} + \mathbf{V} \cdot \boldsymbol{\tau}_{\mathrm{NS}_n} = \mathbf{0}$$

$$\mathbf{v}_{n,0} = \begin{bmatrix} v_{n,1,0} \\ v_{n,2,0} \\ \vdots \\ v_{n,R,0} \end{bmatrix} = \mathbf{A}_{\mathrm{r}} \cdot \mathbf{m}_{\mathbf{Y}} - \mathbf{d}_n \quad ; \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_R^{\mathrm{T}} \end{bmatrix} = \mathbf{A}_{\mathrm{r}} \cdot \mathbf{D}_{\mathbf{S}_{\mathbf{Y}}} \cdot \mathbf{Q} \tag{9.46}$$

$\mathbf{A}_{\mathrm{r}}$ and $\mathbf{V}$ being defined as in Equation 9.26, and $\mathbf{d}_n$being the [R×1] column vector with the observed (or calculated) value for the r-th quality attribute of interest, for the n-th observation, as its r-th element. As long as the number of latent variables used to fit the PLS model is greater than the number of quality attributes considered (i.e. $A > R$) a subspace may exist corresponding to more than one set of inputs $\hat{\mathbf{x}}_{\mathrm{NS}_n}$, and therefore a set of scores $\boldsymbol{\tau}_{\mathrm{NS}_n}$, that theoretically guarantee $\mathbf{d}_n$.

From all such scores, Jaeckle & MacGregor [88] wisely suggest computing the one closest to $\boldsymbol{\tau}_n$, $\tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n}$, and comparing them with each other. In their approach, the 'combined pseudo-NS' is considered, and because of the way it is computed it is expected that such subspace will not necessarily coincide with the intersection of the R hyperplanes whose equations define the elements of $\mathbf{v}_{n,0}$ and the rows of $\mathbf{V}$ in Equation 9.46. In this thesis it can be demonstrated that:

$$\tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n} = [\mathbf{I}_A - \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{V}] \cdot \boldsymbol{\tau}_n - \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{v}_{n,0} \tag{9.47}$$

And therefore:

$$\boldsymbol{\tau}_n - \tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n} = \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{V} \cdot \boldsymbol{\tau}_n + \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{v}_{n,0} \tag{9.48}$$

As in [88], $\tilde{R}^2_{\mathbf{X}}$ and $\tilde{Q}^2_{\mathbf{X}}$ are defined as:

$$\begin{aligned} \tilde{R}^2_{\mathbf{X}} &= 1 - \frac{\sum_{n=1}^{N^*} \mathrm{SS}\left[\boldsymbol{\tau}_n - \tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n}\right]}{\sum_{n=1}^{N^*} \mathrm{SS}[\boldsymbol{\tau}_n - \mathbf{m}_{\mathbf{T}}]} \\ \tilde{Q}^2_{\mathbf{X}} &= 1 - \frac{\sum_{n=1}^{N^*} \mathrm{SS}\left[\boldsymbol{\tau}_n - \tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n}\right]}{\sum_{n=1}^{N^*} \mathrm{SS}[\boldsymbol{\tau}_n - \mathbf{m}_{\mathbf{T}^*}]} \end{aligned} \tag{9.49}$$

Where N^* denotes the number of observations used for the comparison, which includes only those whose prediction would not require extrapolation outside of the convex hull defined by all other remaining observations in a cross-validations exercise; $\mathbf{m}_{\mathbf{T}}$ is the average of the scores for those observations, while $\mathbf{m}_{\mathbf{T}^*}$ is the average for the scores of the observations used to fit the model at the corresponding iteration during the cross-validation to obtain $\tilde{Q}^2$.

The only differences between $\tilde{R}^2_{\mathbf{X}}$ and $\tilde{Q}^2_{\mathbf{X}}$ as defined here with respect to [88] up until now lies in the combination of Equations 9.48 and 9.49. Here all R quality attributes of interest are accounted for, and these quality attributes could here be a linear combination of outputs, and not the outputs per se. Consider now, however, a subset of quality attributes, such that the set of NS equations in matrix form would now be expressed as

$\mathbf{v}_{n,0}^{*} + \mathbf{V}^{*} \cdot \boldsymbol{\tau}_{\mathrm{NS}_n^{*}} = \mathbf{0}^{*}$, where the asterisk is used to denote that $\mathbf{v}_{n,0}^{*}$ and $\mathbf{V}^{*}$ result from extracting some rows from $\mathbf{v}_{n,0}$ and $\mathbf{V}$, respectively. This means that some of the information regarding some outputs may be missing, irrelevant or unreliable. Then $\tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n^{*}}$, equivalent to $\tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n}$ when some of the outputs are not taken into account, can be obtained with Equation 9.47 by substituting $\mathbf{v}_{n,0}$ and $\mathbf{V}$ with $\mathbf{v}_{n,0}^{*}$ and $\mathbf{V}^{*}$, and variations of $\tilde{R}_{\mathbf{X}}^{2}$ and $\tilde{Q}_{\mathbf{X}}^{2}$ can be defined as:

$$\begin{aligned} \tilde{R}_{\mathbf{X}}^{2}\{\mathbf{d}_{R^{*}}\} &= 1 - \frac{\sum_{n=1}^{N^{*}} \mathrm{SS}\left[\boldsymbol{\tau}_n - \tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n^{*}}\right]}{\sum_{n=1}^{N^{*}} \mathrm{SS}[\boldsymbol{\tau}_n - \mathbf{m}_{\mathbf{T}}]} \\ \tilde{Q}_{\mathbf{X}}^{2}\{\mathbf{d}_{R^{*}}\} &= 1 - \frac{\sum_{n=1}^{N^{*}} \mathrm{SS}\left[\boldsymbol{\tau}_n - \tilde{\boldsymbol{\tau}}_{\mathrm{NS}_n^{*}}\right]}{\sum_{n=1}^{N^{*}} \mathrm{SS}[\boldsymbol{\tau}_n - \mathbf{m}_{\mathbf{T}^{*}}]} \end{aligned} \tag{9.50}$$

Where $\{\mathbf{d}_{R^{*}}\}$ denotes that $\tilde{R}_{\mathbf{X}}^{2}$ and $\tilde{Q}_{\mathbf{X}}^{2}$ have been computed only with the information provided by a subset (which must be defined previously) of the quality attributes, and not all of them.

9.7. Additional considerations

This chapter focused on providing efficient tools for the estimation of the DS (as well as the subspaces most likely to contain the TDS and least likely to fall outside of it) when resorting to a PLS-regression model, and the assessment of its performance in its inverse form. However, equipment wear, process modifications etc. will change the TDS of a process and, similarly, the estimated DS will change when new raw materials are tested and process operation is adapted to them (i.e., new conditions not in the historical data used to build the PLS model). A more reasonable approach would seem to be comprised, in this case, by a combination of the methodologies proposed here with recursive PLS-based algorithms for model updating. Nevertheless, this is out of the scope of this thesis and deserves future research.

Chapter 10
Optimization problem formulation in Quality by Design

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

1. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Sequential experimental approach to improve the design space estimation using latent-variable model inversion. Part II. Optimization problem reformulation and sequential experimental approach. *Submitted.*

10.1. Introduction

As mentioned in Section 8.2 and proposed in [90], PLS model inversion can be formulated as an optimization problem in order to find new combinations of inputs leading to a product with the desired quality attributes. In any process, some - but not necessarily all - quality attributes are usually required to meet particular values or some (generally univariate) specifications, while restrictions on the input variables may be imposed, either for feasibility reasons or due to specific combinations of operating conditions being favoured. Regarding the formulation of the optimization problem as presented in [90], which is in the form of a quadratic optimization problem, two similar yet distinct ways in which it can be approached will be presented in this section: the optimization problem *in the original space through the latent space*, and the optimization problem *in the latent space*.

10.1.1. Optimization in the original space through the latent space

The optimization problem formulated this way is shown in Equation 10.1.

$$\min_{\mathbf{x}_{\text{NEW}}} \left[g_0 \cdot (\hat{\mathbf{y}}_{\text{NEW}} - \mathbf{y}_{\text{DES}})^{\text{T}} \cdot \boldsymbol{\Gamma} \cdot \mathbf{D}_{\mathbf{s_Y}}^{-2} \cdot (\hat{\mathbf{y}}_{\text{NEW}} - \mathbf{y}_{\text{DES}}) + g_1 \cdot \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} + g_2 \cdot SPE_{\mathbf{x}_{\text{NEW}}} \right]$$

s.t.

$$\begin{aligned}
&\boldsymbol{\tau} = \mathbf{W}^{*\text{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot (\mathbf{x}_{\text{NEW}} - \mathbf{m}_{\mathbf{X}}) \\
&\hat{\mathbf{y}}_{\text{NEW}} = \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{Y}} \\
&\hat{\mathbf{x}}_{\text{NEW}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{X}} \\
&T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\text{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot \boldsymbol{\tau} \leq T_{\text{lim}}^2 \\
&SPE_{\mathbf{x}_{\text{NEW}}} = (\hat{\mathbf{x}}_{\text{NEW}} - \mathbf{x}_{\text{NEW}})^{\text{T}} \cdot \mathbf{D}_{\mathbf{s_X}}^{-1} \cdot (\hat{\mathbf{x}}_{\text{NEW}} - \mathbf{x}_{\text{NEW}}) \leq g_{2b} \cdot SPE_{\text{lim}} \\
&\mathbf{lb}_{\mathbf{x}_{\text{NEW}}} \leq \mathbf{x}_{\text{NEW}} \leq \mathbf{ub}_{\mathbf{x}_{\text{NEW}}} \\
&\mathbf{lb}_{\hat{\mathbf{y}}_{\text{NEW}}} \leq \hat{\mathbf{y}}_{\text{NEW}} \leq \mathbf{ub}_{\hat{\mathbf{y}}_{\text{NEW}}} \\
&\mathbf{F}_{\mathbf{x}} \cdot \mathbf{x}_{\text{NEW}} = \mathbf{f}_{\mathbf{x}_{\text{NEW}}}
\end{aligned} \tag{10.1}$$

where $\boldsymbol{\Gamma}$ 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 l-th output variable; g_0, g_1 and g_2 are the weights given to each term in the objective function when solving the optimization problem; g_{2b} is a coefficient between 0 and 1 to impose a hard constraint on $SPE_{\mathbf{x}_{\text{NEW}}}$; T_{lim}^2 and SPE_{lim} are the Hotelling T^2 and $SPE_{\mathbf{x}}$ upper limits as calculated in Eq. 8; $\mathbf{lb}$ and $\mathbf{ub}$ are the lower and upper bounds for $\mathbf{x}_{\text{NEW}}$ and $\hat{\mathbf{y}}_{\text{NEW}}$; and $\mathbf{F}_{\mathbf{x}}$ and $\mathbf{f}_{\mathbf{x}_{\text{NEW}}}$ are defined as $\mathbf{F}_{\mathbf{x}}$ and $\mathbf{f}_{\hat{\mathbf{x}}_{\text{NEW}}}$ in Equation 9.16, but defining restriction on $\mathbf{x}_{\text{NEW}}$ instead of $\hat{\mathbf{x}}_{\text{NEW}}$. It should be noted here that $\hat{\mathbf{x}}_{\text{NEW}}$ corresponds to the vector of inputs obtained from $\boldsymbol{\tau}$ through model inversion (i.e. $\hat{\mathbf{x}}_{\text{NEW}} = \mathbf{P} \cdot \boldsymbol{\tau}$), while $\mathbf{x}_{\text{NEW}}$ is the vector of inputs the objective function is expressed as a function of, and whose projection onto the latent space is $\boldsymbol{\tau}$ (i.e. $\boldsymbol{\tau} = \mathbf{W}^{*\text{T}} \cdot \mathbf{x}_{\text{NEW}}$). $SPE_{\mathbf{x}_{\text{NEW}}}$ is, consequently, a measure of the squared difference between $\mathbf{x}_{\text{NEW}}$ with projection $\boldsymbol{\tau}$ and the vector of inputs $\hat{\mathbf{x}}_{\text{NEW}}$ obtained from $\boldsymbol{\tau}$ by indirect inversion of the model when solving Equation 10.1.

A distinction must be made regarding the different restrictions imposed on the solution of this optimization approach. On one hand, the different terms inside the objective function are referred to as soft constraints, since they favour solutions close to the NS, with low Hotelling $T_{\mathbf{x}_{\text{NEW}}}^2$ and low $SPE_{\mathbf{x}_{\text{NEW}}}$ values, but do not define a threshold for such values, nor force the solution to meet any specific values. On the other hand, the different restrictions defined outside the objective function in Equation 10.1 are referred to as hard constraints, since they define the subspace outside of which a solution will not be accepted, but they do not favour any specific solution inside the subspace of acceptable solutions. It is these hard constraints, then, that define the region of acceptable solutions. It should also be noted that the way this optimization problem is formulated does not make use of the concept of the NS presented in [88], or 'combined pseudo-NS' in [89]. Instead, the term $(\hat{\mathbf{y}}_{\text{NEW}} - \mathbf{y}_{\text{DES}})^{\text{T}} \cdot \boldsymbol{\Gamma} \cdot \mathbf{D}_{\mathbf{s_Y}}^{-2} \cdot (\hat{\mathbf{y}}_{\text{NEW}} - \mathbf{y}_{\text{DES}})$ in the objective function implicitly acknowledges the presence of L different NS as referred to in [89], and whose analytical expression have been provided in Section 9.3.3, so that the result from the optimization, $\boldsymbol{\tau}$, has to be as close as possible to these NS. Each of

them is therefore constituted by all combinations of inputs that theoretically guarantee the desired value for a single output variable, but not the rest. If these NS intersect, at least a solution meeting all the desired values for the L out variables, $\mathbf{y}_{\mathrm{DES}}$, exists. If this intersection results in a subspace of dimension higher than 0 (i.e. a line, plane or hyper-plane), then such intersection is coincident with, or very close to, the NS as defined in [88], or combined pseudo-NS in [89].

This formulation of the optimization problem is said to take place *in the original space* because the objective function whose value is to be minimized is ultimately being expressed as a function of a vector of input variables $\mathbf{x}_{\mathrm{NEW}}$ that do not necessarily belong to the subspace of the PLS-regression model that best adjusts the available data (i.e. $SPE_{\mathbf{x}_{\mathrm{NEW}}} \geq 0$); but it is also said to be carried out *through the latent space* because the PLS-regression model is used to predict restrictions $\hat{\mathbf{y}}_{\mathrm{NEW}}$, and restrictions are imposed on both the Hotelling T^2 and $SPE_{\mathbf{x}_{\mathrm{NEW}}}$ to guarantee that any solution obtained will fall within acceptable limits for both parameters (i.e. neither too far away from the centre of projection of the latent subspace, nor from the model's subspace itself). This way, the optimization algorithm will provide a vector of inputs $\mathbf{x}_{\mathrm{NEW}}$ that theoretically guarantees the desired outputs (or values as close to the desired ones as possible) while respecting, to an extent, the correlation structure found in the dataset used to fit the PLS-regression model, but without being forced to comply exactly with it.

For all of its advantages, however, this approach to solving the optimization problem suffers from one severe drawback: as the objective function is expressed as a function of $\mathbf{x}_{\mathrm{NEW}}$, the dimensionality of the space within which it is to be minimized increases linearly with the number of input variables, and the growth of its size is exponential. Being a quadratic optimization problem, an algorithm to solve it may yield one of many local minima as its solution, instead of the global minimum. To avoid this as much as possible, a large number of seeds spanning the space of acceptable solutions (i.e. meeting all restrictions) may be provided to the algorithm, but this may become computationally very costly or even untenable as the number of input variables involved increases.

10.1.2. Optimization in the latent space

The optimization problem formulated this way is shown in Equation 10.2, where $\mathbf{\Gamma}$, g_0 and g_1 are defined as in Equations 10.1; $\mathbf{lb}$ and $\mathbf{ub}$ are the lower and upper bounds for $\hat{\mathbf{x}}_{\mathrm{NEW}}$ and $\hat{\mathbf{y}}_{\mathrm{NEW}}$; and $\mathbf{F}_{\mathbf{x}}$ and $\mathbf{f}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}$ are defined as $\mathbf{F}_{\mathbf{x}}$ and $\mathbf{f}_{\hat{\mathbf{x}}_{\mathrm{NEW}}}$ in Equation 9.16. It should be noted here that $\hat{\mathbf{x}}_{\mathrm{NEW}}$ corresponds to the vector of inputs obtained from $\boldsymbol{\tau}$ through model inversion (i.e. $\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{P} \cdot \boldsymbol{\tau}$), and therefore any solution achieved by solving this optimization problem will be located on the PLS-regression model subspace. Although a restriction is imposed on the Hotelling T^2, none affects the $SPE_{\hat{\mathbf{x}}_{\mathrm{NEW}}}$ since, in this case, $SPE_{\hat{\mathbf{x}}_{\mathrm{NEW}}} = 0$ for any $\boldsymbol{\tau}$ achieved by approaching the optimization problem this way.

$$\min_{\boldsymbol{\tau}} \left[g_0 \cdot (\hat{\mathbf{y}}_{\mathrm{NEW}} - \mathbf{y}_{\mathrm{DES}})^{\mathrm{T}} \cdot \boldsymbol{\Gamma} \cdot \mathbf{D}_{\mathbf{s_Y}}^{-2} \cdot (\hat{\mathbf{y}}_{\mathrm{NEW}} - \mathbf{y}_{\mathrm{DES}}) + g_1 \cdot \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} \right]$$

s.t.

$$\begin{aligned} &\hat{\mathbf{y}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{Y}} \\ &\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{X}} \\ &T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot \boldsymbol{\tau} \leq T_{\mathrm{lim}}^2 \\ &\mathbf{lb}_{\hat{\mathbf{x}}_{\mathrm{NEW}}} \leq \hat{\mathbf{x}}_{\mathrm{NEW}} \leq \mathbf{ub}_{\hat{\mathbf{x}}_{\mathrm{NEW}}} \\ &\mathbf{lb}_{\hat{\mathbf{y}}_{\mathrm{NEW}}} \leq \hat{\mathbf{y}}_{\mathrm{NEW}} \leq \mathbf{ub}_{\hat{\mathbf{y}}_{\mathrm{NEW}}} \\ &\mathbf{F}_{\mathbf{x}} \cdot \hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{f}_{\hat{\mathbf{x}}_{\mathrm{NEW}}} \end{aligned} \tag{10.2}$$

This implies that the optimization algorithm will provide a vector of inputs $\hat{\mathbf{x}}_{\mathrm{NEW}}$ that theoretically guarantees the desired outputs (or values as close to them) and complies exactly with the correlation structure found in the dataset used to fit the PLS-regression model. As opposed to the approach in Section 10.1.1, this may make it impossible to find a set of inputs $\hat{\mathbf{x}}_{\mathrm{NEW}}$ that meets all of the restrictions imposed on them, while an hypothetical $\mathbf{x}_{\mathrm{NEW}}$ outside of the model subspace but close enough to it (i.e. $SPE_{\mathbf{x}_{\mathrm{NEW}}} \leq SPE_{\mathrm{lim}}$) could be found that meets all constraints on the inputs. Consider, for example, the data in Table 9.1 and its projection onto the latent subspace of a PLS-regression model fitted with 2 LV. Trying to solve the optimization problem as formulated in Equation 10.2, and imposing as hard constraints on the solution that all of the inputs must meet exactly the values corresponding to one of the six samples used to fit the PLS-regression model, yields no feasible solution at all. However, a solution can be found for all but one case (the one corresponding to the fourth observation) if the optimization problem is formulated as in Section 10.1.1, since all but the fourth observation meet that $SPE_{\mathbf{x}_n} \leq SPE_{\mathrm{lim}}$. This limitation comes as a consequence of the optimization taking place in the A-dimensional latent subspace instead of the original M-dimensional space ($A \leq M$), as constitutes an example of the issue explained in Section 9.2, where it is stated that at most A equality restrictions (including restrictions imposed on the outputs, and not just on the inputs) may be imposed on the optimization solution as hard constraints without it being highly unlikely that a solution exists at all.

The optimization problem as formulated this way presents, however, a very relevant advantage in practice with respect to the one illustrated in Section 10.1.1, derived precisely from the reduction in dimensionality of the subspace of acceptable solutions, which can significantly reduce the computational cost of the algorithm used for the optimization, as well as the risk of it getting 'trapped' in local minima. Furthermore, given the uncertainty associated to the PLS-regression model's predictions of both inputs and outputs and the fact that, in a number of real scenarios, more hard equality constraints may have to be imposed on some of the input variables than the number of LV variables used to fit the PLS-regression model, a two-steps optimization ought to be resorted to in order to bypass to an extent the limitation regarding how many hard equality constraints can be simultaneously imposed, by i) relaxing some of the equality

constraints into pairs of inequality restrictions when solving Equation 10.2; ii) modifying the values for the restricted variables in $\hat{\mathbf{x}}_{\mathrm{NEW}}$ accordingly to meet the equality constraints while penalising as little as possible the objective function, to get $\mathbf{x}_{\mathrm{NEW}}$; and iii) evaluating if $T^2_{\mathbf{x}_{\mathrm{NEW}}} \leq T^2_{\mathrm{lim}}$ and $SPE_{\mathbf{x}_{\mathrm{NEW}}} \leq SPE_{\mathrm{lim}}$, in which case a valid solution would have been obtained.

In the following sections, the optimization in the latent space (and not the one in the original space) will be addressed in more detail, and some modifications of it and/or alternative approaches will be provided to address different problems of interest. Some of these issues are:

1. The formulation of the objective function is useful when the quality attributes of interest are coincident with the output variables used to fit the PLS-regression model, but not when one or more of them are expressed as a linear combination of such outputs (and they are not explicitly included in the **Y** matrix when fitting the PLS regression model for any of the reasons mentioned in the third point in Section 8.3, and as illustrated in the example addressed in Section 9.4.2.3).
2. The presence of a soft constraint on the Hotelling T^2 will necessarily cause a solution that would otherwise meet all desired values for all of the outputs to not do so (e.g. in the example illustrated in Figure 9.5a, Section 9.4.2.2, a set of scores exists that meets all hard constraints and provides the desired values for the two outputs, but imposing a soft constraint on the Hotelling T^2 would cause the solution of the optimization to drift away from this set of scores). Its absence in this same scenario, however, may lead the optimization algorithm to provide a different solution each time it is executed, provided the combined NS (i.e. a subspace within which all sets of inputs theoretically guarantee the desired values for all of the outputs) exists. Part of this is because of the lack of soft constraints on the inputs being included in the objective function and hard equality constraints on the outputs, the inclusion of both of which may sometimes be advisable. Take as examples the case studies illustrated in Sections 9.4.2.1 and 9.4.2.3, where the NS for the respective quality attributes of interest is seen to exist within the experimental region/region of acceptable solutions for the optimization problem, and be comprised of many different sets values for the inputs/scores:
    a. The optimization problem as formulated in Equation 10.2 with $g_0 > 0$ and $g_1 > 0$ will provide a solution as close as possible to the NS and, at the same time, as close as possible to the centre of projection. How close $\boldsymbol{\tau}_{\mathrm{NEW}}$ is to each will depend on the values given to g_0 and g_1, but it will surely not be a set of scores on the NS.

b. The optimization problem as formulated in Equation 10.2 with $g_0 > 0$ and $g_1 = 0$ will provide any of the many possible solution on the NS, since any of them is equally valid according to the objective function (i.e. the distance to the NS is zero). Which of these is obtained will define, in most cases, on the seed provided to the algorithm as an starting point for the algorithm and from which the 'search' for an optimum begins.

c. Alternatively, the optimization problem as formulated in Equation 10.2 with $g_1 > 0$ and the soft constraint defined by its first term being substituted by a hard constraint on the output variable will provide as a solution the set of scores exactly on the NS with smallest leverage.

3. Even accounting for the issue mentioned in the second point of this list, if more than one solution for the optimization problem is desired, solving it more than once will provide the same or almost equal solutions, unless the weighting of the soft constraints in the objective function or the hard constraints are modified from one iteration to the next one.

4. While the first term in objective function in Equation 10.2 is formulated as a weighted sum of quadratic differences between $\hat{\mathbf{y}}_{\mathrm{NEW}}$ and $\mathbf{y}_{\mathrm{DES}}$, which implies $\mathbf{y}_{\mathrm{DES}}$ must be defined, there are a number of scenarios where the value for one or more of the outputs/quality attributes of interest is to be minimized or maximized. In some of those cases, a minimum or maximum value can be introduced as the 'desired value' for one or more of these quality attributes, when it is already known that such quality attributes cannot achieve values below or above those thresholds, respectively. However, this is not always the case (i.e. because such limits do not exist or, more frequently, are not known) and, even if it is, defining such values as the 'desired ones' may lead to an artificial increase in the weight given to minimizing or maximizing the value of one of the quality attributes with respect to others.

Section 10.2 aims at covering these issues and proposing different ways to deal with each of them, as well as to reduce the uncertainty in the DS estimation, while Sections 10.3 and 10.4 provide alternative formulations to the quadratic one for the optimization problem.

10.2. Quadratic optimization formulation

10.2.1. Optimization of a linear combination of outputs

Consider Equation 9.25 to be the expression corresponding to any set of output variables $\hat{\mathbf{y}}_{\mathrm{NS}_r}$ for which the desired value for the *r*-th quality attribute of interest, d_r, is met, that is $\mathbf{a}_r^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NS}_r} = d_{\mathrm{DES},r}$. The first term in the objective function in Equation 10.2 is expressed as a function of $\hat{\mathbf{y}}_{\mathrm{NEW}}$ and $\mathbf{y}_{\mathrm{DES}}$, and not $\mathbf{a}_r^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NEW}}$ and $d_{\mathrm{DES},r}$. To evaluate the adequacy of this formulation when the *r*-th quality attribute is taken into account, instead of one or more outputs, consider the following objective function:

$$\min_{\boldsymbol{\tau}} \left[g_0 \cdot \left(\mathbf{a}_r^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NEW}} - d_{\mathrm{DES},r} \right)^{\mathrm{T}} \cdot \gamma_r \cdot s_{d_r}^{-2} \cdot \left(\mathbf{a}_r^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NEW}} - d_{\mathrm{DES},r} \right) + g_1 \cdot \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} \right] \quad (10.3)$$

Where γ_r is a scalar that represents the weight given to achieving the desired value for the *r*-th quality attribute of interest, such that the sum of the absolute values for the weights given to achieving the desired values for each quality attribute is one (if only one quality attribute is considered, as in this case, then $\gamma_r = 1$); and $s_{d_r}^{-2}$ is the inverse of the estimated variance for the *r*-th quality attribute from the observations in the calibration dataset.

