1290

UNIVERSIDADE Đ
COIMBRA

Maria Bravo de Almeida Pereira Campos

Multiblock methods for handling complex and heterogeneous data structures in industry

Doctoral Thesis in Chemical Engineering under the supervision of Professor Doctor Marco Paulo Seabra dos Reis, presented to the Department of Chemical Engineering, Faculty of Sciences and Technology of the University of Coimbra

December 2022

Faculty of Sciences and Technology

Multiblock methods for handling complex and heterogeneous data structures in industry

Maria Bravo de Almeida Pereira Campos

Doctoral Thesis in Chemical Engineering under the supervision of Professor Doctor Marco Paulo Seabra dos Reis, presented to the Department of Chemical Engineering, Faculty of Sciences and Technology of the University of Coimbra

December 2022

1 2 9 0

UNIVERSIDADE Ð COIMBRA

"The goal is to turn data into information and information into insight"

Carly Fiorina

Acknowledgments

It has been a long and challenging journey to conclude the work presented in this thesis. While working full time in jobs with increasing responsibility, moving from one country to another and having a son in the meantime, I am very proud that I have finally achieved this important milestone. I would like to thank all who have contributed to my journey, motivating, and challenging me every day. Without their support this would not have been possible.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Doctor Marco Reis, for his invaluable guidance, insights, ideas, and unwavering support and belief in me. His constant feedback and encouragement were key to help me achieve this.

I am grateful to my colleagues and friends that enhanced my experience during these years, providing discussions, support and an environment that was both stimulating and enriching. My special thanks and appreciation go to Ricardo Sousa and Ricardo Rendall. I would also like to include in my acknowledgements Pedro Felizardo and Prof. José Cardoso Menezes for convincing me to do a PhD and for introducing me to the topic and to my supervisor.

I am also very thankful to my family who always gave me their full support and the strength I needed in all the phases of my life. To my son, Vicente, I hope I can pass on the example that "*Success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do*." *—Pele, Brazilian soccer player*

Abstract

The amount of data collected from industrial processes and analytical chemistry have increased by orders of magnitude during the last 50 years due to the fast development of computers and measuring systems. In many areas of industry and life sciences, data sets are collected that can be naturally grouped in multiple blocks. Examples are abundant in many different areas of science: processes with multiple unit operations, bioprocesses conducted through multiple steps, products characterized by several analytical methods, and biological systems of which different-omics measurements are obtained.

These multiple blocks of data (subsets of variables that can be rationally clustered in distinct groups) have their own specific correlation structures and constitute entities within the system that may act or interact in some way to establish the final properties of the product. In this context, the integrity of such blocks of data should be preserved throughout the analysis, i.e., instead of manipulating individual variables during model building, we argue that it is more natural and consistent with the system nature to manipulate blocks of functionally related variables.

Methods that are able to accomplish this endeavor are called multiblock methods and they constitute the focus of this thesis. Multiblock models strive to maintain the natural ordering in the data with the objective of keeping track of the different blocks during the analysis in the same way as one keeps track of individual variables in classical multivariate data analysis. By integrating these different data blocks into the modeling and keeping their integrity intact, more parsimonious and informative models can be developed. Multiblock modeling methods are also useful to explain the relationships between different blocks, and the relative contribution of each block in the model. This leads to more informative insights in the end and increases model interpretability more than any result obtained by the individual analysis of each data set (or data source). Moreover, explicit information about the common and unique variation from each block of predictor can be extracted.

Several multiblock methods based on latent variables have been suggested in the literature since their introduction several decades ago, offering more efficient solutions for a variety of problems many professionals face nowadays. However, there is not a clear workflow for their selection and application, and some important technical aspects have not been addressed properly.

Therefore, in this thesis a systematic workflow for the development of multiblock modelling is proposed including a three-level approach for selecting the adequate pre-processing in multiblock modelling approaches. In multiblock models we face the additional complexity of having to deal with inter-block variability (i.e., between blocks variability) in addition to the intra-block variability (i.e., within blocks variability) in order to avoid the model outcomes to be impacted by aspects that are not directly related to the phenomena of interest but to data related issues (e.g., number of variables in each block, units, etc.). The strategy proposed in this thesis proceeds from handling intra-block effects regarding data quality (Level I) and variables' balancing (Level II) to the equalization and tuning of the inter-block variability (Level III). Moreover, new, and more robust Level III pre-processing methods are proposed and compared with current state-of-the-art block scaling approaches in the scope of two real case studies (Chapter 5).