For the objective functions in Equations 10.2 and 10.3 to be equivalent, the first term in each of them must be so. This first term in Equation 10.3 can be reformulated as:

$$\left(\hat{\mathbf{y}}_{\mathrm{NEW}} - \hat{\mathbf{y}}_{\mathrm{NS}_r} \right)^{\mathrm{T}} \cdot \gamma_r \cdot \mathbf{a}_r \cdot s_{d_r}^{-2} \cdot \mathbf{a}_r^{\mathrm{T}} \cdot \left(\hat{\mathbf{y}}_{\mathrm{NEW}} - \hat{\mathbf{y}}_{\mathrm{NS}_r} \right) \quad (10.4)$$

where $\hat{\mathbf{y}}_{\mathrm{NS}_r}$ is any set of outputs that meets that $\mathbf{a}_r^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NS}_r} = d_{\mathrm{DES},r}$. When $\hat{\mathbf{y}}_{\mathrm{NS}_r}$ in Equation 10.4 coincides with $\mathbf{y}_{\mathrm{DES}}$ in Equation 10.2, then:

$$\boldsymbol{\Gamma} \cdot \mathbf{D}_{\mathbf{SY}}^{-2} = \gamma_r \cdot \mathbf{a}_r \cdot s_{d_r}^{-2} \cdot \mathbf{a}_r^{\mathrm{T}} \quad (10.5)$$

It must be noted, however, that the estimated variance for the quality attribute from the observations in the calibration dataset, $s_{d_r}^2$, is calculated as:

$$s_{d_r}^2 = \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{S}_\mathbf{y} \cdot \mathbf{a}_r$$

$$\mathbf{S}_\mathbf{y} = \begin{pmatrix} s_{y_1}^2 & \widehat{\mathrm{cov}}(y_1, y_2) & \cdots & \widehat{\mathrm{cov}}(y_1, y_L) \\ \widehat{\mathrm{cov}}(y_2, y_1) & s_{y_2}^2 & & \widehat{\mathrm{cov}}(y_2, y_L) \\ \vdots & & \ddots & \vdots \\ \widehat{\mathrm{cov}}(y_L, y_1) & \widehat{\mathrm{cov}}(y_L, y_2) & \cdots & s_{y_L}^2 \end{pmatrix} \tag{10.6}$$

$$s_{y_l}^2 = \frac{\sum_{n=1}^{N} \left(y_{n,l} - \bar{y}_l\right)^2}{N - df}$$

$$\widehat{\mathrm{cov}}(y_l, y_{l'}) = \frac{\sum_{n=1}^{N} \left(y_{n,l} - \bar{y}_l\right) \cdot \left(y_{n,l'} - \bar{y}_{l'}\right)}{N - df}$$

$\bar{y}_l$ and $y_{n,l}$ being, respectively, the average and *n*-th measured value for the *l*-th output of the model calibration dataset, and df the degrees of freedom consumed by the model. Therefore:

$$\mathbf{\Gamma} = \gamma_r \cdot \mathbf{a}_r \cdot \left(\mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{S}_\mathbf{y} \cdot \mathbf{a}_r\right)^{-1} \cdot \mathbf{a}_r^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}}^2 \tag{10.7}$$

Since $\mathbf{\Gamma}$ in Equation 10.2 is defined as a diagonal matrix of weightings, the objective function in Equation 10.2 and the one formulated in Equation 10.3 will only be equivalent, with $\mathbf{\Gamma}$ remaining a diagonal matrix, if the outputs are uncorrelated[xi].

Therefore, a more general reformulation of Equation 10.2 to account for quality attributes of interest expressed as linear combinations of the outputs is:

$$\min_{\boldsymbol{\tau}} [\mathrm{OF}_1(g_{0\mathrm{a}}, g_1, \boldsymbol{\tau})]$$

$$\mathrm{OF}_1(g_{0\mathrm{a}}, g_1, \boldsymbol{\tau}) = g_{0\mathrm{a}} \cdot (\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau})^{\mathrm{T}} \cdot \mathbf{\Gamma}_R \cdot \mathbf{D}_{\mathbf{s_d}}^{-2} \cdot (\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}) + g_1 \cdot \sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2}$$

s.t.

$$\begin{aligned} &\hat{\mathbf{y}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau} + \mathbf{m}_\mathbf{Y} \\ &\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau} + \mathbf{m}_\mathbf{X} \\ &T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\mathrm{T}} \cdot \mathbf{\Lambda}^{-1} \cdot \boldsymbol{\tau} \leq T_{\mathrm{lim}}^2 \\ &\mathbf{A}_{\boldsymbol{\tau}} \cdot \boldsymbol{\tau} \leq \mathbf{d}_{\boldsymbol{\tau}} \\ &\mathbf{F}_{\boldsymbol{\tau}} \cdot \boldsymbol{\tau} = \mathbf{f}_{\boldsymbol{\tau}} \end{aligned} \tag{10.8}$$

where $\mathbf{v}_0$ and $\mathbf{V}$ are as defined in Equation 9.26[xii], and correspond to the NS associated to the desired values of the R quality attributes of interest, $\mathbf{d}_{\mathrm{DES}}$ [R×1]; $\mathbf{\Gamma}_R$ is now a [R×R] diagonal matrix where the *r*-th element in the diagonal represents the weight

[xi] The extension of this demonstration when all R quality attributes of interest are considered is straightforward

[xii] Note that $\mathbf{v}_0$ and $\mathbf{V}$ can be re-defined to also include the 'null spaces' of the different linear combinations of inputs if/whenever necessary. Doing so is straightforward, and has not been done in the objective function presented here for the sake of simplicity

given to the goal of achieving the desired value for *r*-th quality attribute; $\mathbf{D}_{\mathbf{s_d}}^{-2}$ is the $[R\times R]$ diagonal matrix with $s_{d_r}^{-2}$ (calculated as in Equation 10.6) as its *r*-th element; g_{0a} and g_1 are defined as g_0 and g_1 in Equations 10.2; $\mathbf{A}_{\boldsymbol{\tau}}$ and $\mathbf{d}_{\boldsymbol{\tau}}$ as $\mathbf{A}_{\boldsymbol{\tau}}$ and $\mathbf{d}_{\boldsymbol{\tau}_{\mathrm{NEW}}}$ in Equation 9.11; and $\mathbf{F}_{\boldsymbol{\tau}}$ and $\mathbf{f}_{\boldsymbol{\tau}}$ as $\mathbf{F}_{\boldsymbol{\tau}}$ and $\mathbf{f}_{\boldsymbol{\tau}_{\mathrm{NEW}}}$ in Equation 9.13.

10.2.2. Optimization for exploration and DOE in the latent space

As mentioned in Section 10.1, one other drawback of the optimization formulation in Equation 10.2 is the need to actively modify the soft and/or hard constraints (or the weight given to the former in the objective function) in order to guarantee that different iterations will provide solutions different to previous ones, or to already available observations in the calibration dataset used to fit the PLS-regression model. Furthermore, when the optimization approach is resorted to in order to perform new experiments (e.g. in unexplored areas of the knowledge space or the subspace thought to most likely contain the TDS), and additional issue must be considered, and that is the fact that the first term in the objective function favours/penalizes equally acceptable solutions at the same distance (in the latent space) from the NS. However, if it is accepted that the limits of the confidence region for the NS are non-linear (and, in particular and as an approximation, of the form presented in Section 9.3.4) due to the prediction error varying with the leverage associated to each set of inputs/scores, then it makes sense for exploratory purposes to penalize less solutions equally close to the NS but with higher leverage, and vice versa.

When both of these concerns are accounted for, the following optimization problem can be formulated:

$$\min_{\boldsymbol{\tau}}[\mathrm{OF}_2(g_{0\mathrm{a}}, g_{0\mathrm{b}}, g_1, g_2, \boldsymbol{\tau})]$$

$$\begin{aligned}\mathrm{OF}_2(g_{0\mathrm{a}}, g_{0\mathrm{b}}, g_1, g_2, \boldsymbol{\tau}) =& \mathrm{OF}_1(g_{0\mathrm{a}}, g_1, \boldsymbol{\tau}) + g_{0b} \cdot \mathbf{d}_{\mathbf{NS}}^{\mathrm{T}}(\boldsymbol{\tau}, \mathbf{NS}) \cdot \boldsymbol{\Gamma}_R \cdot \mathbf{d}_{\mathbf{NS}}(\boldsymbol{\tau}, \mathbf{NS}) + \\ &+ g_2 \cdot d^{-1}(\boldsymbol{\tau}, \mathbf{T})\end{aligned}$$

s.t.

$$\begin{aligned}&\hat{\mathbf{y}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{Y}} \\ &\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{X}} \\ &\left|d_{\mathrm{NS}_r}(\boldsymbol{\tau}, \mathrm{NS}_r)\right| = \left|\frac{v_{r,0} + \mathbf{v}_r^{\mathrm{T}} \cdot \boldsymbol{\tau}}{s_{\hat{d}_{\mathrm{NEW},r}}}\right| \leq t_{N-A,\frac{\alpha}{2}} \quad \forall\, r \in \{1, 2, \ldots, R\} \\ &T_{\boldsymbol{\tau}}^2 = \boldsymbol{\tau}^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot \boldsymbol{\tau} \leq T_{\mathrm{lim}}^2 \\ &\mathbf{A}_{\boldsymbol{\tau}} \cdot \boldsymbol{\tau} \leq \mathbf{d}_{\boldsymbol{\tau}} \\ &\mathbf{F}_{\boldsymbol{\tau}} \cdot \boldsymbol{\tau} = \mathbf{f}_{\boldsymbol{\tau}}\end{aligned} \tag{10.9}$$

where $\boldsymbol{\Gamma}_R$ holds the same meaning as in Equation 10.8; g_{0b} and g_2 are the weights given to each of the new added terms in the objective function; $\mathbf{d}_{\mathbf{NS}}(\boldsymbol{\tau}, \mathbf{NS})$ is a [R x 1] vector with $d_{\mathrm{NS}_r}(\boldsymbol{\tau}, \mathrm{NS}_r)$ as its *r*-th element, which is a measure of the distance from the projection $\boldsymbol{\tau}$ of the solution to the *r*-th NS pondered by the inverse of $s_{\hat{d}_{\mathrm{obs},r}}$ (as

calculated in Equation 9.31), instead of the *r*-th diagonal element in $\mathbf{D}_{\mathbf{sd}}^{-2}$; and $d^{-1}(\boldsymbol{\tau}, \mathbf{T})$ is the inverse of the squared statistical distance from the projection $\boldsymbol{\tau}$ of the solution to the closest already available sample in the calibration dataset **X** with projection **T** in the latent subspace, calculated as:

$$d^{-1}(\boldsymbol{\tau}, \mathbf{T}) = \frac{1}{\min[(\boldsymbol{\tau} - \boldsymbol{\tau}_n)^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot (\boldsymbol{\tau} - \boldsymbol{\tau}_n)] + const} \; \forall \, n \, \epsilon \, \{1,2,\dots,n\} \qquad (10.10)$$

$const$ being a small positive real number introduced in order to avoid $d^{-1}(\boldsymbol{\tau}, \mathbf{T})$ going to infinite, which may cause computational issues.

It must be noted that the terms of the objective function weighted by g_{0a} and g_{0b} in Equation 10.9 both penalize solutions further away from the NS, but do so in different ways, and may be resorted to for different purposes (e.g. the first one for optimization purposes, as done in the literature, and the second one for DOE/exploratory purposes, as will be illustrated in Chapter 12). Therefore using one or the other, but not both simultaneously, is recommended. A hard constraint defined similarly to the soft one weighted by g_{0b} may also be included in order to guarantee that any solution, if it satisfies all hard constraints, falls within the subspace most likely to contain the TDS as defined in Section 9.4[xiii].At the same time, $d^{-1}(\boldsymbol{\tau}, \mathbf{T})$ guarantees that any solution obtained from the optimization will not be coincident or too close to an already available observation in the dataset used to fit the PLS regression model. One of the main disadvantages of this optimization approach is that the solution achieved may change substantially depending not only on the constraints imposed but also on the weight given to each of the terms in the objective function $\mathrm{OF}_2(g_{0\mathrm{a}}, g_{0\mathrm{b}}, g_1, g_2, \boldsymbol{\tau})$, even when there is only one quality attribute of interest. To illustrate this, consider a PLS-regression model is fitted with 2 LV, using data consisting of six observations following the hypothetical input-output causal relationship presented in Equation 8.1, already used in one of the examples in [87] and in Section 9.4. Then, consider the desired value for the output to be $y_{\mathrm{DES}} = 204.86$, and this problem to be addressed as an optimization problem. Figure 10.1 presents, for the same example shown in Figure 9.4 (Section 9.4.2.1), and at each point in the latent space, the values that the objective function in Equation 10.9 takes when weight equal to 1 is given to one of the terms in it, and 0 to the rest. This highlights the different contribution to each of the terms in the objective function, and allows assessing the effect of different weights being given to each one (g_{0a}, g_{0b}, g_1 and g_2) prior to the optimization.

[xiii] For the more likely scenario of a range of values within specifications limits being accepted for a given quality attribute, instead of a single value, these hard constraints can be substituted by the ones defined in Equation 9.28, extended for a linear combination of outputs as detailed in Section 9.3.4

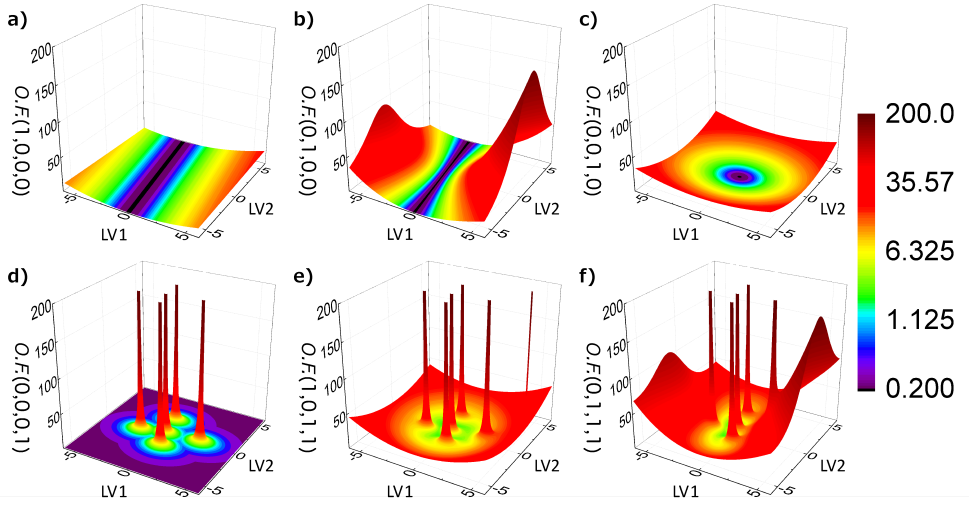


Figure 10.1. Graphical representation of the contribution of each term in the objective function of the optimization framework of Equation 10.9 when using a PLS-regression model with 2 LV fitted with six observations following the input-output causal relationship presented in Equation 8.1, and $y_{\mathrm{DES}} = 204.86$, as well as two combinations of these terms. The z-axis represents the value for the Objective Function (O.F.) in Equation 10.9 referred to in these plots as $O.F.(g_{0a}, g_{0b}, g_1, g_2)$.

In Figure 10.1.a and Figure 10.1.c the contributions of each term in the O.F. of the optimization framework as formulated by Equation 10.2 can be assessed. It can be seen how these terms' contributions to the O.F. are of similar order of magnitude if close to the centre of projection. Since during the optimization the lowest value for the O.F. is desired, the term corresponding to $O.F.(1,0,0,0)$ penalizes solutions further away from the NS (for $y_{DES} = 204.86$), while the term associated to $O.F.(0,0,1,0)$ favours solutions with lower Hotelling $T^2_{\boldsymbol{\tau}}$.

On the other hand, the contribution of one of the new proposed terms can be evaluated through Figure 10.1.b, with $O.F.(0,1,0,0)$ corresponding to $d_{\mathrm{NS}_r}(\boldsymbol{\tau}, \mathrm{NS}_r)$. This term favours solutions in a similar manner as $O.F.(1,0,0,0)$. However, it penalizes solutions farther away from the NS much more. Furthermore, solutions with a higher leverage (and therefore higher uncertainty in the prediction) are less penalized when compared to those with a lower leverage, which implies that this term will allow a wider area for the exploration of different solutions where the predictions are less reliable. Figure 10.1.d permits assessing the impact of the other new proposed term for the O.F. on its value, through $O.F.(0,0,0,1)$. This term heavily penalizes solutions close to the projection onto the latent subspace of already available observations in the calibration dataset.

Finally, Figure 10.1.e corresponds to $O.F.(1,0,1,1)$, that is, the sum of the terms plotted in Figure 10.1.a, Figure 10.1.c and Figure 10.1.d. It must be noted that, because the

contributions of $O.F.(1,0,0,0)$ and $O.F.(0,0,1,0)$ are of a similar order of magnitude close to the centre of projection, there are solution further away from the NS that are equally optimal to solutions closer to it if g_{0a} and g_1 are given the same values, according to the O.F. as formulated this way. On the other hand, Figure 10.1.f corresponds to $O.F.(0,1,1,1)$, that is, the sum of the terms plotted in Figure 10.1.b, Figure 10.1.c and Figure 10.1.d. In this case any solution will fall, even without hard restrictions, closer to the NS than if the weights given to each term were those corresponding to Figure 10.1.e, even if g_{0b} and g_1 are given the same values. This serves to illustrate the importance of giving the appropriate weight to each term in the objective function to guarantee that more desirable solutions are achieved.

This graphical representation also highlights the importance of addressing if a more convenient formulation of the optimization problem would require dismissing the terms associated to soft constraints on the quality attributes in the objective function (i.e. $g_{0a} = g_{0b} = 1$) and imposing, instead, hard constraints on them, such that the solution will be that exactly on the NS and with smallest leverage (and therefore a narrower prediction confidence interval for the prediction of the quality attribute/s) or, ii) if the prediction error is not of concern at this stage, considering all the combinations of inputs/scores on the NS as 'equally good' solutions (which would, in fact, require defining this subspace instead of solving the optimization problem).

Further application of this formulation of the optimization problem will be illustrated in Chapter 11.

10.2.3. Tackling the maximization/minimization problem

As mentioned in Section 10.1.2, and illustrated in Figure 10.1, the first term in the objective function of the optimization problem as formulated up until now penalizes solutions far away from the hyperplanes defined by the NS corresponding to a vector of 'desired values' for the considered outputs/quality attributes. This may present some issues when such 'desired value' is not a specific one, but instead the value for the corresponding quality attribute is to be minimized or maximized. Here three different ways to address this scenario are presented, as well as the reason why they are or not recommended if more than one output variable/quality attribute is present.

10.2.3.1 Defining feasible minimum/maximum values as the desired ones

Consider a scenario for which the information in Table 10.1 is available concerning two hypothetical output variables y_1 and y_2, for their average, minimum and maximum values and their standard deviation in the calibration dataset used to fit the PLS-regression model, as well as their minimum and maximum feasible values.

Table 10.1. Characterization of a hypothetical **Y** dataset with two output variables

Variable	Dataset				Feasible	
	Mean	Std. dev	Minimum	Maximum	Minimum	Maximum
y_1	50	1.7	45	55	30	1000
y_2	95	3.4	90	110	0	120

Consider now three different scenarios:

1. Both y_1 and y_2 are to be maximized. If the vector of desired values for the outputs is defined as $\mathbf{y}_{\mathrm{DES}}^{\mathrm{T}} = [1000 \quad 120]$, the difference between $\boldsymbol{y}_{\mathrm{DES},\mathbf{1}}$ and the maximum observed value for y_1 is ~556 standard deviations, while the difference between $y_{\mathrm{DES},2}$ and the maximum observed value for y_2 is less than 3 standard deviations. Therefore, defining $\boldsymbol{\Gamma}_R = \mathbf{I}_2$, $\mathbf{I}_2$ being the [2×2] identity matrix, makes it look like the same importance is being given to achieved the 'desired' (i.e. maximum) value for both outputs. However, accomplishing this goal for the first output is being given ~189 times as much weight as achieving it for the second one. This will only *not* be a problem when the combinations of inputs (e.g. process conditions) that maximize both y_1 and y_2 are the same.
2. Both y_1 and y_2 are to be minimized. In this case, if the vector of desired values for the outputs is defined as $\mathbf{y}_{\mathrm{DES}}^{\mathrm{T}} = [30 \quad 0]$, the difference between $y_{\mathrm{DES},1}$ and the minimum observed value for y_1 is ~8.8 standard deviations, while the difference between $\boldsymbol{y}_{\mathrm{DES},2}$ and the minimum observed value for y_2 is ~26.5 standard deviations. In contrast to the previous scenario, when $\boldsymbol{\Gamma}_R = \mathbf{I}_2$ the goal of minimizing y_2 is being given roughly 3 times the weight given to minimizing y_1. Again, this will only *not* be a problem when the combinations of inputs that minimize both y_1 and y_2 are the same.
3. y_1 is to be minimized/maximized, and y_2 is to take a specific desired value, or vice versa. A similar issue to that in the two previous cases occurs here, to higher or lower extent depending on which variable is to be minimized/minimized, and the desired value for the other one.

It can then be concluded that this way of defining $\mathbf{y}_{\mathrm{DES}}$ (or $\mathbf{d}_{\mathrm{DES}}$) may be inadequate in most cases, especially if a feasible minimum or maximum is not known or defining one does not make sense from a practical point of view, and therefore an arbitrarily low or high value would be set.

10.2.3.2 Changing the sign of the weight given in the objective function

This second approach relies in the way the first and second terms in the objective function as formulated in Equation 10.9 contribute to the value of the objective function

(which is to be minimised). Note that these terms are quadratic distances (which are always positive) weighted by the elements of the diagonal matrix $\mathbf{\Gamma}_R$. As a consequence, when the *r*-th element in $\mathbf{\Gamma}_R$ is positive, these terms are minimised when the *r*-th quadratic distance (to $d_{\mathrm{DES},r}$) gets smaller. On the other hand, if the *r*-th element in $\mathbf{\Gamma}_R$ is negative, then these terms are minimised when the *r*-th quadratic distance (to $d_{\mathrm{DES},r}$) gets larger. Therefore, making the *r*-th element in $\mathbf{\Gamma}_R$ negative will penalize solutions that provide values *close to the 'desired' one* instead of values *far away from the 'desired' one* for the *r*-th quality attribute of interest. One can take advantage of this for minimization/maximization purposes, by:

i. defining $d_{\mathrm{DES},r}$ as the average value for the *r*-th quality attribute in the calibration dataset, and imposing the hard inequality constraint $d_{\mathrm{NEW},r} \leq d_{\mathrm{DES},r}$ when the *r*-th quality attribute is to be minimised;
ii. defining $d_{\mathrm{DES},r}$ as the average value for the *r*-th quality attribute in the calibration dataset, and imposing the hard inequality constraint $d_{\mathrm{NEW},r} \geq d_{\mathrm{DES},r}$ when the *r*-th quality attribute is to be minimised.

Although apparently slightly better than the approach in Section 10.2.3.1, this one presents some issues derived from having to artificially impose additional constraints on the outputs/quality attributes of interest, which may make finding a solution other than the average in the dataset impossible, specially if maximizing/minimizing different quality attributes are competing objectives, which would not be a particularly interesting nor useful solution. To avoid this, the maximum/minimum feasible value for the *r*-th quality attribute may be used to define $d_{\mathrm{DES},r}$ and to impose the corresponding inequality restrictions on them, but the same issue as the one presented in Section 10.2.3.1 would then occur, although more weight would implicitly be given in this case to achieving the minimum/maximum values for the variables with highest variance in the calibration dataset. This approach is also slightly counter-intuitive, since $d_{\mathrm{DES},r}$ is not, actually, the desired value for the *r*-th quality attribute.

10.2.3.3 Finding extreme achievable values below the Hotelling T^2 limit

This approach is based on the fact that not just any solution of the optimization problem as formulated in Equation 10.9 is equally valid (optimality aside), and particularly no solution outside of the Hotelling T^2 confidence hyperellipsoid, for a given confidence level, will be accepted, independently of any other restrictions imposed on the inputs or outputs or the feasible minimum or maximum values for any variable. Therefore, not accounting for additional constraints, the lowest and highest predicted values for a quality attribute according to the PLS-regression model will be those for two sets of scores located on opposite extremes of the Hotelling T^2 confidence hyperellipsoid. To identify them, consider that the vector $\mathbf{v}_r$ orthogonal to the NS associated to the *r*-th quality attribute, and defined as the transpose of the *r*-th row of $\mathbf{V}$ in Equation 9.26,

provides the direction of maximum variability for such variable. Consider also that the equation of the Hotelling T^2 confidence hyperellipsoid can be written as:

$$\boldsymbol{\tau}^{\mathrm{T}} \cdot \frac{\boldsymbol{\Lambda}^{-1}}{T^2_{\mathrm{lim}}} \cdot \boldsymbol{\tau} = 1 \tag{10.11}$$

And, $\boldsymbol{\Lambda}^{-1}$ being a diagonal matrix, the vector normal to such hyperellipsoid on a point $\boldsymbol{\tau}_r^+$ on its surface, $\mathbf{n}_{\boldsymbol{\tau}_r^+}$, is:

$$\mathbf{n}_{\boldsymbol{\tau}_r^+} = \frac{2}{T^2_{\mathrm{lim}}} \cdot \boldsymbol{\Lambda}^{-1/2} \cdot \boldsymbol{\tau}_r^+ \tag{10.12}$$

If $\boldsymbol{\tau}_r^+$ is the set of scores associated to the combination of inputs theoretically (i.e. according to the PLS-regression model) guaranteeing the maximum achievable value for the *r*-th quality attribute without reaching outside of the Hotelling T^2 hyperellipsoid for a given confidence level, then:

$$\begin{gathered} \frac{\mathbf{n}_{\boldsymbol{\tau}_r^+}}{\sqrt{\mathbf{n}_{\boldsymbol{\tau}_r^+}{}^{\mathrm{T}} \cdot \mathbf{n}_{\boldsymbol{\tau}_r^+}}} = \frac{\mathbf{v}_r}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} \\ \boldsymbol{\tau}_r^{+\mathrm{T}} \cdot \frac{\boldsymbol{\Lambda}^{-1}}{T^2_{\mathrm{lim}}} \cdot \boldsymbol{\tau}_r^+ = 1 \end{gathered} \tag{10.13}$$

On the other hand, if $\boldsymbol{\tau}_r^-$ is the set of scores associated to the combination of inputs theoretically guaranteeing the minimum achievable value for the *r*-th quality attribute without reaching outside of the Hotelling T^2 hyperellipsoid for a given confidence level, then:

$$\left.\begin{gathered} \frac{\mathbf{n}_{\boldsymbol{\tau}_r^-}}{\sqrt{\mathbf{n}_{\boldsymbol{\tau}_r^-}{}^{\mathrm{T}} \cdot \mathbf{n}_{\boldsymbol{\tau}_r^-}}} = \frac{-\mathbf{v}_r}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} \\ \boldsymbol{\tau}_r^{-\mathrm{T}} \cdot \frac{\boldsymbol{\Lambda}^{-1}}{T^2_{\mathrm{lim}}} \cdot \boldsymbol{\tau}_r^- = 1 \end{gathered}\right\} \rightarrow \mathbf{n}_{\boldsymbol{\tau}_r^-} = -\mathbf{n}_{\boldsymbol{\tau}_r^+} \rightarrow \boldsymbol{\tau}_r^- = -\boldsymbol{\tau}_r^+ \tag{10.14}$$

Equations 10.13 and 10.14 can then be solved to obtain $\boldsymbol{\tau}_r^+$ and $\boldsymbol{\tau}_r^-$, and from them the maximum and minimum achievable values for the *r*-th quality attribute while inside the Hotelling T^2 confidence hyperellipsoid, respectively. These values can be used as the *r*-th element of $\mathbf{d}_{\mathrm{DES}}$ in the optimization problem formulated in Equation 10.9. While using them does not necessarily guarantee that unbalanced weightings are implicitly given to the goals of maximizing/minimizing each of the quality attributes, the procedure to define them does not depend on any "feasibility constraints" or on the decision of the person resorting to the optimization problem as formulated in Equation 10.9 to give them arbitrarily large/small values. Incidentally, if a single quality attribute is considered for the optimization, and if $\boldsymbol{\tau}_r^+$ or $\boldsymbol{\tau}_r^-$ meet all other hard constraints imposed

on the solution of the optimization algorithm, these scores can be considered the solution of the algorithm itself when no soft constraint is imposed on its Hotelling $T^2_{\boldsymbol{\tau}}$ (i.e. $g_1 = 0$).

10.3. Linear optimization formulation

Section 10.2 addresses some limitations of the optimization problem as formulated in the literature, such as the optimization of quality attributes defined as linear combinations of output variables, or the way to tackle the minimization/maximization problems. However, one yet unaddressed drawback is the fact that the quadratic optimization formulation suffers from the possibility of the algorithm used to solve it getting stuck in a local minimum. The quadratic formulation is useful in this context because it can deal with non-linear constraints, such as the one imposed on the solution's Hotelling $T^2_{\boldsymbol{\tau}}$, or the ones associated to the NS confidence regions' limits. It must be noted, however, that quadratic formulation of the optimization problem may not be necessary if the soft constraint on the Hotelling $T^2_{\boldsymbol{\tau}}$ in the objective function is dismissed (while keeping the hard constraint to avoid extrapolation) and any non-linear hard constraints happen to be redundant, i.e. less restrictive than any other considered linear constraint. The linear formulation of the optimization problem is, then, as shown in Equation 10.15.

$$\min_{\boldsymbol{\tau}} \left[\boldsymbol{\gamma}_R^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_d}}^{-1} \cdot (\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}) \right]$$

s.t.

$$\begin{aligned} &\hat{\mathbf{y}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{Q} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{Y}} \\ &\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{D}_{\mathbf{s_X}} \cdot \mathbf{P} \cdot \boldsymbol{\tau} + \mathbf{m}_{\mathbf{X}} \\ &\mathbf{A}_{\boldsymbol{\tau}} \cdot \boldsymbol{\tau} \leq \mathbf{d}_{\boldsymbol{\tau}} \\ &\mathbf{F}_{\boldsymbol{\tau}} \cdot \boldsymbol{\tau} = \mathbf{f}_{\boldsymbol{\tau}} \end{aligned} \tag{10.15}$$

where $\boldsymbol{\gamma}_R$ is a [R×1] vector that contains as its r-th element the weight given to minimising r-th quality attribute (if it is to be maximised, the sign of the r-th element in $\boldsymbol{\gamma}_R$ must be changed).

There are some distinctions that must be made between the linear and quadratic optimization. First of all, the way the quadratic optimization is formulated guarantees the existence of at least one minimum (independently of the algorithm being able to reach it, or it being the global minimum), and required no hard constraints to achieve convergence. When resorting to the linear optimization formulation, however, one does not minimize a quadratic distance, but a directed distance (i.e. it may be positive or negative), and therefore hard constraints must be imposed to guarantee the convergence of the algorithm. This means that, at the very least, a lower/upper limit must be defined when minimising/maximising. It also means, however, that no issues will arise when resorting to the linear optimization associated to 'artificially' giving too much/little importance to achieving the minimum/maximum value for one or more of the quality attributes of interest.

On the other hand, the quadratic formulation is especially useful when specific values are desired for the outputs (or even the inputs, or linear combinations of any of them), precisely because of the fact that quadratic distances to those values are minimized. The linear optimization formulation however requires, in this same scenario, artificially imposing a hard lower limit equal to the desired value for the *r*-th quality attribute as an inequality constraint on it.

The main advantage of this formulation of the optimization problem is that it will provide a solution that, if meeting all the inequality restrictions imposed in the quadratic formulation, is also its optimum, while not presenting the same convergence issues as the quadratic ones. One of the possible disadvantages of this formulation, other than not being able to account for non-linear restrictions when optimizing the objective function, is that in very heavily restricted convex spaces of acceptable solutions with narrow 'corners', the algorithm used may not find the global optimum.

Alternatively, an approach to find an optimal solution is proposed which allows taking into account both linear and non-linear hard constraints (except non-linear equality constraints), while avoiding the convergence issues related to local minima for quadratic programming and those related to very narrow spaces of acceptable solutions for linear programming. The steps to follow for this sequential approach are:

1. Define the space of acceptable solutions considering only linear hard constraints. This can be done by using either the variant of the COSIM algorithm or the 'Segmentation by defining and discarding vertices' approach explained in Section 5.2.2. By doing so, the [$I \times A$] matrix and [$I \times 1$] vector of active linear inequality hard constraints $\mathbf{A_\tau}$ and $\mathbf{d_\tau}$, and the [$J \times A$] matrix and [$J \times 1$] vector of I active linear equality hard constraints $\mathbf{F_\tau}$ and $\mathbf{f_\tau}$ can be obtained, as well as a matrix of vertices of the resulting convex space. If they exist (i.e. a subspace of acceptable solution exists given the linear hard constraints imposed), continue to the next step. Otherwise some restrictions may have to be relaxed or eliminated, if possible (if not, there is no solution for the optimization).
2. Check if at least one linear combination of the vertices of the envelope of the subspace of acceptable solutions within this subspace meets all the non-linear hard constraints imposed on the solution. If it does, define such linear combination as the 'initial solution' $\boldsymbol{\tau}_\mathbf{o}$ and continue to the next step (if not, there is no solution for the optimization).
3. Define the 'vector of maximum variability' $\mathbf{v}_{\mathrm{MV}} = \mathbf{V}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_d}}^{-1} \cdot \boldsymbol{\gamma}_R$, where the *r*-th element of $\boldsymbol{\gamma}_R$ is positive if the corresponding quality attribute will be 'maximized', and negative otherwise, as well as the 'original displacement vector' $\mathbf{v}_{\mathrm{disp},0}$, which is the projection of $\mathbf{v}_{\mathrm{MV}}$ onto the subspace defined by the active linear equality hard constraints such that:

$$\mathbf{v}_{\mathrm{disp},0} = [\mathbf{I}_A - \mathbf{F}_{\boldsymbol{\tau}}^{\mathrm{T}} \cdot (\mathbf{F}_{\boldsymbol{\tau}} \cdot \mathbf{F}_{\boldsymbol{\tau}}^{\mathrm{T}})^{-1} \cdot \mathbf{F}_{\boldsymbol{\tau}}] \cdot \mathbf{v}_{\mathrm{disp},0} + \mathbf{F}_{\boldsymbol{\tau}}^{\mathrm{T}} \cdot (\mathbf{F}_{\boldsymbol{\tau}} \cdot \mathbf{F}_{\boldsymbol{\tau}}^{\mathrm{T}})^{-1} \cdot \mathbf{f}_{\boldsymbol{\tau}} \tag{10.16}$$

4. Obtain the projection of $\mathbf{v}_{\text{disp},0}$ on the vectors normal to the i-th hyperplane, $\mathbf{a}_{\boldsymbol{\tau}}^{i}$ (which is the transpose of the i-th row vector in $\mathbf{A}_{\boldsymbol{\tau}}$), $\mathbf{v}_{\text{disp},i}$, as in Equation 10.17, and the scalars δ_i such that $\left(\boldsymbol{\tau}_{\text{o}} + \delta_i \cdot \mathbf{v}_{\text{disp},i}\right) \cdot \mathbf{a}_{\boldsymbol{\tau}}^{i} = d_{\boldsymbol{\tau}}^{i}$, where $d_{\boldsymbol{\tau}}^{i}$ is the i-th element of $\mathbf{d}_{\boldsymbol{\tau}}$.

$$\mathbf{v}_{\text{disp},i} = \frac{\mathbf{a}_{\boldsymbol{\tau}}^{i\,\text{T}} \cdot \mathbf{v}_{\text{disp},0}}{\sqrt{\mathbf{a}_{\boldsymbol{\tau}}^{i\,\text{T}} \cdot \mathbf{a}_{\boldsymbol{\tau}}^{i}}} \cdot \mathbf{a}_{\boldsymbol{\tau}}^{i} \tag{10.17}$$

5. Compute also δ_k^*, such that for the k-th non-linear hard constraint $\boldsymbol{\tau}_{\text{o}} + \delta_i \cdot \mathbf{v}_{\text{disp},0}$ meets it.
6. Arrange all δ_i and δ_k^* from lowest to highest values, and discard all the negative ones.

If the space of acceptable solutions is a convex space, the solution (or solutions) of the optimization problem will be located in the intersection of the linear equality restrictions with the first $(A - J)$ inequality hard constraints (either linear or non-linear, depending which ones are 'reached' first) associated to the ordered list in step 6. Otherwise this solution (or solutions) is not guaranteed to be optimal, but can be used as a good 'seed' for the quadratic optimization algorithm.

10.4. A sequential optimization approach

The linear optimization formulation presented in Section10.3 can be useful to tackle some of the disadvantages associated to the quadratic one. However, it may still prove to not be a powerful enough tool in presence of overly complex, not redundant, non-linear, hard restrictions. At the same time, it may also be unnecessarily complex (from a computational point of view) in the absence of non-redundant non-linear constraints. Therefore, the following sequential procedure is proposed in order to minimize the computational cost of solving an optimization problem. This method makes use of the previous algorithms, from the simplest to the most complex ones, only when necessary:

1. Define the space of acceptable solutions considering only linear hard constraints, as in the first step of the proposed algorithm in Section 10.3, and evaluate the objective function (as in Equation 10.9) for each vertex. If the vertex or vertices that optimize the value of the objective function meet all hard constraints imposed, then the solution/s of the optimization has been achieved. Otherwise, continue to the next step.
2. Find the set of scores on the Hotelling T^2 confidence hyperellipsoid that optimizes the objective function as defined in Equation 10.15 while meeting all linear hard equality constraints. If it meets all other hard inequality constraints, then the solution of the optimization has been achieved. Otherwise, continue to the next step.

3. Solve the optimization problem according to the proposed algorithm in Section 10.3, using as a 'seed' or 'starting solution' the average of the vertices defined in step 1, or any linear combination of them (pondered by a vector of coefficients that sum up to one) that meets all constraints. If the achieved set or sets of scores meet all restrictions, then the solution of the optimization has been achieved. Otherwise, proceed to the next and last step.
4. Solve the quadratic optimization problem as formulates in Equation 10.9, using as 'seeds' or 'starting solutions' the same as in the previous step, and/or the quasi-solution in the previous step.

It is important to note that this procedure is useful as long as no non-linear hard equality constraints are imposed on the solution, in which case caution is advised.

Chapter 11
Two real case studies of optimization in the latent space

11.1. Introduction

In Chapter 9 the procedure to transfer the restriction on the original space (both inputs and outputs) to the common subspace of the PLS-regression model was provided, as well as the definition of the DS, the NS as an estimation of it, and how to estimate the subspace most likely to contain (and least likely to fall out of) the projection of the TDS when PLS-regression is resorted to. Secondly, Chapter 10 provided different tools and methods to efficiently approach the optimization problem in the latent space depending on the goal and constraints imposed on its solution. In this chapter, two real case studies will be presented that make use of most of the concepts and methods illustrated in Chapters 9 and 10, and a comparison of the different applicable approaches to the corresponding optimization problems will be carried out.

11.2. Methods

Linear PLS-regression, explained in Section 3.2.2, will be resorted to, as well as the methods for the transferal of constraints imposed on the original variables onto the latent space, and the definition of analytical expression of the NS both for an output and for a lineal combination of outputs, explained in Chapter 9, and the different approaches to the optimization problem detailed in Chapter 10.

11.3. Datasets

Although for confidentiality reasons the origin and details on the two real datasets used in this chapter cannot be disclosed, the most relevant information concerning them is provided in this section.

11.3.1. Case study 1: minimizing two output variables simultaneously

The dataset used for this example consists of a matrix of inputs with 6609 hourly measurements made of 63 input variables, and a matrix of outputs with the same number of measurements of 2 output variables. The output variables are strongly correlated with each other.

11.3.2. Case study 2: maximizing a linear combination of outputs

The dataset used for this example consists of a matrix of inputs with 2075 hourly measurements made of 135 input variables, and a matrix of outputs with the same number of measurements of 10 output variables. All output variables are strongly correlated among each other, some positively and some negatively, and must meet a hard equality constraint that affects all of them.

11.4. Results and discussion

11.4.1. Case study 1: minimizing two output variables simultaneously

Given the available data, and after a cross validation exercise to select the optimum number of latent variables, a PLS-regression model with $A = 8$ LV was fitted capable of explaining ~ 61% of the variability of the inputs and ~78% of the variability of the outputs. The feasible absolute minima for the two (pre-treated) output variables, y_1 and y_2 are known to be -5.98 and -8.59. Therefore, if the vector of desired values for the outputs is defined as $\mathbf{y}_{\mathrm{DES}}^{\mathrm{T}} = [-5.98 \quad -8.59]$, the NS corresponding to these values can be also defined, which when represented on the space of the latent variables can be visualized as in Figure 11.1.

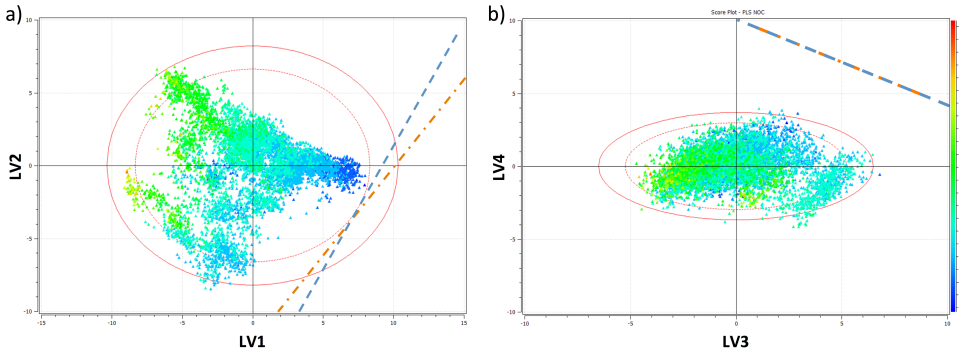


Figure 11.1. Case study 1: Graphical representation of the projection of the calibration dataset (with the observations coloured according to the value of $\boldsymbol{y_1}$) and the intersection of the NS for $\mathbf{y}_{\mathbf{DES}}^{\mathbf{T}} = [-5.98 \quad -8.59]$ associated to the first (blue dashed line) and second (orange dashed line) output variables with a) the hyperplane defined by the two first LV and b) the hyperplane defined by the third and fourth LV

To illustrate the importance of properly formulating the optimization problem, the soft constraints (in this case $\mathbf{y}_{\text{DES}}$ and the restriction on the Hotelling T^2) and the hard constraints (in this case $\mathbf{A}_{\boldsymbol{\tau}}$ and $\mathbf{d}_{\boldsymbol{\tau}}$ as a consequence of restrictions on the inputs, and again the restriction on the Hotelling T^2), assume that the formulation in Equation 10.9 is resorted to, and the following scenarios are considered:

i. $g_{0\text{a}} = 1$ and $g_{0\text{b}} = g_1 = g_2 = 0$; $\boldsymbol{\Gamma} = \mathbf{I}_2$ and $\mathbf{y}_{\text{DES}}^{\text{T}} = [-5.98 \quad -8.59]$; and no hard constraints imposed on any variable, nor on the Hotelling T^2 of the solution.

ii. $g_{0\text{a}} = 1$, $g_1 = 1$ and $g_{0\text{b}} = g_2 = 0$; $\boldsymbol{\Gamma} = \mathbf{I}_2$ and $\mathbf{y}_{\text{DES}}^{\text{T}} = [-5.98 \quad -8.59]$; and hard constraints imposed on the inputs so that the solution will not be outside of the historical operating conditions (defined as the univariate restrictions corresponding to the lowest and highest values for the inputs), but without any hard restriction on the Hotelling T^2 of the solution.

iii. $g_{0\text{a}} = 1$, $g_1 = 1$ and $g_{0\text{b}} = g_2 = 0$; $\boldsymbol{\Gamma} = \mathbf{I}_2$ and $\mathbf{y}_{\text{DES}}^{\text{T}} = [-5.98 \quad -8.59]$; and hard constraints imposed on the inputs so that the solution will not be outside of the historical operating conditions, as well as on the Hotelling T^2 of the solution (so that its Hotelling T^2 is not above its corresponding 95% confidence limit).

The projection of the solutions of these optimization problems are shown in Figure 11.2, where it can be seen that only the third scenario (iii) provides a set of inputs that guarantee that no extrapolation outside of the Hotelling T^2 hyperellipsoid takes place.

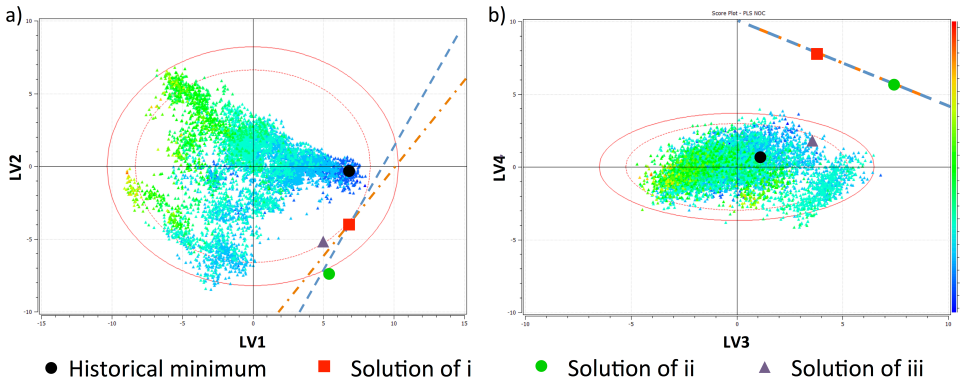


Figure 11.2. Case study 1: Graphical representation of the projection of solutions of the optimization problems in scenarios 'i' (red square), 'ii' (green circle) and 'iii' (violet triangle), as well as the projection of the set of process conditions corresponding to the historical minima for both outputs on a) the hyperplane defined by the two first LV and b) the hyperplane defined by the third and fourth LV

For each of these solutions, the vector $\hat{\mathbf{y}}_{\mathrm{NEW}}^{\mathrm{T}}$ results:

i. $\hat{\mathbf{y}}_{\mathrm{NEW}}^{\mathrm{T}} = \mathbf{y}_{\mathrm{DES}}^{\mathrm{T}} = [-5.98 \quad -8.59]$
ii. $\hat{\mathbf{y}}_{\mathrm{NEW}}^{\mathrm{T}} = [-5.82 \quad -7.02]$
iii. $\hat{\mathbf{y}}_{\mathrm{NEW}}^{\mathrm{T}} = [-3.18 \quad -3.09]$

While the historical minimum is $\mathbf{y}_{\mathrm{min}}^{\mathrm{T}} = [-2.04 \quad -2.63]$

It is important to note the presence of a gradient in the values of the outputs, clearly visible in Figures 11.1a and 11.2b, with the lowest values apparently found for higher values in the first LV and lower values in the second one, which is also the least explored region of the latent space. In fact, all of the solutions obtained fall exactly within this area. Additionally, although the first and second solution seem to be significantly better than the third one in terms of the predicted values for both outputs, the third one is the only reliable one, since it is the only one that falls within the Hotelling T^2 confidence hyperellipsoid and fulfils the restriction of being within the subspace of historical operating conditions. Furthermore, this solution also provides, in theory, better results than the set of process conditions corresponding to the historical minimum.

With respect to the efficiency of the different optimization algorithms presented in Chapter 10, the quadratic optimization algorithm using the formulation in Equation 10.9 provided these solutions consistently (i.e. always the same ones) and quickly (i.e. less than one second of computational time). The linear optimization formulation in Equation 10.15 could not provide, by itself, this same solutions, given that the restriction on the Hotelling T^2 is not redundant, but the slight variation proposed immediately after did. On the other hand, the sequential optimization approach proposed in Section 10.4 was able to provide exactly the same solutions as the quadratic optimization formulation in ~1/10th of that time, by ending in the second of its four steps.

11.4.2. Case study 2: maximizing a linear combination of outputs

After a cross validation exercise to select the optimum number of latent variables, a PLS-regression model with $A = 14$ LV was fitted capable of explaining ~ 75.5% of the variability of the inputs and ~77.7% of the variability of the outputs. The maximum observable value for the quality attribute of interest, which is a linear combination of all of the outputs, was $d_{\mathrm{max_{obs}}} = 171.58$.

Figure 11.3 shows the projection of the calibration dataset onto the subspace of the PLS-regression model for the eight first LV and the direction of maximum variability for the quality attribute of interest.

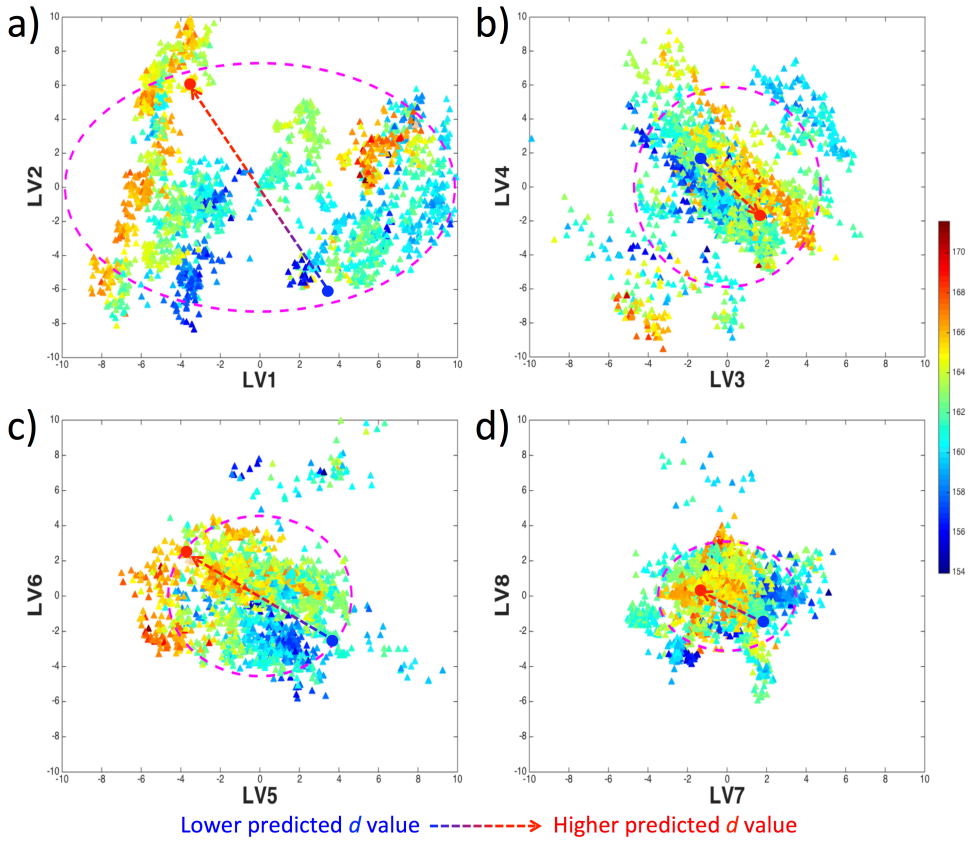


Figure 11.3. Case study 2: Graphical representation of the projection of the calibration dataset (with the observations coloured according to the value of the quality attribute of interest) and the direction of maximum variability for the quality attribute to maximize in a) the hyperplane defined by LV 1 and 2; b) the hyperplane defined by LV 3 and 4; c) the hyperplane defined by LV 5 and 6; and d) the hyperplane defined by LV 7 and 8

In this example, 4 of the 135 inputs are required to meet specific values in the solution. To illustrate the complexity and some of the issues with this optimization exercise, three different scenarios will be considered in which successively more constraints are imposed as in the first case study. Using the formulation in Equation 10.9:

i. $g_{0a} = 1$, $g_1 = 1$ and $g_{0b} = g_2 = 0$; $\mathbf{\Gamma} = 1$ and $d_{\text{DES}} = 174.66$ (i.e. the maximum achievable solution inside the Hotelling T^2 confidence hyperellipsoid without accounting for any other restrictions, determined as explained in Section 10.2.3.3); and no hard constraints imposed on the inputs.

ii. $g_{0a} = 1$, $g_1 = 1$ and $g_{0b} = g_2 = 0$; $\mathbf{\Gamma} = 1$ and $d_{\mathrm{DES}} = 174.66$; and hard inequality constraints imposed on the inputs so that the solution will not be outside of the historical operating conditions.

iii. $g_{0a} = 1$, $g_1 = 1$ and $g_{0b} = g_2 = 0$; $\mathbf{\Gamma} = 1$ and $d_{\mathrm{DES}} = 174.66$; and hard inequality constraints imposed on the inputs so that the solution will not be outside of the historical operating conditions, in addition to the 4 hard equality constraints imposed on the inputs that must meet certain values.

In all of these scenarios the hard constraint on the Hotelling T^2 value for the solution is imposed to guarantee that it is inside the corresponding confidence hyperellipsoid, as well as the hard equality constraint on the outputs.

Figure 11.4 shows the solutions for all of these scenarios, as well as the observation of the calibration dataset for which the historical maximum was observed. For each of these solutions, $\hat{d}_{\mathrm{NEW}}$ results:

i. $\hat{d}_{\mathrm{NEW}} = 174.66$
ii. $\hat{d}_{\mathrm{NEW}} = 165.41$
iii. $\hat{d}_{\mathrm{NEW}} = 165.11$

While the historical maximum was $d_{\mathrm{max_{obs}}} = 171.58$, it must be noted that the historical maximum for the quality attribute of interest among all observations with the input variables meeting the equality constraints imposed during the optimization was $d^*_{\mathrm{max_{obs}}} = 159.34$. Therefore, although this solution offers no improvement with respect to the historical maximum, a significant improvement is theoretically achieved with respect to the maximum observed given the restriction on the four inputs to meet specific values.

Another important aspect of this optimization problem is the fact that the computational cost of the quadratic optimization formulation increased dramatically with respect to the previous example, specially if the 'seeds' provided to it were not defined carefully, which also lead to inconsistent solutions (i.e. the algorithm would provide a different, actually not globally optimal solution most times it was executed). In particular, providing to the quadratic programing algorithm in Matlab 'seeds' that already met all equality and inequality hard constraints significantly improved its performance, reducing the computational time from ~1.5-4 hours (depending on the computer) to ~5-15 minutes, although slight (and arguably not significant) differences in the solutions were observed among solutions. By contrast, using the approach proposed in Section 10.4 reduced the computational time to less than ~30 seconds, while providing consistent solutions every execution.

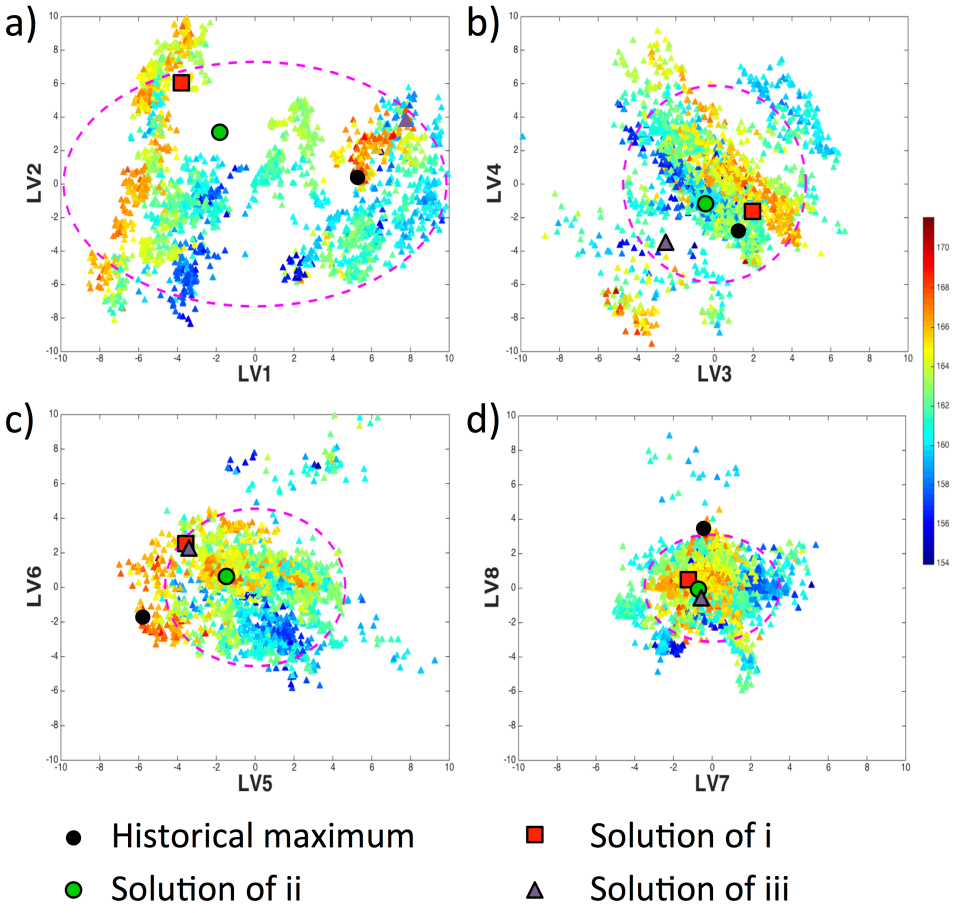


Figure 11.4. Case study 2: Graphical representation of the projection of solutions of the optimization problems in scenarios 'i' (red square), 'ii' (green circle) and 'iii' (violet triangle), as well as the projection of the set of process conditions corresponding to the historical maximum for the quality attribute of interest on a) the hyperplane defined by LV 1 and 2; b) the hyperplane defined by LV 3 and 4; c) the hyperplane defined by LV 5 and 6; and d) the hyperplane defined by LV 7 and 8

11.5. Conclusions

By applying the different optimization approaches proposed in Chapter 10 to two real case studies, it was possible to assess some of the issues they present, the importance of properly defining and imposing restrictions, and the limitations of the quadratic optimization formulation in terms of computational efficiency and convergence issues when the dimension of the space of acceptable solutions increases and several hard equality constraints are imposed, as well as how a proper definition of the 'seeds' pro-

vided to this algorithm can help mitigate some of them. Lastly, the approach proposed in Section 10.4 was observed to outperform the use of the quadratic optimization formulation in Equation 10.9 in such scenario.

It is also worth noting that two of the terms in the objective function for the quadratic optimization formulation were given, in both examples, null weight. These terms are precisely the ones detailed in Section 10.2.2, and proposed in order to be used for exploratory purposes and for DOE in the latent space. Since this was not the goal of any of the cases presented in this section, and a single solution was desired each time, these terms were not necessary. They will, however, be resorted to in Chapter 12.

Chapter 12
On experimentation to improve the design space estimation

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

1. Palací-López, D., Facco, P., Barolo, M. & Ferrer, A. Sequential experimental approach to improve the design space estimation using latent-variable model inversion. Part II. Optimization problem reformulation and sequential experimental approach. *Submitted.*

12.1. Introduction

Chapter 9 focused on the definition of the DS, its relationship with the KS, and how to estimate the subspace most likely to contain the projection of the TDS (as well as the least likely to fall outside of it) in the latent space when PLS-regression is resorted to. On the other hand, Chapter 10 focused on how to efficiently approach the optimization problem in the latent space depending on the goal and the nature of the constraints imposed on the solution of the optimization. Finally, Chapter 11 served to illustrate two real case studies where some of the concepts addressed in Chapters 9 and 10 where applied in order to optimize different quality attributes of interest. None of them, however, aims at solving/minimizing the impact of one of the main issues that is always present when resorting to data-based approaches for the definition of the DS or for optimization purposes: the uncertainty. In this chapter, a sequential experimental approach aimed at reducing the uncertainty in the estimation of the DS is proposed. Reducing the error in the estimation of the DS should result in a smaller subspace of the knowledge space being chosen to bracket the DS, and will improve the accuracy in the definition of different combinations of inputs to obtain the desired outputs.

12.2. Methods

OLS and PLS-regression techniques, explained in Section 3.2.2, will be resorted to, as well as the general concepts of D-optimal DOE addressed in Section 2.2.1, the definition of the envelope of the experimental region discussed in Section 5.2.2, the estimation of the subspace most likely to contain the DS as explained in Chapter 9, and the optimization approach detailed in Chapter 10. This taken into account, three proposed approaches will now be illustrated:

- Method 1: OLS and DOE in the space of the original variables; this method is proposed here in order to establish a comparison between an approach based on 'traditional' methods and PLS-based methods;
- Method 2: combines linear PLS regression model fitting with DOE in the latent space and LVRMI;
- Method 3: differs with the second one in the substitution of the 'explicit' DOE with an 'implicit' one via optimization in the latent space

These methods aim at improving the DS estimation when specific values of the outputs (or quality attributes) are desired. When a range of values are considered acceptable, the procedure can be used to achieve a better estimations of the DS corresponding to the values for the LSL and USL defining such range. If the goal is to minimize/maximize a quality attribute, but a feasible minimum/maximum cannot be clearly defined, this method can also be used to estimate the DS corresponding to the maximum/minimum admissible value. The two PLS-based methods will also allow in some cases, because of its sequential nature, the exploration of poorly known areas of the process, where the DS is (according to the model) expected to lay, specially when a quality attribute of interest is to be minimized/maximized.

12.2.1. Method 1: Classical DOE and OLS model inversion approach

The first method is based on traditional approaches for model fitting such as OLS, and as such happenstance data cannot be used. Instead, data from DOE must be resorted to for its application in order to infer causality and invert the model afterwards. These data shall be pre-treated to get rid of severe outliers, as well as apply the necessary transformations (logarithms, inverse terms, etc…) before fitting a model via OLS. Any improvements in the DS estimate are expected to occur when a more appropriate model structure is chosen and a better estimation of its corresponding parameters is achieved. This method is formulated in such a way that these two points are taken into account, while also proposing a way to evaluate the accuracy in the estimation of the DS and its improvement after subsequent experimentation.

Figure 12.1 shows the different steps to follow for the application of this method.

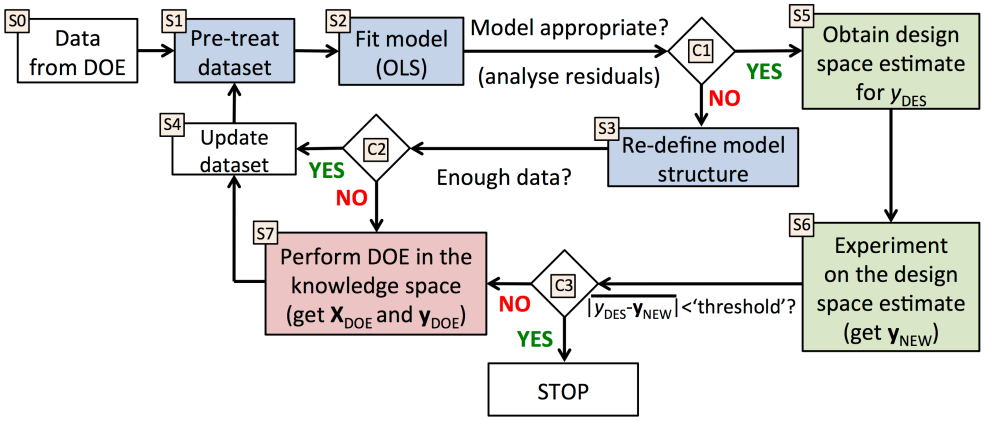


Figure 12.1. Method 1: Classical Design of Experiments and OLS model inversion-based sequential approach to improve the estimation of the design space. Blocks identified with S0, S1, etc. correspond to different 'steps' of the approach, while those identified as C1, C2 or C3 are 'checkpoints' where the decision is made to continue one route or another or stop altogether

Three main groups of steps have been highlighted: model structure definition and model fitting (blue; steps S1, S2 and S3), DS estimate exploration (green; steps S5 and S6) and experimentation to improve the DS estimation (red; step S7). A brief explanation of each of these steps in one iteration (i.e. a cycle starting at step S1 and ending at S4, or once it is decided that there is no need to continue at checkpoint 3), for a general scenario with a single output/quality attribute involved[xiv], is as follows:

0. Step S0: this method permits only the use of data from which causality can be inferred between the inputs and outputs used to fit and invert the OLS regression model. Therefore happenstance/historical data cannot be resorted to.
1. Step S1: additional terms (i.e. exponential or logarithmic terms, interaction terms among input variables, quadratic and higher order terms, etc.) are included or discarded (e.g. due to not being considered relevant to the study) if deemed convenient prior to the next step. Missing data and/or outliers are also appropriately dealt with, if necessary.
2. Step S2: an OLS-regression model is fitted relating the pre-treated matrix of inputs with the vector of observations for the pre-treated output variable/quality attribute of interest.
3. Checkpoint C1: the statistical significance of the different terms included in the fitted model is analysed. The residuals from fitting the regression model are also evaluated to detect outliers and assess if more complex or different model structures should be considered.

[xiv] If more than one output or quality attribute is considered, each iteration should be repeated for every output variable.

4. Step S3: if necessary, the model structure is redefined.
5. Checkpoint C2: if enough data is available to fit the newly defined model, step S4 follows. Otherwise, step S7 goes next.
6. Step S4: the dataset is updated by removing, if necessary, observations considered to be outliers, or adding new observations (if the previous step was S7)
7. Step S5: the DS estimate is computed by simply inverting the OLS model, and refers to the subspace constituted by all possible combinations of inputs that, according to the fitted model, lead to the desired value for the quality attribute (y_{DES} for an output, or d_{DES} for a quality attribute of interest)
8. Step S6: the experimental region is comprised by the section of this DS estimate within the KS, i.e. within the limits imposed by any defined constraints on the inputs. Such KS, also mentioned in step S7, can be defined by following the procedure to obtain the envelope of the experimental region discussed in Section 5.2.2. Experimentation is carried out for a total of n_{DS} samples distributed as uniformly as possible along this DS estimate, obtaining $\mathbf{y}_{\mathrm{NEW}}$ (or $\mathbf{d}_{\mathrm{NEW}}$), a $[n_{\mathrm{DS}}\times 1]$ vector of values of the output (or quality attribute of interest) for n_{DS} observations.
9. Checkpoint C3: the values in $\mathbf{y}_{\mathrm{NEW}}$ (or $\mathbf{d}_{\mathrm{NEW}}$) are compared to y_{DES} (or d_{DES}). Then the 'average distance to the DS', or *ADDS* (defined as $\overline{|y_{\mathrm{DES}} - \mathbf{y}_{\mathrm{NEW}}|}$ in the block diagrams in Figures 12.1, 12.2 and 12.3), is estimated as:

$$ADDS = \frac{1}{n_{\mathrm{DS}}} \cdot \sum_{i=1}^{n_{\mathrm{DS}}} \left| y_{\mathrm{DES}} - y_{\mathrm{NEW},i} \right| \tag{12.1}$$

   This value is compared to a 'threshold' representing the maximum allowed 'average deviation' from the TDS when resorting to its estimation via model inversion. A value of *ADDS* below this limit would indicate that no more iterations are required, since a good enough estimation of the DS has been achieved. Otherwise, step S7 follows. Note that the *ADDS* has the same units as the quality attribute of interest, and provides a measure of the expected average deviation of the values of this quality attribute along the estimated DS from the desired one. It must also be noted that an adaptation to OLS of the methodology to assess the adequacy of a PLS-regression model for inversion (see Section 9.6) is not suitable here, because that assessment concerns how good a model is for inversion in general, while here the only interest is in the determination of the DS for a specific y_{DES}.
10. Step S7: since, when using this method, any improvement in the DS estimate is expected to occur as a consequence of a proper selection of the model structure and a better estimation of its parameters, the D-optimality criterion is suggested for initial/subsequent experimentation, especially if a lot of restrictions are imposed on the process variables in the definition of the experimental region. If only a few process variables are involved, and little restrictions are

imposed, implementing a full factorial design may be feasible, but this is not expected to be the case in most scenarios. A total o n_{DOE} new experiments are carried out, resulting in $\mathbf{X}_{\mathrm{DOE}}$ $[n_{\mathrm{DOE}} \times M]$ and $\mathbf{y}_{\mathrm{DOE}}$ $[n_{\mathrm{DOE}} \times 1]$, which would be added to previous data for the next iteration.

This method presents some drawbacks:

i) since data from DOE is required for its application, happenstance data cannot be resorted to;

ii) using OLS requires resorting to either variable selection or more extensive experimentation as the number of variables and the complexity of the fitted model increases. Since a candidate set of potential experiments has to be defined in the DOE step of the algorithm, computational cost increases along with the dimensionality of the problem;

iii) regression methods such as OLS suffer from correlation among the regressors, which is usually present in real, complex problems (and, unavoidably, in data from mixture design problems). Furthermore, if multiple response variables are involved their corresponding models must be fitted one at a time, and the information regarding any existing correlation among them is lost.

12.2.2. Method 2: DOE in the latent space and LVRMI approach

Although this method follows a similar path as Method 1, since now PLS-regression based methods are used, happenstance data (e.g. historical data) can be exploited in order to infer causality and later invert the model. Because in this study linear PLS is resorted to, these data must still be pre-treated as done in Method 1 prior to fitting a PLS-regression model.

In practice, it is not uncommon that a large number of inputs are involved, and when making use of PLS-based methods the initial dataset is usually not expanded to account for possible interaction or non-linear terms among inputs before proceeding to fit the corresponding LVRM. Failing to account for these non-linearities in the formulation of the model may cause the LVRM to not be appropriate for predictive nor optimization purposes in presence of severe non-linearities, even if its overall performance along the knowledge space seems reasonable. In such situation local modeling techniques [105] may be more appropriate. A similar issue may arise when trying to obtain a good estimation of the DS, since then the main goal is to accurately define a subspace of the KS, and not as much the construction of a model with good performance along the whole KS. This method, the steps of which are illustrated in Figure 12.2, is formulated so this is taken into account.

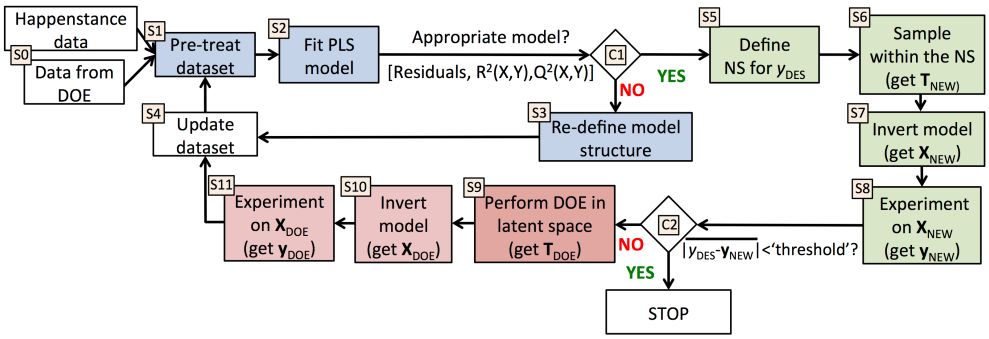


Figure 12.2. Method 2: latent variable regression model inversion plus Design of Experiments in the latent space-based sequential approach to improve the estimation of the design space

A brief explanation of the different steps in one iteration is presented below:

0. Steps S0 and S1 are equivalent to the same steps with Method 1.
1. Step S2: a PLS-regression model is fitted relating the pre-treated matrix of inputs with the matrix of observations for the pre-treated output variables/quality attributes of interest. It is important to note here that a single PLS-regression model may suffice in this case to account for all outputs/quality attributes of interest simultaneously.
2. Checkpoint C1: the residuals from fitting the PLS-regression model are evaluated to detect outliers and assess if more complex or different model structures should be considered. The model performance for prediction (of both matrices of inputs and outputs) and optimization purposes is also assessed.
3. Step S3 and S4 are equivalent to those in Method 1.
4. Step S5: the NS here refers to the projection in the latent space of the DS estimate corresponding to the desired value for the quality attribute of interest (y_{DES} or d_{DES}), whose general expression is provided by Equation 9.26. This subspace is further delimited by the constraints imposed on the input and output variables, whose projection onto the latent space is done via Equation 9.11 and 9.13, as well as the Hotelling T^2 confidence hyperellipsoid.
5. Step S6: $\mathbf{T}_{\mathrm{NEW}}$ in this step is a $[n_{\mathrm{DS}} \times A]$ matrix with the scores of n_{DS} samples distributed as uniformly as possible along the NS as defined in step S5.
6. Step S7: $\mathbf{X}_{\mathrm{NEW}}$ is a $[n_{\mathrm{DS}} \times M]$ matrix obtained from $\mathbf{T}_{\mathrm{NEW}}$ through the inversion of the PLS-regression model.
7. Step S8: experimentation is carried out for the n_{DS} samples defined in $\mathbf{X}_{\mathrm{NEW}}$, obtaining $\mathbf{y}_{\mathrm{NEW}}$ (or $\mathbf{d}_{\mathrm{NEW}}$), a $[n_{\mathrm{DS}} \times 1]$ vector of values of the output (or quality attribute of interest) for those n_{DS} observations.
8. Checkpoint C2: this checkpoint is equivalent to C3 in Method 1.
9. Step 9: a space filling DOE is constructed in the latent space via e.g. the Kennard-Stone approach [106] as it allows uniformly spanning the area of feasible experimentation, given the maximum number of experiments that can be per-

formed (namely, the resources that the experimenter could invest in terms of time and cost), without having to define any particular model structure (i.e. the 'optimality criterion' of this sort of design does not depend on the model to be fit afterwards). With regards to the experimental region, the subspace more likely to contain the TDS, as defined in Section 9.4, is chosen for this method. This way, n_{DOE} different sets of scores are defined, and the matrix $\mathbf{T}_{\mathrm{DOE}}$ $[n_{\mathrm{DOE}} \times A]$ is obtained.

10. Steps S10 and S11 are carried out as steps S7 and S8, but starting from the matrix $\mathbf{T}_{\mathrm{DOE}}$, and obtaining $\mathbf{X}_{\mathrm{DOE}}$ $[n_{\mathrm{DOE}} \times M]$ and $\mathbf{y}_{\mathrm{NEW}}$ (or $\mathbf{d}_{\mathrm{NEW}}$) $[n_{\mathrm{DOE}} \times 1]$. These new data would be added to previous data for the next iteration.

The main drawback of this method is that it requires defining a candidate set for the DOE in step S9 in order to apply the Kennard-Stone algorithm (or any point-exchange algorithm), and therefore the computational cost increases according to the dimensionality of the latent space. However, since this dimensionality is not expected to be too high in most cases, the computational cost will rarely be a real hindrance in practice.

12.2.3. Method 3: Optimization in the latent space and LVRMI approach

This third method is almost identical to Method 2, being based on the same concepts. For the purpose of comparison, Figure 12.3 shows the different steps to follow for the application of this method.

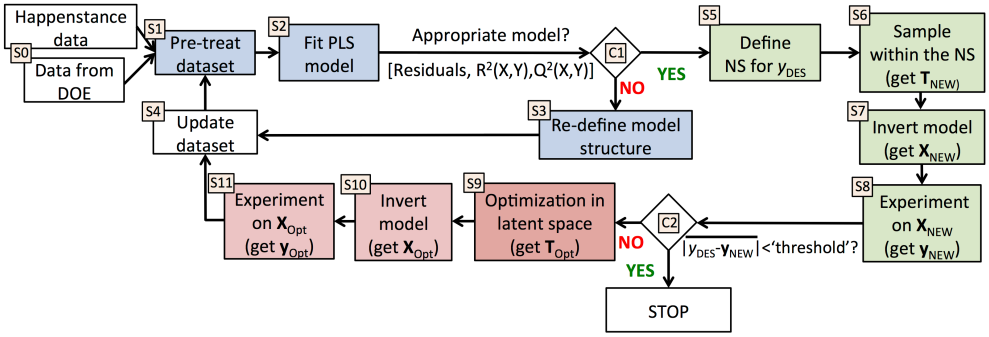


Figure 12.3. Method 3: latent variable regression model inversion plus optimization in the latent space-based sequential approach to improve the estimation of the design space

As it can be seen in Figure 12.3, the only appreciable difference with respect to Method 2 lays in step S9, where the optimization problem as formulated in Equation 10.9 (Section 10.2.2) is resorted to. To obtain n_{Opt} different solutions, the set of scores obtained as a solution of the optimization problem each time is added to the matrix of scores corresponding to the observations used to fit the PLS-regression model, before solving the optimization problem again, until it has been solved n_{Opt} times (to obtain the last set of scores, $n_{\mathrm{Opt}} - 1$ different sets of scores will have been concatenated to the PLS

scores matrix before carrying out the optimization one last time). Note that, because of the way this optimization problem is formulated, there are two elements that set this method apart from Method 2:

i) although the experimental space is the same, 'soft constraints' are also included in this method in the form of the various terms of the objective function. This implies that the new experiments to perform in each iteration, if any, will correspond to solutions of the optimization problem that are not simply required to fall within the experimental space, but will also be 'pushed' towards the NS and the centre of projection.

ii) there is no need to define a candidate set for experimentation, which makes Method 3 less dependant on the proper definition of such candidate set, and less computationally costly.

On the other hand, and as pointed out in Section 10.2.2 (see Figure 10.1), the contribution of each term to the objective function during the optimization step should be assessed in order to properly select the weight give to each of these terms.

12.3. Datasets

12.3.1. Case study 1: mathematical model

For the first case study, the same dataset as the one illustrated in Section 9.4.1.1 will be resorted to. As a reminder, this dataset is constituted by 6 observations from a D-optimal DOE where the data were simulated according to the model in Equation 8.1, and x_1 and x_2 are treated as manipulated variables, while x_3, x_4 and x_5 are three measured variables, with y being the quality attribute of interest. Furthermore, 5% of the variability of each variable is added as noise to the values of x_3, x_4, x_5 and y, both for the generation of the initial dataset and for the simulation of future samples.

12.3.2. Case study 2: simulated Vinyl-Chloride Monomer manufacturing

For the second case study, the same dataset as the one illustrated in Section 9.4.1.3 will be resorted to. As a reminder, this dataset contains 20 simulated samples corresponding to the simulated production of ethylene dichloride (EDC) as an intermediate product to VCM according to the information detailed in [101] in Pro/II. To allow some variability in the process without loss of validity of any assumption required for the simulation, manipulated variables were made to vary independently a maximum of 5% around the assigned values in [101] for the generation of the dataset. Additionally, 5% of the variability of each variable is added as noise to the values of all non-manipulated variables, both for the generation of the initial dataset and for the simulation of future samples. It must be noted that, although the simulated dataset is not a result of a DOE, the data used has been generated following first-principle models and by changing manipulated variables independently from one another. Therefore causality can be inferred in the

relationship among the controlled input variables and the outputs involved. As a consequence, Method 1 can still be applied in this case, since only the controlled variables are accounted for when fitting and inverting the OLS-regression model, and for the construction of the DOE in the original space.

12.4. Results and discussion

The mathematical example will be used to illustrate the three methods in an easier, more intuitive way, as well as to mention their main advantages and limitations. The purpose of the second example is to show how these same methods can be used in scenarios with a higher complexity and practical relevance.

12.4.1. Case study 1: mathematical model

As mentioned in Section 12.3.1, an initial dataset with 6 samples from a DOE is used for the application of all three methods. This dataset includes four samples located in the vertices defined by the univariate limits imposed on x_1 and x_2, which are the two only manipulable variables and therefore the factors used to construct the DOE (with the lower and upper univariate limits corresponding to their respective low and high factor levels). Two replicates of the overall centroid of the KS in the original variables are also included in the DOE. This corresponds to a D-optimal DOE for a model with linear and two-factor interaction terms, plus the two added centroid replicates.

For illustration purposes, Section 12.4.1.1 details the procedure used for the application of each of the proposed methods to this case study, and in Section 12.4.1.2 the results achieved through the use of each algorithm are compared. Some additional considerations are discussed in Section 12.4.1.3.

12.4.1.1 Detailed procedure for the application of all three methods

For the application of Method 1, the detailed procedure is as follows:

1. Given the initial dataset (step S0), the most simple regression model, the linear polynomial, is fitted using x_1 and x_2 (step S2). This reveals (checkpoint C1) that non-linearities may be present, and since enough data is available to do so the decision is made to re-define the model as the quadratic polynomial (S3).
2. The fit of the quadratic model indicates significantly better results, and is therefore selected as the appropriate model. Coincidentally, this implies that, from now on, variables x_1 and x_2, their interaction and their corresponding quadratic terms are used to fit a model (step S2) with the same structure as the mathematical model used to simulate the data.
3. In step S3, the NS for $y_{\mathrm{DES}} = 204.86$ is almost coincident with the TDS (the only difference resulting from the presence of noise in the data). It is important to note that in almost any other practical situation this will not happen, since

the true relationship between manipulable and observed input variables is likely to be more complex, and not included in the fitted regression model.

4. In step S5, the matrix $\mathbf{X}_{\mathrm{NEW}}$ is built by taking 100 samples uniformly spanning the NS (as an estimate of the DS).
5. The DOE in step S7 is constructed on x_1 and x_2, since variables x_3 to x_5 cannot be freely manipulated when taking new samples. The DOE consists of a follow-up D-optimal design for a model with all linear, first-order interaction and quadratic terms, considering the already available samples. A sufficiently large candidate set of uniformly distributed points inside the knowledge space to span it 'in its entirety' (10000 in this case) is generated, and the point exchange algorithm [7,107,108] is used to select which experiments to perform. Although a matrix ($\mathbf{X}_{\mathrm{DOE}}$) is presented in figure 12.1, indicating that usually more than one experiment will be carried out, in this example a single new experiment ($\mathbf{x}_{\mathrm{DOE}}$) is considered.
6. To update the dataset prior to the next iteration of the method, the newly acquired sample is added to the dataset used at the beginning of this iteration.

For the application of Method 2, the detailed procedure is:

1. An initial assessment of the relationship among the manipulated and measured input variables reveals a very strong correlation between x_3 and x_1^2, x_4 and x_2^2, and x_5 and $x_1 \cdot x_2$. Because of this, the restrictions $x_3 = x_1^2$, $x_4 = x_2^2$ and $x_5 = x_1 \cdot x_2$ are applied later on for the estimation of the DS (this being an approximation, since differences exist as a consequence of the noise in the data)
2. Data is mean-centred and scaled to unit variance in step S1.
3. A PLS-regression model with 2 LV is fitted in order to allow easier graphical representation of some of the results. Since 4 LV would be a much better option according to the leave-one-out cross validation method (Q^2(2LV)=0.336; Q^2(4LV)=0.955), a second PLS-regression model with 4 LV is also fitted to illustrate additional results afterwards.
4. In step S6, 100 sets of scores meeting the condition that that $x_3 = x_1^2$, $x_4 = x_2^2$ and $x_5 = x_1 \cdot x_2$, and spanning the NS as uniformly as possible are sampled ($\mathbf{T}_{\mathrm{NEW}}$), and through model inversion $\mathbf{X}_{\mathrm{NEW}}$ is obtained in step S7.
5. The DOE in step S9 is a follow-up space filling design using the Kennard-Stone algorithm after defining a sufficiently large candidate dataset (containing again 10000 potential samples in this case) of sets of scores distributed as uniformly as possible inside the subspace most likely to contain the TDS (as explained in Chapter 9) for $y_{\mathrm{DES}} = 204.86$. As with Method 1, although the matrix $\mathbf{T}_{\mathrm{DOE}}$ in Figure 12.2 implies the definition of more than one set of scores, in this example only one ($\boldsymbol{\tau}_{\mathrm{DOE}}$) is obtained each iteration.
6. Dataset updating in step S4 before the next iteration is done by simply adding $\mathbf{x}_{\mathrm{DOE}}$ and y_{DOE} to the dataset used at the beginning of the current iteration.

For the application of Method 3, the detailed procedure is almost identical to the one used for Method 2. The only notable difference lies in step 9, where the optimization problem as formulated in Equation 10.9 is resorted to, in which all the lower and upper limits in Table 9.3 are imposed as hard inequality restrictions on the solution, once transferred to the latent space, and $g_{0a} = 0$ and $g_{0b} = g_1 = g_2 = 1$. Note that here $g_{0a} = 0$ and $g_{0b} = 1$ since the proposed sequential experimental approach aims at exploring the subspace most likely to contain the TDS and as a consequence, given two sets of scores at the same Euclidean distance from the NS, it is desired to penalize less the one whose point in the NS closest to it has higher leverage (i.e. higher uncertainty in the NS estimation around it).

12.4.1.2 Assessing the performance of each of the three methods

To assess the performance of each of the three methods, the evolution of the similarities between the TDS and the DS estimate can be visualized for e.g. 30 subsequent iterations. In Figure 12.4 such visualization can be done for the 10th and 30th iteration, where a clear improvement in the DS estimation is achieved by using Method 1 (i.e. the amplitude of the corresponding confidence intervals is visibly reduced), which is to be expected since the structure of the model being fitted matches perfectly the hypothetical input-output causal relationship presented in Equation 8.1, and so the discrepancies between the NS and the TDS will be due mostly to the noise in the data, whose effect gets diminished as more observations are available to fit the regression model. Method 2 and Method 3 also led to a better DS estimation, but their corresponding representations cannot be directly compared to those for Method 1, and the accomplished improvement is not visually as clear. Another severe drawback of using Figure 12.4 to evaluate the performance of the algorithms presented in Section 12.2 is that, although it can be used in this particular example to assess the performance of all three methods, such thing will not be possible in most practical situations, since the DS is not known a priori (which is precisely the reason why these methods are proposed: to reduce the uncertainty in the DS estimation).

A distinction is made in Figure 12.4 between the 'theoretical new experiment' and the 'feasible new experiment' at each iteration. The difference between one and the other is that the former corresponds to $\mathbf{x}_{\mathrm{DOE}}$ ($\boldsymbol{\tau}_{\mathrm{DOE}}$) as obtained in step S7 for Method 1 or S10 for Method 2, or to $\mathbf{x}_{\mathrm{Opt}}$ ($\boldsymbol{\tau}_{\mathrm{Opt}}$) in step S10 for Method 3, which implicitly assumes some values for variables x_3, x_4 and x_5, while the later corresponds to the vector inputs/scores with the same values for x_1 and x_2, but the 'real' measured values for x_3, x_4 and x_5.

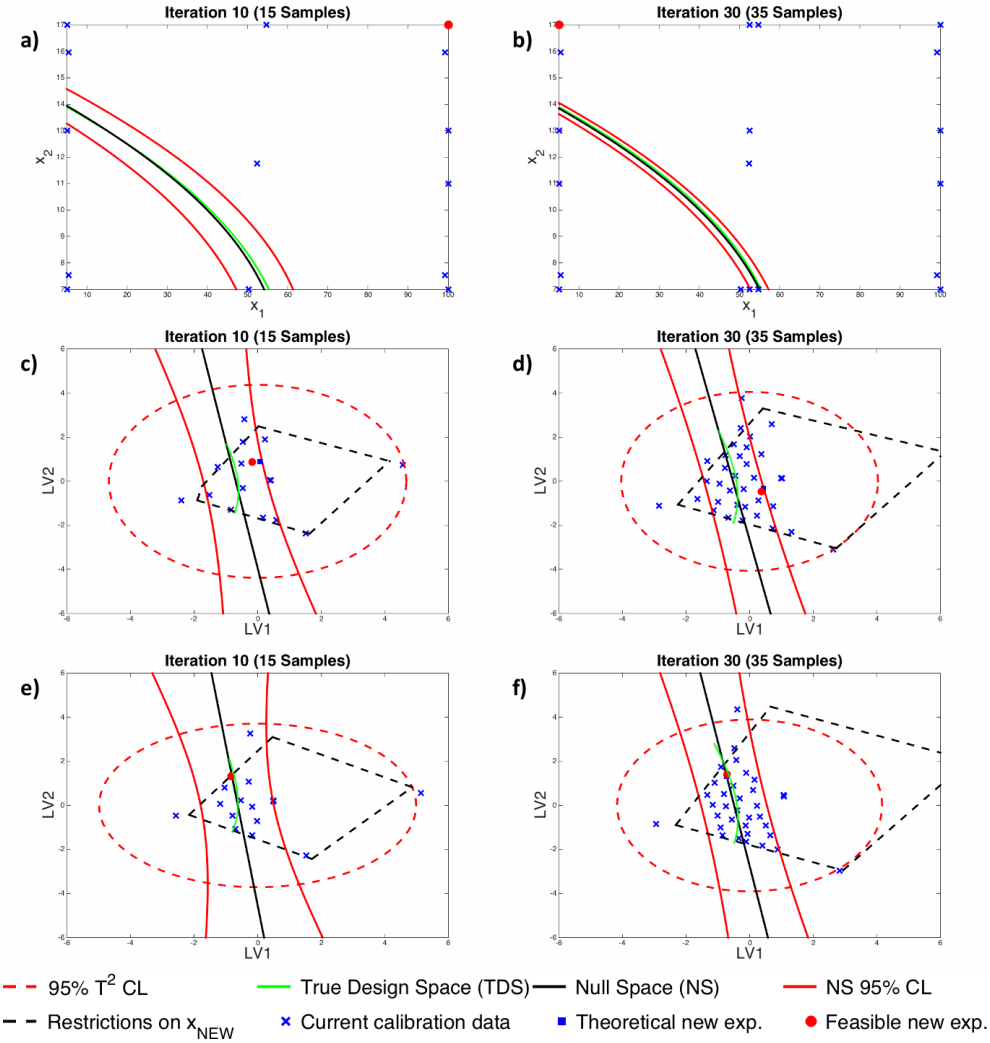


Figure 12.4. Graphical representation of the calibration dataset, TDS and NS with its corresponding 95% confidence limits for Method 1 (Figures 12.4.a and 12.4.b), Method 2 (Figures 12.4.c and 12.4.d) and Method 3 (Figures 12.4.e and 12.4.f) after 9 (Figures 12.4.a, 12.4.c and 12.4.e) and 29 (Figures 12.4.b, 12.4.d and 12.4.f) additional experiments have been performed. All of the methods are applied starting from an initial calibration dataset with 6 samples affected by 5% noise on x_3, x_4, x_5 and y, and only one experiment is performed per iteration, with the experimentation carried out within the KS for Method 1 and within the subspace most likely to contain the TDS as defined in Chapter 9 for Method 2 and Method 3. Dataset updating consists in just adding each new observation.

Alternatively, the value of the *ADDS* can also be observed for the same 30 subsequent iterations. It should be noted that, in spite of the way the *ADDS* has been defined, such concept may have at least, the following two interpretations:

a) Average absolute distance from each possible combination of process inputs that actually provide the desired values for the quality attributes of interest (process inputs from the DS) to the closest theoretical combination of process inputs leading to the same values for these quality attributes (process inputs from the NS).

b) Average absolute difference between the desired values for the quality attributes of interest and those achieved by operating the process under the different combinations of process inputs from the DS estimate.

The first of these two interpretations, however, requires the DS to be known, and is therefore not suitable for the purpose of this study. Because of this, the second meaning is applied in this work when referring to the *ADDS*.

Figure 12.5 shows the average of these absolute differences per iteration for each algorithm.

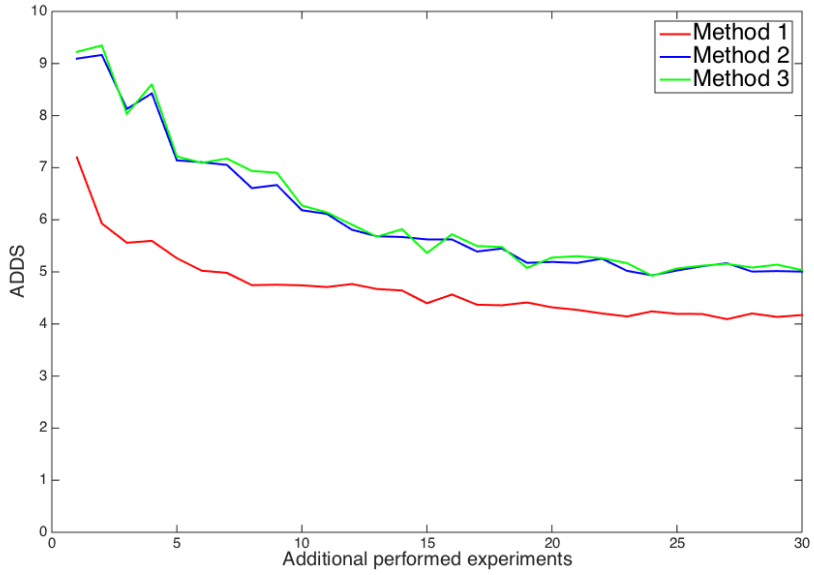


Figure 12.5. Case study 1: Average distance to the DS (*ADDS*), defined in Equation 12.1, for $n_{\mathrm{DS}} = 100$**,** plotted against the number of iterations (coincident, in this case, with the number of additional performed experiments) using all three methods, with a PLS-regression model fitted with 4 LV for Method 2 and Method 3.

It must be noted from Figure 12.5 that almost no significant improvement seems to be achieved after 5 additional experiments are performed with Method 1, nor after 10 additional experiments for Method 2 or Method 3. Since random noise was added when simulating the data, and to assess if observable differences in Figure 12.5 may be statistically significant among different methods (specially Method 2 and Method 3), Method 1, Method 2 and Method 3 were applied 100 times, with 4 LV variables chosen to fit the PLS-regression model in Method 2 and Method 3. 100 curves as the ones shown in Figure 12.5 are obtained per algorithm, and a model with the following structure is fitted with this data:

$$\begin{aligned} ADDS = & b_0 + b_1 \cdot i + b_2 \cdot m_1 + b_3 \cdot m_3 + b_4 \cdot i \cdot m_1 + b_5 \cdot i \cdot m_3 + \\ & + b_6 \cdot i^2 + b_7 \cdot i^2 \cdot m_1 + b_8 \cdot i^2 \cdot m_3 \end{aligned} \tag{12.2}$$

Where i is the number of the corresponding iteration (in Figure 12.5, additional performed experiments), and m_1 and m_3 are dummy variables that take the value 1 when Method 1 and Method 3 are used, respectively. To assess potential statistically significant differences, the data corresponding to all iterations for each method are used to fit the model in Equation 12.2. Values for b_2, b_4, and b_7 significantly different from 0 would signal statistically significant differences between Method 1 and Method 2, while non-zero significant values for b_3, b_5, and b_8 would point out statistically significant differences between Method 3 and Method 2. Table 12.1 shows the results from such model fitting, where only statistically significant effects remain in the model.

Table 12.1. Case Study 1: Study of potential significant differences among methods in terms of the accuracy in the estimation of the DS

Parameter	Estimation	p-value
b_0	9.38	0.0000
b_1	-0.37	0.0000
b_2	-3.03	0.0000
b_4	0.18	0.0001
b_6	$7.81 \cdot 10^{-3}$	0.0000
b_7	$-3.75 \cdot 10^{-3}$	0.0076

As shown in Table 12.1, no statistically significant differences were found between Method 2 and Method 3. As expected, differences between Method 1 and Method 2 are statistically significant, with Method 1 improving the estimation of the DS at a lower rate ($b_4 > 0$) compared to Method 2 and Method 3. It should also be noted that the initial *ADDS* is very different when using Method 1, being statistically significantly lower than when using Method 2 or Method 3 ($b_2 < 0$). Additionally, although 100

samples were taken along the NS in this example ($n_{\mathrm{DS}} = 100$), taking as little as 3 of them ($n_{\mathrm{DS}} = 3$), as uniformly distributed as possible along the DS estimate (which was a line in this case) provided very similar numerical results and the same conclusions.

These results seem to show an overall better performance of Method 1 given the simulated data for the input-output causal relationship in Equation 8.1. However, three points should be made to this respect, which greatly bias the results shown here in favour of Method 1:

i. since all variables affecting the quality attribute of interest (*y*) were used to fit the regression model, the only expected discrepancies between the expected and the real response would be those caused by the noise introduced into the data. Furthermore, four LV were chosen to fit the PLS regression models with Method 2 and Method 3, as suggested by the cross-validation. Had five LV been selected, any statistically significant differences between Method 2 or 3 and Method 1 would have disappeared;
ii. a good estimation of the DS and experimentation in this DS estimate was possible when using Method 1 only because the quadratic model structure including x_1, x_2, their square terms and interaction was a good approximation of the true model structure (it was, in fact the exact same model structure used to simulate the data). However, expecting such favourable results would be unwise in most practical situations, where a large number of variables are involved, and more complex relationships among them exist. This is because detecting the existence of this sort of relationships among them would be difficult and require extensive experimentation, and only manipulable factors can be considered when Method 1 is resorted to;
iii. in most of the experiments performed during the application of Method 1, the resulting values for the quality attribute of interest were not close to the desired one, while the opposite was true with Method 2 and Method 3. See as an example Figures 12.4.a and 12.4.b, for Method 1, where most of the samples used to fit the regression model (which are selected during the DOE step) are located at the extremes of the KS (hence far away from the TDS), while in figures 12.4.c to 12.4.f, for Method 2 and 3, most of these samples (also selected during the DOE/Optimization step) are located close to the TDS. This means that, in practice, Method 2 and Method 3 are more likely to provide products satisfying the specifications on the quality attribute of interest for which a better estimation of the corresponding TDS is desired.

Note that the first of these items also highlights the importance of properly selecting the number of latent variables to extract when fitting a PLS-regression model when it is to be used in its inverse form, when applying Method 2 and Method 3. Although the cross-validation approach suggested 4 LV to be extracted, this resulted in results worse than those for the OLS-based approach, when 5 LV would have resulted in identical outcomes.

12.4.1.3 Additional considerations

Several factors affect these algorithms, and their effect on the performance of each method was also assessed to evaluate their robustness. Some of these factors and the results from this study are listed below:

i. Noise in the data:

Regarding the effect of the amount of noise, the same procedure was carried out with 0%, 1% and 10% of the natural variability of the different variables added as random noise to the data. All three methods were affected, increasing the amount of experimentation required to achieve the same level of accuracy in the DS estimate as more noise was present, which was to be expected, but a clear improvement still occurred with relatively little experimentation.

ii. Size of the initial dataset and level of dispersion around the TDS:

With respect to the size of the initial dataset, the larger the amount of initial samples was, the better the initial DS estimation, and the slowest was the improvement achieved by using Method 2 and Method 3, although Method 1 was not affected by this. Furthermore, for a given size of the initial dataset and a set amount of noise, the DS estimate with Method 2 and Method 3 was always better whenever the samples were located closer to the DS. However, if all initial samples were selected such that they were located exactly on the DS corresponding to a specific value for the quality attribute of interest, no regression model could be fitted (no matter the method), as expected, since no variation in the output/quality attribute would exist in the dataset, in spite of any variations in the inputs (i.e. it would look as if variations in the inputs cause no variation in the output/quality attribute, and therefore no relationship exist among them, despite this not being true).

iii. Number of experiments performed per iteration and dataset updating method:

Instead of performing one single additional experiment per iteration (step S7 for Method 1, or steps S9-S11 for Method 2 and Method 3), the same procedure was followed, but 2, 3 or 5 new samples were taken per iteration. No significant differences were found for Method 1, but the performance of Method 2 and Method 3 were deteriorating as more new samples were taken per iteration due to the fact that more iterations were needed to achieved the same performance in terms of ADDS. This deterioration was more critical as the amount of noise present in the data was incremented.

For the dataset updating method with Method 2 and Method 3 (step S4), eliminating samples that provided values for y ($\mathbf{y}_{\mathrm{DOE}}$ or $\mathbf{y}_{\mathrm{Opt}}$) 'too far from y_{DES}', as an alternative to simply adding the new experiments performed, was also tested. This significantly improved the results achieved if done in the first iteration. However, being too strict when using this criterion at the beginning or during subsequent iterations led to PLS regression models with a high uncertainty in the prediction of y, which would back-

propagate during the model inversion step. Ultimately, none of these two methods provided desirable results in such scenario.

iv. Severity of the non-linearities in the process:

Finally, regarding the severity of the non-linearities in the process, the first quadratic term ($x_3 = x_1^2$) in the mathematical model was substituted by an inverse exponential term ($x_3 = e^{-x_1}$), an exponential term ($x_3 = e^{x_1}$) and a logarithmic term ($x_3 = log(x_1)$) to account for three different cases to the original one. In all of them, Method 2 and Method 3 led to an improvement in the DS estimation only when a sufficiently narrow range of variability was allowed for x_1. Large ranges of variability for x_1, on the other hand, led to scenarios where the DS estimation would improve or worsen apparently at random from one iteration to the next. On the other hand, Method 1 did not provide any significant improvement, but no instability in the accuracy of the DS estimate was observed either.

12.4.2. Case study 2: simulated Vinyl-Chloride Monomer manufacturing

As done in Section 9.4.2.3, and previous to the application of the proposed methods, the initial fitting of the corresponding models relating input variables x_1 to x_8 and z_1 with the output variables y_1 and y_2 via PLS was observed to lead to models with relatively poor explanatory and predictive capability with regards to y_2, which is the most relevant quality attribute of interest in this example. Because of this z_1 was transformed into a new variable $x_9 = z_1 \cdot x_8$, and y_2 is substituted by $y_3 = y_1 \cdot y_2$. By doing so, a PLS regression models fitted with these 'new' variables can be obtained that provides much better explanatory and predictive capabilities (R^2(**Y**)>0.965 and R^2(**X**)>0.863 when using five LVs).

Since now the quality attribute of interest, y_2, is no longer included when fitting the PLS-regression model, however, the NS for a linear combination of outputs (y_1 and y_3) must be defined following the procedure detailed in Section 9.3.3, such that for a given desired value for y_2, $y_{2,\mathrm{DES}}$, such linear combination must satisfy Equation 9.33. The value $y_{2,\mathrm{DES}} = 0.979$ is used in this case as the *minimum* acceptable purity for EDC4 (i.e. the DS is defined in this case as the subspace comprised by all possible combinations of inputs guaranteeing at least this purity, and therefore determining the subspace of inputs guaranteeing exactly it is equivalent to defining the frontier of such DS).

12.4.2.1 Detailed procedure for the application of all three methods

For the application of Method 1, the detailed procedure is as follows:

1. Since only manipulated variables (x_1 to x_5) can be used for the application of Method 1, and in order to account for possible non-linearities (fitting a second order polynomial model would be preferable to fitting a first order one), an initial evaluation of the variables most likely to influence the process outputs was

carried out. This was done via stepwise OLS regression considering different combinations of manipulated variables. As a result, only variables x_1, x_4 and x_5 were chosen as factors with a significant impact on y_1 and y_3. The values for x_2 and x_3, which were not found to have a significant effect on the two quality attribute of interests according to the respective OLS regression models, were fixed. Specifically, x_2 was set to $x_2 = 0.996 \cdot x_1$, and x_3 was set to its lower bound (see Table 9.7).

2. Given the initial dataset (step S0), only the aforementioned selected variables x_1, x_4 and x_5 are used in step S1 to construct the matrix of inputs used to fit the OLS-regression models, together with two-factor interaction and second order terms, to fit the OLS-regression models for y_1 and y_3. This model structure is selected for both models (step S2).
3. Equation 9.33 is used to combine both OLS-regression models fitted in step S2 so that the DS estimate can be obtained for $y_{2,\mathrm{DES}} = 0.979$ in step S3.
4. In step S5, the matrix $\mathbf{X}_{\mathrm{NEW}}$ is built taking 12 samples uniformly spanning the DS estimate as obtained in step S3. Values for x_2, x_3 are computed from them as mentioned in the previous item of this list.
5. The DOE in step S7 is constructed on x_1, x_4 and x_5, since variables x_2, x_3 and x_6 to x_8 cannot be freely manipulated when taking new samples. The DOE consists of a follow-up D-optimal design for a model with all linear, first-order interaction and quadratic terms, considering the already available samples. A sufficiently large candidate set of uniformly distributed points inside the knowledge space to span it 'in its entirety' (10000 in this case) is generated, and the point exchange algorithm is used to select a single new experiment ($\mathbf{x}_{\mathrm{DOE}}$) to carry out.
6. To update the dataset prior to the next iteration of the method, the newly acquired sample is added to the dataset used at the beginning of this iteration.

For the application of Method 2, the detailed procedure is:

1. Data, including all input variables (as opposed to Method 1) and y_1 and y_3, is mean-centred and scaled to unit variance by rows in step S1.
2. The linear PLS-regression technique is used to fit a PLS-regression model with 5 LV.
3. In step S6, 12 sets of scores spanning the NS as uniformly as possible are sampled ($\mathbf{T}_{\mathrm{NEW}}$), and through model inversion $\mathbf{X}_{\mathrm{NEW}}$ is obtained in step S7. The 'experimentation' in step S8 is done to obtain the feasible or 'real' $\mathbf{X}_{\mathrm{NEW}}$ as well as $\mathbf{y}_{\mathrm{NEW}}$.
4. The DOE in step S9 is a follow-up space filling design using the Kennard-Stone algorithm after defining a sufficiently large candidate dataset (containing again 10000 potential samples in this case) of sets of scores distributed as uniformly as possible inside the subspace most likely to contain the TDS (as

explained in Chapter 9) for $y_{2,\mathrm{DES}} = 0.979$, from which a single new set of scores $\boldsymbol{\tau}_{\mathrm{DOE}}$ is selected.

5. $\mathbf{x}_{\mathrm{DOE}}$ is obtained through model inversion in step S10, and the 'real' $\mathbf{x}_{\mathrm{DOE}}$ and y_{DOE} are obtained in step S11.
6. Dataset updating in step S4 before the next iteration is done by simply adding $\mathbf{x}_{\mathrm{DOE}}$ and y_{DOE} to the dataset used at the beginning of the current iteration.

The same differentiation made for the first case study, with regards to the application of Method 3 with respect to Method 2, can be done here.

12.4.2.2 Assessing the performance of each of the three methods

In this case the DS is not known, and the number of variables involved (either in the original or the latent space) would not allow for an easily interpretable graphical representation of the DS estimate, such as the one in Figure 12.4. In its place, and similarly to Figure 12.5, Figure 12.6 shows the *ADDS* for $y_{2,\mathrm{DES}} = 0.979$ per iteration for each algorithm.

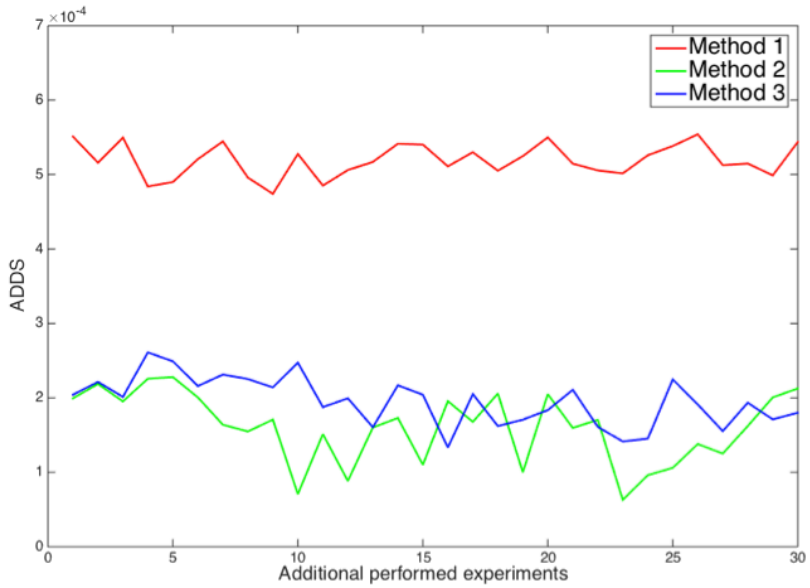


Figure 12.6. Case study 2: Average distance to the DS (*ADDS*), defined in Equation 12.1, for $n_{\mathrm{DS}} = 12$**,** plotted against the number of iterations (coincident, in this case, with the number of additional performed experiments) using all three methods.

The same procedure as with the first case study was carried out in order to identify statistically significant differences among the methodologies, carrying out all three

methods thrice with 4 different datasets (for a total of 12 times per method). A model with the same structure as the one in Equation 12.2 was fitted, and the statistical significance of its terms assessed. Table 12.2 shows the results obtained from this evaluation.

Table 12.2. Case Study 2: Study of potential significant differences among methods in terms of the accuracy in the estimation of the DS

Parameter	Estimation	p-value
b_0	$1.96 \cdot 10^{-4}$	0.0000
b_1	$-2.95 \cdot 10^{-7}$	0.0459
b_2	$3 \cdot 10^{-4}$	0.0000
b_4	$9.37 \cdot 10^{-7}$	0.0475

Again, values for b_2, b_4, and b_7 different from 0 would signal statistically significant differences between Method 1 and Method 2, while non-zero values for b_3, b_5, and b_8 would point out statistically significant differences between Method 2 and Method 3. Taking this into account, Table 12.2 seems to point at statistically significant differences between Method 1 and the other two algorithms, with statistically significant differences in the rate of improvement in the accuracy in the definition of the DS, according to which Method 1 is outperformed by the other two (although in practice this difference may be negligible).

A more obvious statistically significant difference is observed between Method 1 and Method 2 in terms of the 'baseline' *ADDS*. This is in part due to PLS-regression methods being able to capture a larger part of the variability in the process when compared to OLS-regression methods. Note, with respect to this, that only manipulated variables can be taken into account for the application of Method 1, which also contributes to this difference.

On the other hand, the statistical significance of b_1 and its negative value signal a decrease of the *ADDS* when resorting to Method 2 and Method 3. However, this improvement in the accuracy in the estimation of the DS is probably not relevant in practice. This is related to the size of the initial dataset when applying these algorithms. Note that, for the first case study, the initial dataset was comprised by six samples, and no apparent significant decrease of the *ADDS* was observed after 10 or 15 additional experiments were performed (i.e. when the PLS-regression model was fitted with 16 or 21 experiments). In this second case study, the three methods were applied with an initial dataset with 20 observations. It may then be suspected that no improvement is observed in the accuracy of the DS estimate because it is already as good as it can be. In fact, Method 1 cannot be applied with an initial dataset with as few as 5 or 6 obser-

vations, but Method 2 and Method 3 show, in this scenario, a fast initial decrease of the *ADDS* similar to that seen in Figure 12.5 (not shown).

Finally, and similarly as in the first case study, values very different to $y_{2,\mathrm{DES}} = 0.979$ where obtained for the quality attribute of interest (molar rate of EDC in stream EDC4) in most of the experiments having to be carried out with Method 1 (from 0.9778 to 0.9809), being the opposite true with Method 2 (from 0.9781 to 0.9793) and Method 3 (from 0.9789 to 0.9792).

12.5. Conclusions

Three different algorithms have been presented aiming at the improvement in the estimation of the DS, one of which is based on traditional approaches such as OLS-regression and classical DOE in the space of the original variables, and the other two in the concepts of LVRMI and DOE and optimization in the latent space.

The performance of these three methods has been tested in two case studies, and the following has been concluded:

a) The first method (OLS-based) may work well whenever the problem is simple, well known and easy to model. However, it is outperformed by the other two methods if this is not the case. Furthermore, happenstance (e.g. historical) data cannot be exploited and further experimentation requires working under processing conditions that will most probably not lead to the desired values for the quality attribute of interest. On the other hand, the other two methods (PLS-based) allow historical data to be used and obtaining new processing conditions that provide values close to the desired ones for the quality attribute of interest even while improving the accuracy of the DS estimate.

b) The sensitivity to noise, the size of the initial dataset and the level of dispersion of the samples in this dataset around the DS has been assessed. Regarding this, the two PLS-based methods provided better initial results when the initial samples were located close to the DS, and a better rate of improvement in the estimation of the DS was observed when i) few observations close to the DS were present in the initial dataset, and ii) a smaller initial dataset, as opposed to a larger one, was available. This is most probably in line with the idea of local modeling whenever more or less severe non-linearities are present in the 'true' model structure.

c) All of the algorithms were observed to be quite sensitive to the presence of severe non-linearities in the process, and are negatively affected if this is the case.

With respect to point *b*, it must be pointed out that the datasets used in the two examples illustrated in Sections 5 and 6 were comprised by observations located more or less uniformly along the whole KS. In most practical scenarios, where happenstance

data is resorted to, the TDS may be located in unexplored areas of the process, and the proposed methods are expected to provide a way to iteratively explore them and improve their estimation. This, however, is yet to be assessed in a real case study.

Finally, it is important to consider that only Method 2 and Method 3 can make complete use of the available data through the latent space, since both manipulated and non-manipulated variables can be accounted for when fitting a PLS-regression model and inverting it. This is not possible with OLS-based approaches.

PART IV

Epilogue

Chapter 13
Conclusions and perspectives

13.1. Accomplishment of the objectives

Given the results shown in previous chapters, the degree of accomplishment of the different objectives formulated in Section 1.2 can be assessed from the following general conclusions.

13.1.1. *Objective I - Traditional and latent variable-based approaches applied to mixture design problems*

Chapter 5 allowed assessing most of the complexities associated to the traditional approaches for mixture design of experiments and model definition and fitting strategies, also presenting some proposals for the definition of the experimental region and the construction of DOEs in irregular mixture/experimental spaces (i.e. both applicable not only to mixture problems, but also to mixture-process variable and/or mixture-amount problems).

On the other hand, Chapter 6 illustrated some of the most relevant advantages in the use of latent variable-based regression techniques in terms of interpretability and easiness of usage, including a novel approach making use of K-PLS and pseudo-samples trajectories for the retrieval, whenever applicable, of a mixture model's Scheffé coefficients. Some of the advantages presented by these methods include:

- latent variable-based regression methods allow the study of the effect of mixture and non-mixture variables simultaneously without having to treat them differently;

- latent variable-based methods such as PLS allow an easier/more efficiently detection of equally contributing mixture constituents in terms of their effect in the properties of a blend;
- while, given mixture data not affected by severe nonlinearities and/or with a high number of observations, PLS and K-PLS with pseudo-sample trajectories yielded very similar results to classical Scheffé model fitting by means of OLS, when more severe non-linearities and/or a small data structures had to be analysed, K-PLS proved to be a valid alternative for overcoming the main limitation of both OLS and PLS in terms of model-fitting for prediction purposes;
- furthermore, a way to recover the parameters of a Scheffé model from the trend of the pseudo-sample trajectories was also derived and validated via a simulated case-study when using K-PLS, provided that fitting such model makes sense;
- finally, although the performance of the OLS- and PLS-based methodologies was shown to improve in some cases by taking into account inverse terms, K-PLS did not suffer from some of the drawbacks they did, such the lack of enough degrees of freedom for a stable estimation of the coefficients of these augmented models, while (for RBF K-PLS) allowing different types of complex nonlinear relationships to be modelled, making it specially suitable when combinations of unknown non-linearities affect the sort of interdependence between constituent proportions and output variables.

Lastly, a user-friendly software tool was developed to allow easily assessing the differences in performance of several of the methodologies and algorithms illustrated in Chapters 5 and 6.

13.1.2. *Objective II - Latent variable-based approaches for efficient processes optimization*

Chapter 10 in this Thesis presented one of the most commonly resorted to algorithms for process optimization using latent variable-based approaches, the quadratic optimization formulation, as well as some of its limitations or shortcomings as carried out in the literature. Then some extensions and proposed ways to standardize the way the optimization can be made in different scenarios were provided, in order to tackle them, such as:

- the optimization problem as proposed in the literature was extended to include the possibility of quality attributes being expressed as linear combination of outputs, after demonstrating the impossibility of formulating such optimization problem for such quality attribute in terms of a multiple-objective problem concerning the different outputs separately (without the matrix of weights for them losing its original meaning);

- a discussion was presented regarding the importance of providing proper weights to each of the terms in the objective function in the quadratic formulation of the optimization problem, as well as the kind of (both soft and hard) constraints to be imposed on the solution depending on the goal. In addition, two new soft constraints were proposed which can be useful for DOE and exploration purposes in the latent space, and one of which can also be resorted to in order to guarantee that different yet acceptable solutions may be achieved when the optimization it done sequentially/more than once without modifying any other restrictions;
- when addressing a problem where one or more quality attributes are to be maximized or minimized, a method to objectively determine the values one should provide as the "desired ones" has been developed and illustrated, which may also be used, in some cases, to determine the solution of the optimization problem in a much more efficient way that resorting to the quadratic or linear formulations themselves, and that helps in avoiding involuntarily assigning excessive/insufficient importance (with respect to the intended one) to maximizing or minimizing some quality attribute or output;
- lastly, a sequential approach to the optimization problem has been proposed and demonstrated to be more efficient than directly resorting to the quadratic or even the linear formulation, while avoiding (or at the very least mitigating) the issues some of those present with regards to the convergence to the global optimum in a given space of acceptable solutions.

The application of some of these contributions has been shown for two case studies in Chapter 11 and, to a lesser extent, another two in Chapter 12.

13.1.3. *Objective III - Latent variable-based approaches applied to the Quality by Design initiative, to increase processes flexibility and guarantee the desired quality*

Chapter 9 deals with some basic concepts regarding the transferal of restrictions on the original space to the latent space and their connection to the Knowledge Space, as well as the way in which the computation of the direct inversion, as a way to find a set of inputs (if any) that theoretically guarantees the desired quality, and the so-called null space, as a way to obtain several of them (if theoretically possible), have been presented in past literature. In this context:

- the explicit formulation of the projection of the restrictions on the original variables onto the latent space has been shown, and a new analytical formulation of the null space has been provided and extended to any quality attributes of interest that can be expressed as a linear combination of outputs;

- a novel formulation of the confidence region for the NS, which coupled to the transferral of constraints onto the latent space allows a more robust estimation of the subspace most likely to contain the TDS with respect to past proposals, as well as the estimation of the subspace less likely to fall outside of the TDS, has been developed, with the application of the first having been illustrated with three simulated case studies;
- finally, a novel sequential experimental approach to increase the accuracy in the estimation of the DS has been developed and applied to two simulated case studies in Chapter 12, and its robustness tested against several factors, presenting this way an efficient way to perform experimentation that provides products with quality characteristics close to the desired ones while simultaneously helping better define the DS.

13.2. Future research lines

The results achieved through the completion of this Ph.D. thesis open new and relevant perspectives that may merit further consideration in the near future:

- Development and potential improvement of already existing algorithms for the analysis of mixture data including raw material properties.
- Extension and/or adaptation of the different presented and proposed algorithms for the constructions of DOEs for mixture design problems (including mixture-process variables and mixture-amount problems) to more rigorously define DOE construction methods in the latent space and for mixture problems including the properties of the raw materials in L/T-shaped datasets. This may be achieved by e.g. considering the relationship between raw material properties and scores in the latent subspace of models constructed as in [63–65].
- Extension of the optimization and DS estimation methodologies coupled to model-updating algorithms for process optimization outside of the KS, and for situations with frequent changes in raw materials, process conditions, etc.
- Application of the idea behind the definition of the subspace least likely to fall outside of the DS to the definition of regions of acceptable raw material properties, and to the assessment of a production process capability.
- For mixture design including raw materials' properties, and batch processes:
    - Extension of the proposed approach for the estimation of the subspace most likely to contain (and least likely to fall outside of) the DS.
    - Further development of the proposed optimization formulation.
    - Improvement of methodologies to increase the accuracy in the estimation of the DS, and in the assessment of such accuracy.

Chapter 14
Appendices

14.1. Annex to Part I

14.1.1. Relationship between the Euclidean distance matrix, $\mathbf{D}$, and the inner product matrix, $\mathbf{X} \cdot \mathbf{X}^{\mathrm{T}}$

The Euclidean distance between two observations, $\mathbf{x}_n^{\mathrm{T}}$ and $\mathbf{x}_{n^*}^{\mathrm{T}}$, of a generic dataset, $\mathbf{X}$ $[N \times J]$, can be expressed as:

$$d_{n,n^*} = \|\mathbf{x}_n - \mathbf{x}_{n^*}\|^2 = (\mathbf{x}_n - \mathbf{x}_{n^*})^{\mathrm{T}} \cdot (\mathbf{x}_n - \mathbf{x}_{n^*}) = \mathbf{x}_n{}^{\mathrm{T}} \cdot \mathbf{x}_n + \mathbf{x}_{n^*}{}^{\mathrm{T}} \cdot \mathbf{x}_{n^*} - 2 \cdot \mathbf{x}_n{}^{\mathrm{T}} \cdot \mathbf{x}_{n^*} \quad (14.1)$$

Let $\mathbf{F}$ $[N \times N]$ be the inner product matrix and $\mathbf{D}$ $[N \times N]$ the Euclidean distance matrix, defined as:

$$\begin{aligned} \mathbf{F} &= \mathbf{X} \cdot \mathbf{X}^{\mathrm{T}} \\ \mathbf{D} &= \mathbf{f} \cdot \mathbf{1}^{\mathrm{T}} + \mathbf{1} \cdot \mathbf{f}^{\mathrm{T}} - 2 \cdot \mathbf{F} \end{aligned} \quad (14.2)$$

where $\mathbf{f}$ $[N \times 1]$ denotes the diagonal vector of $\mathbf{F}$ and $\mathbf{1}$ $[N \times 1]$ is a vector of ones.

Centring $\mathbf{X}$ such that:

$$\overline{\mathbf{X}} = \mathbf{X} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{X} \quad (14.3)$$

then:

$$\begin{aligned} \overline{\mathbf{F}} = \overline{\mathbf{X}} \cdot \overline{\mathbf{X}}^{\mathrm{T}} &= \left(\mathbf{X} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{X}\right) \cdot \left(\mathbf{X} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{X}\right)^{\mathrm{T}} = \\ &= \mathbf{F} - \frac{1}{N} \cdot \mathbf{F} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{F} + \frac{1}{N^2} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{F} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \end{aligned} \quad (14.4)$$

If **D** is double-centred as:

$$\mathbf{B} = -\frac{1}{2} \cdot \mathbf{H} \cdot \mathbf{D} \cdot \mathbf{H}^{\mathrm{T}} \tag{14.5}$$

where **H** represents the operator

$$\mathbf{H} = \mathbf{I} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \tag{14.6}$$

and **I** is the $[N \times N]$ identity matrix, it follows:

$$\mathbf{B} = -\frac{1}{2} \cdot \mathbf{H} \cdot (\mathbf{f} \cdot \mathbf{1}^{\mathrm{T}} + \mathbf{1} \cdot \mathbf{f}^{\mathrm{T}} - 2 \cdot \mathbf{F}) \cdot \mathbf{H}^{\mathrm{T}} \tag{14.7}$$

Since:

$$\mathbf{f} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{H}^{\mathrm{T}} = \mathbf{f} \cdot \mathbf{1}^{\mathrm{T}} \cdot \left(\mathbf{I} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{f} \cdot \mathbf{1}^{\mathrm{T}} - \mathbf{f} \cdot \frac{\mathbf{1}^{\mathrm{T}} \cdot \mathbf{1}}{N} \cdot \mathbf{1}^{\mathrm{T}} = \mathbf{0} \tag{14.8}$$

it is verified:

$$\mathbf{H} \cdot \mathbf{f} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{H}^{\mathrm{T}} = \mathbf{0} = \mathbf{H} \cdot \mathbf{1} \cdot \mathbf{f}^{\mathrm{T}} \cdot \mathbf{H}^{\mathrm{T}} \tag{14.9}$$

Therefore:

$$\begin{aligned} \mathbf{B} = \mathbf{H} \cdot \mathbf{F} \cdot \mathbf{H}^{\mathrm{T}} = \left(\mathbf{I} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}}\right) \cdot \mathbf{F} \cdot \left(\mathbf{I} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}} = \\ = \mathbf{F} - \frac{1}{N} \cdot \mathbf{F} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} - \frac{1}{N} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{F} + \frac{1}{N^2} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} \cdot \mathbf{F} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} = \bar{\mathbf{F}} \end{aligned} \tag{14.10}$$

that is:

$$\mathbf{B} = -\frac{1}{2} \cdot \mathbf{H} \cdot \mathbf{D} \cdot \mathbf{H}^{\mathrm{T}} = \bar{\mathbf{X}} \cdot \bar{\mathbf{X}}^{\mathrm{T}} \tag{14.11}$$

And so it has been demonstrated that the double-centred Euclidean distance matrix **B** is equal to the inner product matrix, $\bar{\mathbf{X}} \cdot \bar{\mathbf{X}}^{\mathrm{T}}$.

14.1.2. Practical meaning of the pseudo-samples in the feature space

Consider that $\mathbf{B} = \bar{\mathbf{X}} \cdot \bar{\mathbf{X}}^{\mathrm{T}}$ has been used for calibrating a 1-latent variable PLS model. The scores of the *N* objects under study, $\mathbf{t}^{\mathbf{B}}$ $[N \times 1]$, can be written as:

$$\mathbf{t}^{\mathbf{B}} = \mathbf{B} \cdot \mathbf{w}^{*\mathbf{B}} \tag{14.12}$$

where $\mathbf{w}^{*\mathbf{B}}$ $[N \times 1]$ represents the PLS vector of weights, which, in this case, does not contain any useful information about the *J* original variables in **X** . Substituting Equation 14.11 in Equation 14.12 it follows:

$$\mathbf{t}^{\mathbf{B}} = \bar{\mathbf{X}} \cdot \bar{\mathbf{X}}^{\mathrm{T}} \cdot \mathbf{w}^{*\mathbf{B}} = \bar{\mathbf{X}} \cdot \mathbf{w}^{*\prime} \tag{14.13}$$

where $\mathbf{w}^{*\prime}$ $[J\times 1]$ is now relevant for interpretation purposes. Projecting the $[1\times J]$ vector of pseudo-samples $\mathbf{g}^{\mathrm{T}} = [0,\ 0,\ ...,\ 1\ \ 0,\ 0,\ 0]$ onto the PLS model subspace:

$$t_{\mathbf{g}^{\mathrm{T}}} = \mathbf{g}^{\mathrm{T}} \cdot \overline{\mathbf{X}}^{\mathrm{T}} \cdot \mathbf{w}^{*\mathbf{B}} = \mathbf{g}^{\mathrm{T}} \cdot \mathbf{w}^{*\prime} = w_j^{*\prime} \tag{14.14}$$

which permits obtaining the *j*-th element of $\mathbf{w}^{*\prime}$.

14.2. Annex to Part II

14.2.1. Relationship between the Scheffé and Cox models coefficients

Consider formulation of the second-order Scheffé polynomial:

$$E(y) \quad = \sum_{i=1}^{Q} \beta_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \beta_{ij} \cdot x_i \cdot x_j \tag{14.15}$$

And that of the second-order Cox polynomial:

$$E(y) = \beta'_0 + \sum_{i=1}^{Q} \beta'_i \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \beta'_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \beta'_{ii} \cdot x_i^2$$

$$s.t. \begin{cases} \sum_{i=1}^{Q} \beta'_i \cdot s_i = 0 \\ \sum_{j=1}^{Q} c_{ij} \cdot \beta'_{ij} \cdot s_j = 0 \ \ \forall\, i \in \{1,2,\dots,Q\} \\ c_{ij} = \begin{cases} 1/2 & if\ i \neq j \\ 1 & if\ i = j \end{cases} \end{cases} \tag{14.16}$$

Consider also the perfect collinearity constraint in Equation 5.1:

$$\sum_{i=1}^{Q} x_i = 1 \tag{14.17}$$

The second order terms x_i^2 can then be reformulated as:

$$x_i^2 = x_i \cdot \left(1 - \sum_{\substack{j=1 \\ j\neq i}}^{Q} x_j \right) \tag{14.18}$$

Then, from Equation 14.16:

$$E(y) = \beta'_0 \cdot \sum_{i=1}^{Q} x_i + \sum_{i=1}^{Q} \beta'_i \cdot x_i + \sum_{i=1}^{Q} \sum_{\substack{j=1 \\ j \neq i}}^{Q} \beta'^*_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \beta'_{ii} \cdot x_i \cdot \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^{Q} x_j \right)$$

$$E(y) = \sum_{i=1}^{Q} (\beta'_0 + \beta'_i) \cdot x_i + \sum_{i=1}^{Q} \sum_{\substack{j=1 \\ j \neq i}}^{Q} \beta'^*_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^{Q} \beta'_{ii} \cdot x_i - \sum_{i=1}^{Q} \sum_{\substack{j=1 \\ j \neq i}}^{Q} \beta'_{ii} \cdot x_i \cdot x_j \quad (14.19)$$

$$E(y) = \sum_{i=1}^{Q} (\beta'_0 + \beta'_i + \beta'_{ii}) \cdot x_i + \sum_{i=1}^{Q} \sum_{\substack{j=1 \\ j \neq i}}^{Q} \left(\beta'^*_{ij} - \beta'_{ii} \right) \cdot x_i \cdot x_j$$

where the notation β'^*_{ij} permits to explicitly differentiate the interaction terms $x_i \cdot x_j$ and $x_j \cdot x_i$. Note that, since $\beta'^*_{ij} = \beta'^*_{ji}$, $\beta'_{ij} = 2 \cdot \beta'^*_{ij} = 2 \cdot \beta'^*_{ji}$ whenever $i \neq j$. Therefore, Equation 14.19 can also be formulated as:

$$E(y) = \sum_{i=1}^{Q} (\beta'_0 + \beta'_i + \beta'_{ii}) \cdot x_i + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \left(\beta'_{ij} - \beta'_{ii} - \beta'_{jj} \right) \cdot x_i \cdot x_j \quad (14.20)$$

Since Equation 14.15 and 14.20 must be equivalent, then:

$$\begin{aligned} \beta_i &= \beta'_0 + \beta'_i + \beta'_{ii} \\ \beta_{ij} &= \beta'_{ij} - \beta'_{ii} - \beta'_{jj} \end{aligned} \quad (14.21)$$

If, instead of the second-order, the first-order polynomial is considered, then $\beta'_{ij} = \beta'_{ii} = \beta'_{jj} = 0$, and therefore:

$$\beta_i = \beta'_0 + \beta'_i \quad (14.22)$$

For higher-degree polynomials the reader is referred to [57], where the corresponding relationship between the Scheffé and Cox models' coefficients it is shown.

14.2.2. Projection of a point/vector onto the intersection of a group of hyperplanes

Consider the existence of *R* hyperplanes of dimension *A*-1 in a space of dimension *A*, with *R*<*A*, and let the equation of the *r*-th hyperplane, NS_r, to be expressed as:

$$v_0 + \sum_{a=1}^{A} v_a \cdot \tau_{\mathrm{NS}_r,a} = 0 \; ; \mathbf{v} = [v_1, v_2, \ldots, v_A]^{\mathrm{T}} \quad (14.23)$$

Consider the intersection of these hyperplanes to exist and to be of dimension *A*-*R*. Let $\boldsymbol{\tau}$ be a point outside of this intersection, and $\tilde{\boldsymbol{\tau}}$ the point closest to $\boldsymbol{\tau}$ that belongs to the intersection of all *R* hyperplanes. The directed distance between $\boldsymbol{\tau}$ and the *r*-th hyperplane NS, can be calculated as:

$$dd(\boldsymbol{\tau}, \mathrm{NS}_r) = \frac{v_{r,0}}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} + \mathbf{u}_{\mathrm{NS}_r}^{\mathrm{T}} \cdot \boldsymbol{\tau} \quad ; \quad \mathbf{u}_{\mathrm{NS}_r} = \frac{\mathbf{v}_r}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} \tag{14.24}$$

where $\mathbf{u}_{\mathrm{NS}_r}$ is the unitary vector orthogonal to NS_r and $dd(\boldsymbol{\tau}, \mathrm{NS}_r)$ has the same sign and magnitude as the Euclidean distance when the direction of the vector from the *r*-th NS to $\boldsymbol{\tau}$ is the same as that of $\mathbf{u}_{\mathrm{NS}_r}$, but is negative if the have opposing directions. Since $\tilde{\boldsymbol{\tau}}$ is a point in the intersection of the *R* hyperplanes, then $dd(\tilde{\boldsymbol{\tau}}, \mathrm{NS}_r) = 0$.

On the other hand, the vector that connects $\tilde{\boldsymbol{\tau}}$ to $\boldsymbol{\tau}$ must be orthogonal to any vector that is also orthogonal to every $\mathbf{v}_r$ in order to guarantee that $\tilde{\boldsymbol{\tau}}$ is the point closest to $\boldsymbol{\tau}$ that is also in the intersection of the *R* hyperplanes. As a consequence, $\tilde{\boldsymbol{\tau}}$ can be expressed as a function of $\boldsymbol{\tau}$ of the form

$$\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau} + \sum_{r=1}^{R} c_r \cdot \mathbf{v}_r = \boldsymbol{\tau} + \mathbf{V}^{\mathrm{T}} \cdot \mathbf{c} \quad ; \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_R^{\mathrm{T}} \end{bmatrix} \tag{14.25}$$

where $\mathbf{c}$ is the vector that relates $\tilde{\boldsymbol{\tau}}$ and $\boldsymbol{\tau}$. Taking into account Equations 14.24 and 14.25:

$$dd(\tilde{\boldsymbol{\tau}}, \mathrm{NS}_r) = \frac{v_{r,0}}{\sqrt{\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{v}_r}} + \mathbf{u}_{\mathrm{NS}_r}^{\mathrm{T}} \cdot (\boldsymbol{\tau} + \mathbf{V}^{\mathrm{T}} \cdot \mathbf{c}) = 0 \tag{14.26}$$

This equality can be also expressed as

$$\mathbf{v}_r^{\mathrm{T}} \cdot \mathbf{V}^{\mathrm{T}} \cdot \mathbf{c} = -\left(v_{r,0} + \mathbf{v}_r^{\mathrm{T}} \cdot \boldsymbol{\tau}\right) \tag{14.27}$$

Therefore, the following system of *R* linear equations with *R* unknowns can be defined:

$$\begin{gathered} \mathbf{V} \cdot \mathbf{V}^{\mathrm{T}} \cdot \mathbf{c} = -(\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}) \\ \mathbf{v}_0 = \begin{bmatrix} v_{1,0} \\ v_{2,0} \\ \vdots \\ v_{R,0} \end{bmatrix} \quad ; \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_R^{\mathrm{T}} \end{bmatrix} \end{gathered} \tag{14.28}$$

Since the rank of $\mathbf{V}$ is *R*, $\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}}$ is a [*R*×*R*] matrix with rank *R* and can therefore be inverted. Then:

$$\mathbf{c} = -(\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot (\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}) \tag{14.29}$$

Substituting Equation 14.29 in Equation 14.25:

$$\begin{gathered}\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau} - \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot (\mathbf{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}) = \\ = [\mathbf{I}_A - \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{V}] \cdot \boldsymbol{\tau} - \mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{v}_0\end{gathered} \tag{14.30}$$

Where $\mathbf{I}_A$ is the [A×A] identity matrix.

Therefore $\tilde{\boldsymbol{\tau}}$ is the projection of $\boldsymbol{\tau}$ onto the intersection of R ($R < A$) hyperplanes in an A-dimensional space. The extension of Equation 14.30 for a vector, instead of a point, is straightforward.

14.3. Annex to Part III

14.3.1. Relationship between the result of the PLS-regression direct inversion and the point in the combined null space closest to the centre of projection/with lowest leverage

Consider that, for the hyperplanes corresponding to the NS for $\mathbf{y}_{\mathrm{DES}}$:

$$\begin{gathered}\mathbf{v}_0 = \mathbf{m}_{\mathbf{Y}} - \mathbf{y}_{\mathrm{DES}} \\ \mathbf{V} = \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}} \cdot \mathbf{Q}\end{gathered} \tag{14.31}$$

$\mathbf{m}_{\mathbf{Y}}$ and $\mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}$ being the [L×1] column vector of centring factors and the [L×L] diagonal matrix with the scaling factors applied to the L output variables before fitting the PLS-regression model, respectively.

If $\boldsymbol{\tau} = \mathbf{0}$ in Equation 14.30, then:

$$\tilde{\boldsymbol{\tau}} = -\mathbf{V}^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathrm{T}})^{-1} \cdot \mathbf{v}_0 = \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S}_{\mathbf{Y}}} \cdot \left(\mathbf{D}_{\mathbf{s}_{\mathbf{Y}}} \cdot \mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}\right)^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{14.32}$$

where

$$\left(\mathbf{D}_{\mathbf{s}_{\mathbf{Y}}} \cdot \mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}\right)^{-1} = \left(\mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}\right)^{-1} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}{}^{-1} = \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}{}^{-1} \cdot (\mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}})^{-1} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}{}^{-1} \tag{14.33}$$

And therefore:

$$\tilde{\boldsymbol{\tau}} = \mathbf{Q}^{\mathrm{T}} \cdot (\mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}})^{-1} \cdot \mathbf{D}_{\mathbf{s}_{\mathbf{Y}}}{}^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{14.34}$$

Which is, exactly, the expression of the scores $\boldsymbol{\tau}_{\mathrm{NEW}}$ for the direct inversion, as seen in Equation 9.16. It must be noted that the point in the intersection of NS for a given $\mathbf{y}_{\mathrm{DES}}$ that is closest to the centre of projection is not necessarily the same as the point in that intersection with the smallest leverage. To demonstrate this, consider that, for the hyperplanes corresponding to the NS for $\mathbf{y}_{\mathrm{DES}}$, with standardized scores:

$$\begin{gathered}\mathbf{v}_0 = \mathbf{m}_{\mathbf{Y}} - \mathbf{y}_{\mathrm{DES}} \\ \mathbf{V} = \mathbf{D}_{\mathbf{S}_{\mathbf{Y}}} \cdot \mathbf{Q} \cdot \boldsymbol{\Lambda}^{1/2}\end{gathered} \tag{14.35}$$

with $\mathbf{\Lambda}^{1/2}$ defined as the [A×A] diagonal matrix containing the A standard deviations of the scores associated to the LVs. If $\boldsymbol{\tau} = \mathbf{0}$ in Equation 14.30, then:

$$\tilde{\boldsymbol{\tau}} = \mathbf{\Lambda}^{1/2} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}} \cdot \left(\mathbf{D}_{\mathbf{S_Y}} \cdot \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{S_Y}}\right)^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{14.36}$$

Simplifying, as in Equation 14.34:

$$\tilde{\boldsymbol{\tau}} = \mathbf{\Lambda}^{1/2} \cdot \mathbf{Q}^{\mathrm{T}} \cdot (\mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^{\mathrm{T}})^{-1} \cdot \mathbf{D}_{\mathbf{S_Y}}{}^{-1} \cdot (\mathbf{y}_{\mathrm{DES}} - \mathbf{m}_{\mathbf{Y}}) \tag{14.37}$$

Therefore, Equations 14.34 and 14.37 are only equivalent if $\mathbf{\Lambda}$ is the [A×A] identity matrix.

14.3.2. Analytical expression for the confidence region of the null space for a linear combination of outputs using OLS-type expression for the prediction's confidence interval

Let $\mathbf{x}_{\mathrm{obs}}$ be the vector of inputs for an observation, and $\hat{\mathbf{y}}_{\mathrm{obs}}$ the prediction of the outputs, given $\mathbf{x}_{\mathrm{obs}}$, by a PLS regression model fitted with A latent variables relating the [N×M] matrix of inputs $\mathbf{X}$ and the [N×L] matrix of outputs $\mathbf{Y}$. Let d_{obs} be the value for the quality attribute of interest, which can be expressed as a linear combination of the outputs $\hat{\mathbf{y}}_{\mathrm{obs}}$ such that

$$d_{\mathrm{obs}} = \mathbf{a}^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{obs}} = \mathbf{a}^{\mathrm{T}} \cdot \left(\mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \boldsymbol{\tau}_{\mathrm{obs}} + \mathbf{m}_{\mathbf{Y}}\right) \tag{14.38}$$

where $\mathbf{a}$ is the the [L×1] vector with the coefficients that relate d_{obs} with $\hat{\mathbf{y}}_{\mathrm{obs}}$; $\widehat{\mathbf{Q}}$[xv] is the estimated [L×A] matrix of loadings that relate $\boldsymbol{\tau}_{\mathrm{obs}}$ with $\mathbf{y}_{\mathrm{obs}}$; and $\boldsymbol{\tau}_{\mathrm{obs}}$ is the [A×1] vector of scores corresponding to the projection of $\mathbf{x}_{\mathrm{obs}}$ onto the latent space; $\mathbf{m}_{\mathbf{Y}}$ and $\mathbf{D}_{\mathbf{s_Y}}$ being the [L×1] column vector of centring factors and the [L×L] diagonal matrix with the scaling factors applied to the L output variables when fitting the PLS model, respectively.

Let NS represent the Null Space (NS), that is, a subspace such that any combination of inputs inside it will theoretically guarantee the desired value for the quality attribute of interest, d_{DES}. Any vector of outputs $\hat{\mathbf{y}}_{\mathrm{NS}}$ with projection $\boldsymbol{\tau}_{\mathrm{NS}}$ in this subspace will satisfy the following:

$$\mathbf{a}^{\mathrm{T}} \cdot \hat{\mathbf{y}}_{\mathrm{NS}} = \mathbf{a}^{\mathrm{T}} \cdot \left(\mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \boldsymbol{\tau}_{\mathrm{NS}} + \mathbf{m}_{\mathbf{Y}}\right) = d_{\mathrm{DES}} \tag{14.39}$$

Equation 14.39 can be reorganized as

$$\mathbf{a}^{\mathrm{T}} \cdot \mathbf{m}_{\mathbf{Y}} - d_{\mathrm{DES}} + \mathbf{a}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \boldsymbol{\tau}_{\mathrm{NS}} = 0 \tag{14.40}$$

which is the equation of a hyper-plane in the latent space associated to NS of the form:

[xv] $\widehat{\mathbf{Q}}$ is used here to differentiate it from the theoretical loadings matrix $\mathbf{Q}$, as in the appendix in Section 14.3.3, to keep it consistent with it since such distinction is required for the definition of the confidence region of the NS.

$$\begin{gathered} v_0 + \sum_{a=1}^{\mathrm{A}} v_a \cdot \tau_{\mathrm{NS},a} = 0 \,;\, \mathbf{v} = [v_1, v_2, \ldots, v_A]^{\mathrm{T}} \\ v_0 = \mathbf{a}^{\mathrm{T}} \cdot \mathbf{m}_{\mathbf{Y}} - d_{\mathrm{DES}} \\ \mathbf{v} = \widehat{\mathbf{Q}}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \mathbf{a} \end{gathered} \tag{14.41}$$

Consider $\tilde{\boldsymbol{\tau}}_{\mathrm{obs}}$ to be a vector of scores on the hyper-plane defining the NS. Then, the $100 \cdot (1-\alpha)\%$ confidence interval (CI) of the prediction of a linear combination $\tilde{d}_{\mathrm{obs}} = d_{\mathrm{DES}}$, given an observation $\tilde{\mathbf{x}}_{\mathrm{obs}}$ with projection $\tilde{\boldsymbol{\tau}}_{\mathrm{obs}}$**,** and using OLS type expressions, is:

$$CI_{d_{\mathrm{DES}}} = \tilde{d}_{\mathrm{obs}} \pm t_{N-A,\alpha/2} \cdot s_{\tilde{d}_{\mathrm{obs}}} \tag{14.42}$$

Now consider $\boldsymbol{\tau}_{\mathrm{obs}}$ to be a vector of scores within the limits of the confidence region for the NS, such that $\tilde{\boldsymbol{\tau}}_{\mathrm{obs}}$ is the vector of scores closest to $\boldsymbol{\tau}_{\mathrm{obs}}$ and exactly on the hyper-plane defining NS_r. Then:

$$\begin{gathered} \mathbf{a}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \tilde{\boldsymbol{\tau}}_{\mathrm{obs}} - t_{N-A,\alpha/2} \cdot s_{\tilde{d}_{\mathrm{obs}}} \leq \mathbf{a}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \boldsymbol{\tau}_{\mathrm{obs}} \\ \mathbf{a}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \boldsymbol{\tau}_{\mathrm{obs}} \leq \mathbf{a}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot \tilde{\boldsymbol{\tau}}_{\mathrm{obs}} + t_{N-A,\alpha/2} \cdot s_{\tilde{d}_{\mathrm{obs}}} \end{gathered} \tag{14.43}$$

Or, equivalently:

$$\left|\mathbf{a}^{\mathrm{T}} \cdot \mathbf{D}_{\mathbf{s_Y}} \cdot \widehat{\mathbf{Q}} \cdot (\boldsymbol{\tau}_{\mathrm{obs}} - \tilde{\boldsymbol{\tau}}_{\mathrm{obs}})\right| \leq t_{N-A,\alpha/2} \cdot s_{\tilde{d}_{\mathrm{obs}}} \tag{14.44}$$

where $t_{N-A,\alpha/2}$ is the $100 \cdot (1-\alpha/2)$ percentile of a Student's t-distribution with (*N-A*) degrees of freedom, and $s_{\tilde{d}_{\mathrm{obs}}}$ the estimated standard deviation for the prediction error of the linear combination of outputs:

$$s_{\tilde{d}_{\mathrm{obs}}} = SE_d \cdot \sqrt{1 + h_{\tilde{\tau}_{\mathrm{obs}}} + \frac{1}{N}} \tag{14.45}$$

SE_d being the standard error of calibration and $h_{\tilde{\boldsymbol{\tau}}_{\mathrm{obs}}}$ the leverage of $\tilde{\boldsymbol{\tau}}_{\mathrm{obs}}$

As illustrated in Section 14.3.3, SE_d can be calculated as:

$$SE_d = \sqrt{\mathbf{a}^{\mathrm{T}} \cdot \mathbf{S}_{\mathbf{f}} \cdot \mathbf{a}} \tag{14.46}$$

where $\mathbf{S}_{\mathrm{f}}$ is the estimated $[L \times L]$ variance-covariance matrix of prediction errors:

$$\mathbf{S_f} = \begin{pmatrix} s_{f_1}^2 & \widehat{\text{cov}}(f_1, f_2) & \cdots & \widehat{\text{cov}}(f_1, f_L) \\ \widehat{\text{cov}}(f_2, f_1) & s_{f_2}^2 & & \widehat{\text{cov}}(f_2, f_L) \\ \vdots & & \ddots & \vdots \\ \widehat{\text{cov}}(f_L, f_1) & \widehat{\text{cov}}(f_L, f_2) & \cdots & s_{f_L}^2 \end{pmatrix}$$
$$s_{f_l}^2 = \frac{\sum_{n=1}^{N} \left(y_{n,l} - \hat{y}_{n,l}\right)^2}{rdf} \tag{14.47}$$
$$\widehat{\text{cov}}(f_l, f_{l'}) = \frac{\sum_{n=1}^{N} \left(y_{n,l} - \hat{y}_{n,l}\right) \cdot \left(y_{n,l'} - \hat{y}_{n,l'}\right)}{rdf}$$

$y_{n,l}$ and $\hat{y}_{n,l}$ being the observed and average predicted values of the *l*-th output variable for the *n*-th sample in the calibration dataset used to fit the regression model, and *rdf* the residual degrees of freedom (frequently *rdf*=*N*-*A*).

On the other hand, $h_{\tilde{\boldsymbol{\tau}}_{\text{obs}}}$ is computed as:

$$h_{\tilde{\boldsymbol{\tau}}_{\text{obs}}} = \tilde{\boldsymbol{\tau}}_{\text{obs}}^{\text{T}} \cdot (\mathbf{T}^{\text{T}} \cdot \mathbf{T})^{-1} \cdot \tilde{\boldsymbol{\tau}}_{\text{obs}} \tag{14.48}$$

$\mathbf{T}$ being the [N×A] matrix of scores corresponding to the projection onto the latent space of the observations in the calibration dataset used to fit the PLS-regression model. The relationship between $\boldsymbol{\tau}_{\text{obs}}$ and $\tilde{\boldsymbol{\tau}}_{\text{obs}}$, considering $\tilde{\boldsymbol{\tau}}_{\text{obs}}$ to be the vector of scores closest to $\boldsymbol{\tau}_{\text{obs}}$ and exactly on the hyper-plane defining the NS, can be expressed as:

$$\boldsymbol{\tau}_{\text{obs}} - \tilde{\boldsymbol{\tau}}_{\text{obs}} = dd(\boldsymbol{\tau}_{\text{obs}}, \tilde{\boldsymbol{\tau}}_{\text{obs}}) \cdot \mathbf{u}_{\text{NS}}$$
$$dd(\boldsymbol{\tau}_{\text{obs}}, \tilde{\boldsymbol{\tau}}_{\text{obs}}) = \frac{v_0}{\sqrt{\mathbf{v}^{\text{T}} \cdot \mathbf{v}}} + \mathbf{u}_{\text{NS}}^{\text{T}} \cdot \boldsymbol{\tau}_{\text{obs}} \quad ; \quad \mathbf{u}_{\text{NS}} = \frac{\mathbf{v}}{\sqrt{\mathbf{v}^{\text{T}} \cdot \mathbf{v}}} \tag{14.49}$$

where $\mathbf{u}_{\text{NS}}$ is the unitary vector orthogonal to the NS and $dd(\boldsymbol{\tau}_{\text{obs}}, \tilde{\boldsymbol{\tau}}_{\text{obs}})$ is the directed distance between $\boldsymbol{\tau}_{\text{obs}}$ and $\tilde{\boldsymbol{\tau}}_{\text{obs}}$, which has the same sign and magnitude as the Euclidean distance when the direction of the vector $\boldsymbol{\tau}_{\text{obs}} - \tilde{\boldsymbol{\tau}}_{\text{obs}}$ is the same as that of $\mathbf{u}_{\text{NS}}$, but is negative if the have opposing directions. From this, $\tilde{\boldsymbol{\tau}}_{\text{obs}}$ can be re-formulated as a function of $\boldsymbol{\tau}_{\text{obs}}$ such that

$$\tilde{\boldsymbol{\tau}}_{\text{obs}} = \boldsymbol{\tau}_{\text{obs}} - \frac{v_0 + \mathbf{v}^{\text{T}} \cdot \boldsymbol{\tau}_{\text{obs}}}{\mathbf{v}^{\text{T}} \cdot \mathbf{v}} \cdot \mathbf{v} \tag{14.50}$$

Re-organizing Equation 14.44 and substituting Equation 14.50 in it:

$$-t_{N-A,\alpha/2} \leq \frac{v_0 + \mathbf{v}^{\text{T}} \cdot \boldsymbol{\tau}_{\text{obs}}}{s_{\tilde{d}_{\text{obs}}}} \leq t_{N-A,\alpha/2}$$
$$s_{\tilde{d}_{\text{obs}}} = SE_d \cdot \sqrt{1 + h_{\boldsymbol{\tau}_{\text{obs}}} - \left(\frac{v_0 + \mathbf{v}^{\text{T}} \cdot \boldsymbol{\tau}_{\text{obs}}}{\mathbf{v}^{\text{T}} \cdot \mathbf{v}}\right)^2 \cdot \mathbf{v}^{\text{T}} \cdot (\mathbf{T}^{\text{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v} + \frac{1}{N}} \tag{14.51}$$

14.3.3. Confidence interval for the prediction of a linear combination of outputs in PLS using OLS type expression

Let $\mathbf{x}_{\text{obs}}$ be the vector of inputs for an observation, and $\hat{\mathbf{y}}_{\text{obs}}$ the vector of average predictions of the outputs $\mathbf{y}_{\text{obs}}$, given $\mathbf{x}_{\text{obs}}$, by a PLS regression model fitted with A latent variables relating the [N×M] matrix of inputs $\mathbf{X}$ and the [N×L] matrix of outputs $\mathbf{Y}$. Let $\hat{d}_{\text{obs}}$ be a linear combination of $\hat{\mathbf{y}}_{\text{obs}}$ such that:

$$\hat{d}_{\text{obs}} = \sum_{l=1}^{\text{L}} a_l \cdot \hat{y}_{obs,l} = \mathbf{a}^{\text{T}} \cdot \hat{\mathbf{y}}_{\text{obs}} = \mathbf{a}^{\text{T}} \cdot \widehat{\mathbf{Q}} \cdot \boldsymbol{\tau}_{\text{obs}} \tag{14.52}$$

where $\mathbf{a}$ is the the [L×1] vector with the coefficients that relate $\hat{d}_{\text{obs}}$ with $\hat{\mathbf{y}}_{\text{obs}}$, $\widehat{\mathbf{Q}}$ is the estimated [L×A] matrix of loadings that relate $\boldsymbol{\tau}_{\text{obs}}$ with $\mathbf{y}_{obs}$, and $\boldsymbol{\tau}_{\text{obs}}$ is the [A×1] vector of scores corresponding to the projection of $\mathbf{x}_{\text{obs}}$ onto the latent space.

And the variance of $\hat{d}_{\text{obs}}$ is:

$$\sigma^2_{\hat{d}_{\text{obs}}} = \mathbf{a}^{\text{T}} \cdot \mathbf{V}_{\hat{\mathbf{y}}_{\text{obs}}} \cdot \mathbf{a} \tag{14.53}$$

where $\mathbf{V}_{\hat{\mathbf{y}}_{\text{obs}}}$ is the [L×L] variance-covariance matrix corresponding to the average predictions of the output variables. In order to obtain the estimate of this matrix, $\mathbf{S}_{\hat{\mathbf{y}}_{\text{obs}}}$, Ordinary Least Squares (OLS) type expressions will be used, as done by Faber and Kowalski [95], also assuming normalized scores. To obtain the variances, consider

$$\begin{aligned} \mathbf{y}_l &= \mathbf{T} \cdot \mathbf{q}_l + \mathbf{u}_l \\ \mathbf{y}_l &= \mathbf{T} \cdot \hat{\mathbf{q}}_l + \mathbf{f}_l \\ \hat{\mathbf{y}}_l &= \mathbf{T} \cdot \hat{\mathbf{q}}_l \end{aligned} \tag{14.54}$$

$\mathbf{T}$ being the [N×A] scores matrix with $\boldsymbol{\tau}_1^{\text{T}}$, $\boldsymbol{\tau}_2^{\text{T}}$,…, $\boldsymbol{\tau}_{\text{N}}^{\text{T}}$ as its rows, $\mathbf{q}_l$ and $\hat{\mathbf{q}}_l$ are the [A×1] vectors corresponding to the l-th rows of the theoretical loadings matrix $\mathbf{Q}$ and its estimation, $\widehat{\mathbf{Q}}$, respectively, and $\mathbf{u}_l$and $\mathbf{f}_l$ are the [N×1] vectors of disturbances and prediction errors for the N observations[xvi].

The variance of average prediction for the conditions (scores) corresponding to the n-th observation and l-th output $\hat{y}_{n,l}$ is, from the previous expression:

[xvi] Note that, in this section only, and for simplification purposes, it has been assumed that both input and output variables are already pre-treated (e.g. mean-centred and scaled to unit variance)

$$\begin{aligned}\sigma^2_{\hat{y}_{n,l}} &= \mathrm{E}\left[\left(\hat{y}_{n,l} - \mathrm{E}(\hat{y}_{n,l})\right)\cdot\left(\hat{y}_{n,l} - \mathrm{E}(\hat{y}_{n,l})\right)^{\mathrm{T}}\right] \\ \sigma^2_{\hat{y}_{n,l}} &= \mathrm{E}\left[\left(\boldsymbol{\tau}_n^{\mathrm{T}}\cdot\hat{\mathbf{q}}_l - \boldsymbol{\tau}_n^{\mathrm{T}}\cdot\mathrm{E}(\hat{\mathbf{q}}_l)\right)\cdot\left(\boldsymbol{\tau}_n^{\mathrm{T}}\cdot\hat{\mathbf{q}}_l - \boldsymbol{\tau}_n^{\mathrm{T}}\cdot\mathrm{E}(\hat{\mathbf{q}}_l)\right)^{\mathrm{T}}\right] \\ \sigma^2_{\hat{y}_{n,l}} &= \boldsymbol{\tau}_n^{\mathrm{T}}\cdot\mathrm{E}\left[\left(\hat{\mathbf{q}}_l - \mathrm{E}(\hat{\mathbf{q}}_l)\right)\cdot\left(\hat{\mathbf{q}}_l - \mathrm{E}(\hat{\mathbf{q}}_l)\right)^{\mathrm{T}}\right]\cdot\boldsymbol{\tau}_n = \boldsymbol{\tau}_n^{\mathrm{T}}\cdot\mathbf{V}_{\hat{\mathbf{q}}_l}\cdot\boldsymbol{\tau}_n\end{aligned} \tag{14.55}$$

where $\mathrm{E}[\cdot]$ is the expectation operator and $\mathbf{V}_{\hat{\mathbf{q}}_l}$ is the $[A\times A]$ variance-covariance matrix corresponding to loadings of the matrix $\hat{\mathbf{Q}}$ associated with the l-th output variable.

Then, if the formulation for the OLS estimate is applied:

$$\mathbf{V}_{\hat{\mathbf{q}}_l} = \mathrm{E}\left[\left(\hat{\mathbf{q}}_l - \mathrm{E}(\hat{\mathbf{q}}_l)\right)\cdot\left(\hat{\mathbf{q}}_l - \mathrm{E}(\hat{\mathbf{q}}_l)\right)^{\mathrm{T}}\right] = \mathrm{E}[(\hat{\mathbf{q}}_l - \mathbf{q}_l)\cdot(\hat{\mathbf{q}}_l - \mathbf{q}_l)^{\mathrm{T}}] \tag{14.56}$$

Since

$$\begin{aligned}\hat{\mathbf{q}}_l = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\mathbf{y}_l = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot(\mathbf{T}\cdot\mathbf{q}_l + \mathbf{u}_l) = \\ = \mathbf{q}_l + (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\mathbf{u}_l\end{aligned} \tag{14.57}$$

Then

$$\begin{aligned}\mathbf{V}_{\hat{\mathbf{q}}_l} &= \mathrm{E}\left[((\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\mathbf{u}_l)\cdot(\mathbf{u}_l^{\mathrm{T}}\cdot\mathbf{T}\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1})\right] \\ \mathbf{V}_{\hat{\mathbf{q}}_l} &= (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\mathrm{E}(\mathbf{u}_l\cdot\mathbf{u}_l^{\mathrm{T}})\cdot\mathbf{T}\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\end{aligned} \tag{14.58}$$

$\mathrm{E}(\mathbf{u}_l\cdot\mathbf{u}_l^{\mathrm{T}})$ can be expressed as:

$$\begin{aligned}\mathrm{E}(\mathbf{u}_l\cdot\mathbf{u}_l^{\mathrm{T}}) &= \mathrm{E}\begin{pmatrix} u_{1,\mathrm{l}}^2 & u_{1,l}\cdot u_{2,l} & \cdots & u_{1,l}\cdot u_{N,l} \\ u_{2,l}\cdot u_{1,l} & u_{2,l}^2 & & u_{2,l}\cdot u_{N,l} \\ \vdots & & \ddots & \vdots \\ u_{N,l}\cdot u_{1,l} & u_{N,l}\cdot u_{2,l} & \cdots & u_{N,l}^2 \end{pmatrix} \\ \mathrm{E}(\mathbf{u}_l\cdot\mathbf{u}_l^{\mathrm{T}}) &= \begin{pmatrix} \sigma^2_{u_{1,l}} & cov(u_{1,l},u_{2,l}) & \cdots & cov(u_{1,l},u_{N,l}) \\ cov(u_{2,l},u_{1,l}) & \sigma^2_{u_{2,l}} & & cov(u_{2,l},u_{N,l}) \\ \vdots & & \ddots & \vdots \\ cov(u_{N,l},u_{1,l}) & cov(u_{N,l},u_{2,l}) & \cdots & \sigma^2_{u_{N,l}} \end{pmatrix}\end{aligned} \tag{14.59}$$

Assuming uncertainty to be the same for all observations, $\sigma^2_{u_{n,l}} = \sigma^2_{u_l}\ \forall\ n$, and $u_{n,l}$ and $u_{n',l}$ to be uncorrelated ($n \neq n'$) so that $cov(u_{n,l},u_{n',l}) = 0$, then:

$$\mathrm{E}(\mathbf{u}_l\cdot\mathbf{u}_l^{\mathrm{T}}) = \sigma^2_{u_l}\cdot\mathbf{I}_N \tag{14.60}$$

where $\mathbf{I}_N$ is the $[N\times N]$ Identity matrix. Therefore:

$$\mathbf{V}_{\hat{\mathbf{q}}_l} = (\mathbf{T}^{\mathrm{T}}\cdot\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\sigma^2_{u_l}\cdot\mathbf{I}_{N\times N}\cdot\mathbf{T}\cdot(\mathbf{T}^{\mathrm{T}}\cdot\mathbf{T})^{-1} \tag{14.61}$$

Reorganizing terms and simplifying:

$$\mathbf{V}_{\hat{\mathbf{q}}_l} = \sigma_{u_l}^2 \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \tag{14.62}$$

Substituting Equation 14.62 in Equation 14.55:

$$\sigma_{\hat{y}_{n,l}}^2 = \sigma_{u_l}^2 \cdot \boldsymbol{\tau}_n^{\mathrm{T}} \cdot (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} \cdot \boldsymbol{\tau}_n = \sigma_{u_l}^2 \cdot h_n \tag{14.63}$$

Where $\sigma_{u_l}^2$ is the variance of the disturbances for the *l*-th output variable, while h_n is the leverage for the *n*-th observation, which is a measure of the distance of the projection of the *n*-th observation to the centre of projection.

The variance of the disturbances for the *l*-th output can be estimated from the prediction errors as:

$$s_{f_l}^2 = \hat{\sigma}_{u_l}^2 = \frac{\sum_{n=1}^{N}\left(y_{n,l} - \hat{y}_{n,l}\right)^2}{N - df} \tag{14.64}$$

Being *df* the degrees of freedom consumed by the model (usually *df=A*).

Thus, the estimation $s_{\hat{y}_{n,l}}^2$ of the variance of the average prediction, $\sigma_{\hat{y}_{n,l}}^2$:

$$s_{\hat{y}_{n,l}}^2 = \hat{\sigma}_{\hat{y}_{n,l}}^2 = s_{f_l}^2 \cdot h_n \tag{14.65}$$

If, as it is usually the case when fitting a PLS regression model, the output variables have been centred, the error associated with the estimation of the mean has to be considered, resulting in the following modification to the previous expression:

$$s_{\hat{y}_{n,l}^o}^2 = s_{f_l}^2 \cdot \left(h_n + \frac{1}{N}\right) \tag{14.66}$$

Where $\hat{y}_{n,l}^o$ denotes that output variables are not centred, as opposed to $s_{\hat{y}_l}^2$.

And for an individual observation the error of prediction (i.e. the estimation of the uncertainty) is included, which results in:

$$s_{\hat{y}_{\mathrm{i},n,l}^o}^2 = s_{f_l}^2 \cdot \left(1 + h_n + \frac{1}{N}\right) \tag{14.67}$$

Following the same procedure now for the covariance of average prediction for the conditions (scores) corresponding to the *n*-th covariate observation for the *l*-th output $\hat{y}_{n,l}$ and *l'*-th output $\hat{y}_{n,l'}$ $(l \neq l')$, it follows:

$$\begin{gathered} cov(\hat{y}_l\,, \hat{y}_{l'})_n = \mathrm{E}\left[\left(\hat{y}_{n,l} - \mathrm{E}(\hat{y}_{n,l})\right) \cdot \left(\hat{y}_{n,l'} - \mathrm{E}(\hat{y}_{n,l'})\right)^{\mathrm{T}}\right] \\ cov(\hat{y}_l\,, \hat{y}_{l'})_n = \boldsymbol{\tau}_n^{\mathrm{T}} \cdot \mathrm{E}[(\hat{\mathbf{q}}_l - \mathbf{q}_l) \cdot (\hat{\mathbf{q}}_{l'} - \mathbf{q}_{l'})^{\mathrm{T}}] \cdot \boldsymbol{\tau}_n \\ cov(\hat{y}_l\,, \hat{y}_{l'})_n = \boldsymbol{\tau}_n^{\mathrm{T}} \cdot \mathbf{V}_{\hat{\mathbf{q}}_l, \hat{\mathbf{q}}_{l'}} \cdot \boldsymbol{\tau}_n \end{gathered} \tag{14.68}$$

In a similar way as for the variance:

$$\begin{aligned}\mathbf{V}_{\hat{\mathbf{q}}_l,\hat{\mathbf{q}}_{l'}} &= \mathrm{E}\left[((\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\mathbf{u}_l)\cdot\left(\mathbf{u}_{l'}^{\mathrm{T}}\cdot\mathbf{T}\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\right)\right] \\ \mathbf{V}_{\hat{\mathbf{q}}_l,\hat{\mathbf{q}}_{l'}} &= (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot\mathrm{E}\left(\mathbf{u}_l\cdot\mathbf{u}_{l'}^{\mathrm{T}}\right)\cdot\mathbf{T}\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\end{aligned} \tag{14.69}$$

And, since $\mathrm{E}\left(\mathbf{u}_l\cdot\mathbf{u}_{l'}^{\mathrm{T}}\right)$ can be expressed as:

$$\begin{aligned}\mathrm{E}\left(\mathbf{u}_l\cdot\mathbf{u}_{l'}^{\mathrm{T}}\right) &= \mathrm{E}\begin{pmatrix} u_{1,l}\cdot u_{1,l'} & u_{1,l}\cdot u_{2,l'} & \cdots & u_{1,l}\cdot u_{N,l'} \\ u_{2,l}\cdot u_{1,l'} & u_{2,l}\cdot u_{2,l'} & \cdots & u_{2,l}\cdot u_{N,l'} \\ \vdots & & \ddots & \vdots \\ u_{N,l}\cdot u_{1,l'} & u_{N,l}\cdot u_{2,l'} & \cdots & u_{N,l}\cdot u_{N,l'} \end{pmatrix} \\ \mathrm{E}\left(\mathbf{u}_l\cdot\mathbf{u}_{l'}^{\mathrm{T}}\right) &= \begin{pmatrix} cov(u_{1,l},u_{1,l'}) & cov(\mathrm{u}_{1,l},u_{2,l'}) & \cdots & cov(u_{1,l},u_{N,l'}) \\ cov(u_{2,l},u_{1,l'}) & cov(u_{2,l},u_{2,l'}) & & cov(u_{2,l},u_{N,l'}) \\ \vdots & & \ddots & \vdots \\ cov(u_{N,l},u_{1,l'}) & cov(u_{N,l},u_{2,l'}) & \cdots & cov(\mathrm{u}_{N,l},u_{N,l'}) \end{pmatrix}\end{aligned} \tag{14.70}$$

Assuming uncertainty to be the same for all observations, $cov(u_{n,l},u_{n,l'}) = cov(u_l,u_{l'})\ \forall\ n$, and $u_{n,l}$ and $u_{n',l'}$ to be uncorrelated ($n \neq n'$) so that $cov(u_{n,l},u_{n',l'}) = 0$, then:

$$\mathrm{E}\left(\mathbf{u}_l\cdot\mathbf{u}_{l'}^{\mathrm{T}}\right) = cov(u_l,u_{l'})\cdot\mathbf{I}_N \tag{14.71}$$

Therefore:

$$\mathbf{V}_{\hat{\mathbf{q}}_l,\hat{\mathbf{q}}_{l'}} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\mathbf{T}^{\mathrm{T}}\cdot cov(u_l,u_{l'})\cdot\mathbf{I}_{N\times N}\cdot\mathbf{T}\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} \tag{14.72}$$

Reorganizing terms and simplifying:

$$\mathbf{V}_{\hat{\mathbf{q}}_l,\hat{\mathbf{q}}_{l'}} = cov(u_l,u_{l'})\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} \tag{14.73}$$

Substituting Equation 14.73 in Equation 14.68:

$$cov(\hat{y}_l,\hat{y}_{l'})_n = cov(u_l,u_{l'})\cdot\boldsymbol{\tau}_n^{\mathrm{T}}\cdot(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\cdot\boldsymbol{\tau}_n = cov\left(u_i,u_j\right)\cdot h_n \tag{14.74}$$

Again, the estimation of the covariance of the disturbances for the *l*-th and *l'*-th output variables can be obtained as:

$$\widehat{cov}(f_l,f_{l'}) = \frac{\sum_{n=1}^{N}\left(y_{n,l}-\hat{y}_{n,l}\right)\cdot\left(y_{n,l'}-\hat{y}_{n,l'}\right)}{N-df} \tag{14.75}$$

Finally, the estimated variance-covariance matrix of the average outputs prediction vector for the *n*-th observation $\hat{\mathbf{y}}_n$, $\mathbf{V}_{\hat{\mathbf{y}}_n}$, can be expressed as:

$$\mathbf{S}_{\hat{\mathbf{y}}_n} = \mathbf{S}_{\mathbf{f}}\cdot h_n = \begin{pmatrix} s_{f_1}^2 & \widehat{\mathrm{cov}}(f_1,f_2) & \cdots & \widehat{\mathrm{cov}}(f_1,f_L) \\ \widehat{\mathrm{cov}}(f_2,f_1) & s_{f_2}^2 & & \widehat{\mathrm{cov}}(f_2,f_L) \\ \vdots & & \ddots & \vdots \\ \widehat{\mathrm{cov}}(f_L,f_1) & \widehat{\mathrm{cov}}(f_L,f_2) & \cdots & s_{f_L}^2 \end{pmatrix}\cdot h_n \tag{14.76}$$

In order to account for the centring of the output variables:

$$\mathbf{S}_{\hat{\mathbf{y}}_n^{\circ}} = \mathbf{S}_{\mathbf{f}} \cdot \left(h_n + \frac{1}{N}\right) \tag{14.77}$$

And for the prediction of an individual observation:

$$\mathbf{S}_{\hat{\mathbf{y}}_{\mathrm{i},n}^{\circ}} = \mathbf{S}_{\mathbf{f}} \cdot \left(1 + h_n + \frac{1}{N}\right) \tag{14.78}$$

Finally, since centring of the output variables is performed in almost all cases, the estimated variance of $\hat{d}_n$ will be obtained for an individual observation as:

$$s_{\hat{d}_n}^2 = \mathbf{a}^{\mathrm{T}} \cdot \mathbf{S}_{\mathbf{f}} \cdot \mathbf{a} \cdot \left(1 + h_n + \frac{1}{N}\right) \tag{14.79}$$

Therefore, the confidence interval for the prediction of a linear combination of output variables, corresponding to a given observation $\boldsymbol{x}_{\mathrm{obs}}$ with projection $\boldsymbol{\tau}_{\mathrm{obs}}$ and predicted outputs $\widehat{\boldsymbol{y}}_{\mathrm{obs}}$, can be calculated as:

$$\begin{aligned} CI_{d_{\mathrm{obs}}} &= \hat{d}_{\mathrm{obs}} \pm t_{N-A,\alpha/2} \cdot s_{\hat{d}_{\mathrm{obs}}} \\ s_{\hat{d}_{\mathrm{obs}}} &= \sqrt{\mathbf{a}^{\mathrm{T}} \cdot \mathbf{S}_{\mathbf{f}} \cdot \mathbf{a} \cdot \left(1 + h_{\mathrm{obs}} + \frac{1}{N}\right)} \end{aligned} \tag{14.80}$$

As a particular case, the confidence interval for the prediction of the *l*-th of output variable for the same observation results:

$$\begin{aligned} CI_{y_{\mathrm{obs},l}} &= \hat{y}_{\mathrm{obs},l} \pm t_{N-A,\alpha/2} \cdot s_{\hat{y}_{\mathrm{obs},l}} \\ s_{\hat{y}_{\mathrm{obs},l}} &= \sqrt{s_{f_l}^2 \cdot \left(1 + h_{\mathrm{obs}} + \frac{1}{N}\right)} \end{aligned} \tag{14.81}$$

Note that, although here $\mathbf{S}_{\mathbf{f}}$ was computed assuming already pre-treated output variables (e.g. centred to zero mean and unitary variance), this matrix may also be obtained similarly without considering such pre-treatment in order to obtain the confidence intervals for the prediction in the same units as the original **Y** matrix (this must be accounted for also for $\hat{d}_{\mathrm{obs}}/\hat{y}_{\mathrm{obs},l}$).

Bibliography

[1] J.J. Liu, J.F. MacGregor, Modeling and Optimization of Product Appearance: Application to Injection-Molded Plastic Panels, Ind. Eng. Chem. Res. 44 (2005) 4687–4696.

[2] D. Bonvin, C. Georgakis, C.C. Pantelides, M. Barolo, M.A. Grover, D. Rodrigues, R. Schneider, D. Dochain, Linking Models and Experiments, Ind. Eng. Chem. Res. 55 (2016) 6891–6903.

[3] G.E.P. Box, J.S. Hunter, W.G. Hunter, Statistics for Experimenters: design, discovery, and innovation, Second, John Wiley & Sons, INC., 2005.

[4] A.C. Aitken, IV.—On Least Squares and Linear Combination of Observations, Proc. R. Soc. Edinburgh. 55 (1936) 42–48.

[5] C. Rao, Linear statistical inference and its applications, Second Edi, John Wiley & Sons, INC., New York, USA, 1973.

[6] D.C. Montgomery, Applied Statistics and Probability for Engineers Third Edition, 2003.

[7] V. Fedorov, Theory of Optimal Experiments Designs, New York: Academic Press, 1972.

[8] A. Atkinson, A. Donev, R. Tobias, Optimum Experimental Designs, with SAS, 2007.

[9] H.P. Wynn, Results in the Theory and Construction of D-Optimum Experimental Designs, J. R. Stat. Soc. Ser. B. 34 (1972) 133–147. doi:10.1079/IVPt200454()IN.

[10] T.J. Mitchell, An Algorithm for the Construction of " D -Optimal" Experimental Designs, Technometrics. 16 (1974) 203–210.

[11] A. Heredia-Langner, W.M. Carlyle, D.C. Montgomery, C.M. Borror, G.C. Runger, Genetic Algorithms for the Construction of D-Optimal Designs, J. Qual. Technol. 35 (2003) 28–46.

[12] R.K. Meyer, C.J. Nachtsheim, The Coordinate-Exchange for Algorithm Exact Constructing Optimal Experimental Designs, Technometrics. 37 (1995) 60–69.

[13] M. Rodriguez, B. Jones, C. Borror, D. C. Montgomery, Generating and Assessing Exact G-Optimal Designs, J. Qual. Technol. 42 (2010) 3–20.

[14] W.R. Wesley, J.R. Simpson, P.A. Parker, J.J.P. Jr., Exact Calculation of Integrated Prediction Variance for Response Surface Designs on Cuboidal and Spherical Regions, J. Qual. Technol. 41 (2009) 165–180.

[15] D. Lee, B. J. Schachter, Two Algorithms for Constructing a Delaunay Triangulation, Int. J. Parallel Program. 9 (1980) 219–242.

[16] Z. Jian-Ming, S. Ke-Ran, Z. Ke-Ding, Z. Qiong-Hua, Computing constrained triangulation and Delaunay triangulation: a new algorithm, IEEE Trans. Magn. 26 (1990) 694–697.

[17] L. De Floriani, E. Puppo, An on-line algorithm for constrained Delaunay triangulation, CVGIP Graph. Model. Image Process. 54 (1992) 290–300.

[18] X. Li, N. Sudarsanam, D. Frey, Regularities in data from factorial experiments: Research Articles, Complexity. 11 (2006) 32–45.

[19] L.M.C. Buydens, Towards Tsunami-Resistant Chemometrics, Anal. Sci. (2013) 401.

[20] M. Reis, G. Gins, Industrial Process Monitoring in the Big Data/Industry 4.0 Era: from Detection, to Diagnosis, to Prognosis, Processes. 5 (2017) 35.

[21] K. Pearson, LIII. *On lines and planes of closest fit to systems of points in space*, Philos. Mag. Ser. 6. 2 (1901) 559–572.

[22] H. Hotelling, Analysis of a complex of statistical variables into Principal Components. Jour. Educ. Psych., 24, 417-441, 498-520, J. Educ. Psychol. 24 (1933) 417–441.

[23] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika. 1 (1936) 211–218.

[24] S. Wold, A. Ruhe, H. Wold, W. Dunn, The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses, Siam J. Sci. Stat. Comput. 5 (1984).

[25] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228.

[26] S. Wold, M. Sjostrom, PLS-Regression: A basic tool of chemometrics.pdf, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.

[27] Z. Liu, M.-J. Bruwer, J. F. MacGregor, S. Rathore, D. E. Reed, M. J. Champagne, Modeling and Optimization of a Tablet Manufacturing Line, J. Pharm. Innov. 6 (2011) 170–180.

[28] J.F. MacGregor, M.J. Bruwer, I. Miletic, M. Cardin, Z. Liu, Latent variable models and big data in the process industries, IFAC-PapersOnLine. 28 (2015) 520–524.

[29] B. Walczak, D.L. Massart, The Radial Basis Functions — Partial Least Squares approach as a flexible non-linear regression technique, Anal. Chim. Acta. 331 (1996) 177–185.

[30] B. Walczak, D.L. Massart, Application of Radial Basis Functions - Partial Least Squares to non-linear pattern recognition problems: Diagnosis of process faults, Anal. Chim. Acta. 331 (1996) 187–193.

[31] F. Yacoub, J.F. MacGregor, Product optimization and control in the latent variable space of nonlinear PLS models, Chemom. Intell. Lab. Syst. 70 (2004) 63–74.

[32] S. Wold, N. Kettaneh-Wold, B. Skagerberg, Nonlinear PLS modeling, Chemom. Intell. Lab. Syst. 7 (1989) 53–65.

[33] S. Wold, Nonlinear partial least squares modelling II. Spline inner relation, Chemom. Intell. Lab. Syst. 14 (1992) 71–84.

[34] I.E. Frank, A nonlinear PLS model, Chemom. Intell. Lab. Syst. 8 (1990) 109–119.

[35] A. Hoskuldsson, Quadratic PLS regression, J. Chemom. 6 (1992) 307–334.

[36] A. Berglund, S. Wold, INLR (Implicit Non-linear Latent variable Regression). II. Blockscaling of Expanded Terms with QSAR Examples, Comput. Lead Find. Optim. Curr. Tools Med. Chem. 11 (2007) 65–79.

[37] J. Zupan, J. Gasteiger, Neural Networks in chemistry, (1993).

[38] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, 2002.

[39] D.S. Cao, Y.Z. Liang, Q.S. Xu, Q.N. Hu, L.X. Zhang, G.H. Fu, Exploring nonlinear relationships in chemical data using kernel-based methods, Chemom. Intell. Lab. Syst. 107 (2011) 106–115.

[40] P. Williams, Influence of water on prediction of composition and quality factors: The Aquaphotomics of low moisture agricultural materials, J. Near Infrared Spectrosc. 17 (2009) 315–328.

[41] C. Tan, M. Li, Mutual information-induced interval selection combined with kernel partial least squares for near-infrared spectral calibration, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 71 (2008) 1266–1273.

[42] M.J. Embrechts, S. Ekins, Classification of metabolites with kernel-partial least squares (K-PLS), Drug Metab. Dispos. 35 (2007) 325–327.

[43] J. Arenas-García, G. Camps-Valls, Efficient kernel orthonormalized PLS for remote sensing applications, IEEE Trans. Geosci. Remote Sens. 46 (2008) 2872–2881.

[44] N.P. Štruc, Gabor-Based Kernel Partial-Least-Squares Discrimination Features for Face Recognition, Informatica. 20 (2009) 115–138.

[45] R. Sun, F. Tsung, A kernel-distance-based multivariate control chart using support vector methods, Int. J. Prod. Res. 41 (2003) 2975–2989.

[46] J.M. Lee, C.K. Yoo, S.W. Choi, P.A. Vanrolleghem, I.B. Lee, Nonlinear process monitoring using kernel principal component analysis, Chem. Eng. Sci. 59 (2004) 223–234.

[47] K.P. Bennett, M.J. Embrechts, Advances in Learning Theory: Methods, Models and Applications, First Edit, IOS Press, Amsterdam, the Netherlands, 2003.

[48] R. H. Kewley, M. Embrechts, C. Breneman, Data strip mining for the virtual design of pharmaceuticals with neural networks, Neural Networks, IEEE Trans. 11 (2000) 668–679.

[49] B. Üstün, W.J. Melssen, L.M.C. Buydens, Visualisation and interpretation of Support Vector Regression models, Anal. Chim. Acta. 595 (2007) 299–309.

[50] C.F. Alcala, S.J. Qin, Reconstruction-based contribution for process monitoring with kernel principal component analysis, Am. Control Conf. (ACC), 2010. (2010) 7022–7027.

[51] J.C. Gower, S.A. Harding, Nonlinear biplots, Biometrika. 75 (1988) 445–455.

[52] P. W T Krooshof, B. Ustün, G. J Postma, L. Buydens, Visualization and Recovery of the (Bio)Chemical Interesting Variables in Data Analysis with Support Vector Machine Classification, Anal. Chem. 82 (2010) 7000–7007.

[53] G.J. Postma, P.W.T. Krooshof, L.M.C. Buydens, Opening the kernel of kernel partial least squares and support vector machines, Anal. Chim. Acta. 705 (2011) 123–134.

[54] A. Smolinska, L. Blanchet, L. Coulier, K.A.M. Ampt, T. Luider, R.Q. Hintzen, S.S. Wijmenga, L.M.C. Buydens, Interpretation and Visualization of Non-Linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis, PLoS One. 7 (2012) 1–12.

[55] J. Engel, G.J. Postma, I. van Peufflik, L. Blanchet, L.M.C. Buydens, Pseudo-sample trajectories for variable interaction detection in Dissimilarity Partial Least Squares, Chemom. Intell. Lab. Syst. 146 (2015) 89–101.

[56] R. Vitale, O.E. de Noord, A. Ferrer, A kernel-based approach for fault diagnosis in batch processes, J. Chemom. 28 (2014) 697–707.

[57] J.A. Cornell, Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data, Third, John Wiley & Sons, INC., 2011.

[58] R.N. Kackar, D.A. Harville, Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, Commun. Stat. - Theory Methods. 10 (1981) 1249–1261.

[59] D.A. Harville, D.R. Jeske, Mean Squared Error of Estimation or Prediction Under a General Linear Model, J. Am. Stat. Assoc. 87 (1992) 724–731.

[60] M.G. Kenward, J.H. Roger, Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood, Biometrics. 53 (1997) 983–997.

[61] L. Eriksson, E. Johansson, C. Wikstrom, Mixture design — design generation , PLS analysis , and model usage, (1998).

[62] P. Goos, B. Jones, Optimal Design of Experiments: A Case-Study Approach, 2011.

[63] K. Muteki, J.F. MacGregor, Multi-block PLS modeling for L-shape data structures with applications to mixture modeling, Chemom. Intell. Lab. Syst. 85 (2007) 186–194.

[64] S. García-Muñoz, M. Polizzi, WSPLS - A new approach towards mixture modeling and accelerated product development, Chemom. Intell. Lab. Syst. 114 (2012) 116–121.

[65] S. Garcia-Munoz, Two novel methods to analyze the combined effect of multiple raw-materials and processing conditions on the product's final attributes: JRPLS and TPLS, Chemom. Intell. Lab. Syst. 133 (2014) 49–62.

[66] G.F. Piepel, S.K. Cooley, B. Jones, Construction of a 21-Component Layered Mixture Experiment Design Using a New Mixture Coordinate-Exchange Algorithm, Qual. Eng. 17 (2005) 579–594.

[67] W. Smith, Experimental Design for Formulation, Society for Industrial and Applied Mathematics, 2005.

[68] R.D. Snee, M. D.W., Extreme vertices designs for linear mixture models, Technometrics. 16 (1974) 399–408.

[69] R.A. McLean, V.L. Anderson, Extreme Vertices Design of Mixture Experiments, Technometrics. 8 (1966) 447–454.

[70] H. Scheffe, Experiments with mixtures, J. R. Stat. Soc. B-20 (1958) 344–360.

[71] D.R. COX, A note on polynomial response functions for mixtures, Biometrika. 58 (1971) 155–159.

[72] S. Kowalski, J. A. Cornell, G. Geoffrey Vining, A new model and class of designs for mixture experiments with process variables, Commun. Stat. Methods - COMMUN Stat. METHOD. 29 (2000) 2255–2280.

[73] P. Prescott, Modelling in Mixture Experiments Including Interactions with Process Variables, Qual. Technol. Quant. Manag. 1 (2004) 87–103.

[74] T. Næs, E.M. Færgestad, J. Cornell, A comparison of methods for analyzing data from a three component mixture experiment in the presence of variation created by two process variables, Chemom. Intell. Lab. Syst. 41 (1998) 221–235.

[75] I. Måge, T. Næs, Split-plot design for mixture experiments with process variables: A comparison of design strategies, Chemom. Intell. Lab. Syst. 78 (2005) 81–95.

[76] P. Goos, A.N. Donev, Tailor-made split-plot designs for mixture and process variables, J. Qual. Technol. 39 (2007) 326–339.

[77] P. Gritzmann, V. Klee, On the complexity of some basic problems in computational convexity: I. Containment problems, Discrete Math. 136 (1994) 129–174.

[78] J.A. Cornell, Experiments with Mixtures, Second Edi, Wiley, New York, USA, 1990.

[79] R.D. Snee, Developing Blending Models for Gasoline and Other Mixtures, Technometrics. 23 (1981) 119–130.

[80] D. Alman, C. Pfeifer, Empirical colorant mixture models, Color Res. Appl. 12 (1987) 210–222.

[81] J.A. Cornell, How to run mixture experiments for product quality, Asq Press, 1990.

[82] L. Eriksson, T. Byrne, E. Johansson, J. Trygg, C. Vikström, Multi- and Megavariate Data Analysis - Basic Principles and Applications, Third Edit, MKS Umetrics AB, Malmö, Sweden, 2013.

[83] FDA, Pharmaceutical CGMPs for the 21s Century - A risk-based approach, 2004.

[84] ICH, Pharmaceutical Development Q8, ICH Harmon. Tripart. Guidel. 8 (2009) 1–28.

[85] J.F. MacGregor, M.-J. Bruwer, A Framework for the Development of Design and Control Spaces, J. Pharm. Innov. 3 (2008) 15–22.

[86] C. Jaeckle, J. Macgregor, Product design through multivariate statistical analysis of process data, Comput. Chem. Eng. 20 (1996) S1047–S1052.

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

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

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

[90] E. Tomba, M. Barolo, S. García-Muñoz, General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products, Ind. Eng. Chem. Res. 51 (2012) 12886–12900.

[91] L. Zhang, S. Garcia-Munoz, A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective, Chemom. Intell. Lab. Syst. 97 (2009) 152–158.

[92] C.C. Pantelides, N. Shah, C. Adjiman, Design Space, Models, and Model Uncertainty, Comprehensive Quality by Design in Pharmaceutical Development and Manufacture, Am. Inst. Chem. Eng. (2009).

[93] G. Bano, P. Facco, N. Meneghetti, F. Bezzo, M. Barolo, Uncertainty back-propagation in PLS model inversion for design space determination in pharmaceutical product development, Comput. Chem. Eng. 101 (2017) 110–124.

[94] A. Ferrer, Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process, Qual. Eng. 19 (2007) 311–325.

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

[96] A.J. Burnham, J.F. MacGregor, R. Viveros, Interpretation of regression coefficients under a latent variable regression model, J. Chemom. 15 (2001) 265–284.

[97] J. Ye, On measuring and correcting the effects of data mining and model selection, J. Am. Stat. Assoc. 93 (1998) 120–131.

[98] R. Romera, Prediction intervals in Partial Least Squares regression via a new local linearization approach, Chemom. Intell. Lab. Syst. 103 (2010) 122–128.

[99] J.J. Roffel, J.F. MacGregor, T.W. Hoffman, The Design and Implementation of a Multivariable Internal Model Controller for a Continuous Polybutadiene Polymerization Train, IFAC Proc. Vol. 22 (1989) 9–15.

[100] G. Bano, P. Facco, F. Bezzo, M. Barolo, Probabilistic Design space determination in pharmaceutical product development: A Bayesian/latent variable approach, AIChE J. 64 (2018) 2438–2449.

[101] PRO/II Casebook #1 Vinyl Chloride Monomer Plant, in: PRO/II Caseb. #1, SIMULATION SCIENCES INC, 1992.

[102] F. Arteaga, A. Ferrer, Building covariance matrices with the desired structure, Chemom. Intell. Lab. Syst. 127 (2013) 80–88.

[103] F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemom. Intell. Lab. Syst. 101 (2010) 38–42.

[104] J. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. Vis, A. K. Smilde, E.J.J. Velzen, J. Duynhoven, F. van Dorsten, Assessment of PLSDA cross validation, Metabolomics. 4 (2008) 81–89.

[105] P. Zerzucha, M. Daszykowski, B. Walczak, Dissimilarity partial least squares applied to non-linear modeling problems, Chemom. Intell. Lab. Syst. 110 (2012) 156–162.

[106] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148.

[107] R.D. Cook, C.J. Nachtrheim, A comparison of algorithms for constructing exact d-optimal designs, Technometrics. 22 (1980) 315–324.

[108] M.E. Johnson, C.J. Nachtsheim, Some guidelines for constructing exact d-optimal designs on convex design spaces, Technometrics. 25 (1983) 271–277.