This thesis also provides a comprehensive and critical literature review focused on multiblock approaches followed by an extensive comparison assessment of state-of-the-art multiblock methodologies with regard to their prediction and interpretability capabilities by means of a robust statistical framework (Chapter 6). The following state-of-the-art methods are explored: Concatenated PLS method, Hierarchical PLS (HPLS) (Slama, 1991, Wold et al., 1987b, Wold et al., 1996), Multiblock PLS (MBPLS) (Wangen and Kowalski, 1989, Wold et al., 1983), Network-Induced Supervised Learning (NI-SL) (Reis, 2013b), and Sequential Orthogonalized PLS (SO-PLS) (Jørgensen et al., 2004, Næs et al., 2011b, Jørgensen et al., 2007). Limitations and improvement opportunities of these methods are highlighted and discussed.

Furthermore, as part of the contributions of this thesis, a new and more efficient multiblock methodology is presented, called Stepwise SO-PLS. This methodology conducts multiblock predictive analysis, overcoming several limitations found in current state-of-the-art methods, such as the issues of selecting the proper block order, finding out the blocks to left out, and mitigating the impact of inter-block scaling. The new method is tested on real data and the results are fully discussed (Chapter 7).

In summary, the contributions of this thesis represent an effort towards bringing multiblock data analysis to the forefront of advanced analytical methods to adopt in modern industrial data science problems, empowering practitioners with systematic and efficient frameworks to handle problems where variables can be naturally organized in blocks.

Keywords: Multiblock modelling, Latent variable methods, inter-block and intra-block pre-processing, Concatenated PLS method, Hierarchical PLS (HPLS), Multiblock PLS (MBPLS), Network-Induced Supervised Learning (NI-SL), and Sequential Orthogonalized PLS (SO-PLS), Stepwise SO-PLS.

Resumo

A quantidade de dados recolhidos em processos industriais e a tecnologia de instrumentação em química analítica tem aumentado consideravelmente nas últimas cinco décadas devido ao desenvolvimento rápido da computação e sistemas de medição. Em muitas áreas da indústria, e das ciências naturais e da vida, os conjuntos de dados coletados estão naturalmente organizados em blocos de variáveis. Exemplos abundam e podem ser encontrados em diferentes áreas da ciência: processos industriais constituídos por uma sequência de operações unitárias, bioprocessos que têm lugar em várias fases, produtos caracterizados por vários métodos analíticos ou sistemas biológicos dos quais são obtidas diferentes medidas -ómicas, entre outros.

Estes blocos de dados (subconjuntos de variáveis que podem racionalmente ser reunidas em diferentes grupos) têm a sua própria estrutura de correlação e constituem entidades naturais dentro do sistema em análise que podem atuar ou interagir de alguma forma para estabelecer as propriedades finais do produto. Deste modo, a sua integridade deve ser levada em conta na construção de modelos e na sua análise subsequente. Ou seja, ao invés de se manipular variáveis individualmente durante a construção do modelo, o mais natural e consistente com a natureza do sistema é manipular blocos de variáveis funcionalmente relacionados.

Métodos que são capazes de modelar os dados preservando a sua estrutura natural são chamados de métodos multibloco e são o tema principal desta tese. Os modelos multibloco caracterizam-se por manter a ordem natural dos dados com o objetivo de preservar os diferentes blocos durante a análise, da mesma forma que se mantém o controle sobre as variáveis individuais numa análise de dados multivariada clássica. Ao integrar esses diferentes blocos de dados na modelação mantendo a sua integridade, criam-se condições para obter modelos mais parcimoniosos e informativos. Estes modelos também são úteis para explicar a relação entre os diferentes blocos, e a contribuição relativa de cada bloco no modelo, potenciando uma maior extração de maior informação e a uma maior interpretabilidade dos resultados do modelo relativamente a uma análise clássica. Além disso, informações explícitas sobre a variação comum e única de cada bloco passam a ser conhecidas.

Vários métodos multibloco baseados em variáveis latentes têm vindo a ser descritos na literatura desde a sua introdução várias décadas atrás, oferecendo soluções mais eficientes para os problemas que muitos profissionais continuam a enfrentar atualmente. No entanto, não há uma metodologia sistemática estabelecida para sua seleção e aplicação, e alguns aspetos técnicos importantes não foram ainda abordados adequadamente.

Nesta tese, propõe-se uma metodologia sistemática para o desenvolvimento de modelos multibloco, incluindo uma abordagem em três níveis para selecionar o pré-processamento adequado em contextos de multibloco. Em modelos multibloco, para além da variabilidade intra-bloco (ou seja, variabilidade dentro de cada bloco) existe a complexidade adicional de lidar com a variabilidade inter-bloco (ou seja, variabilidade entre blocos), para evitar que os resultados do modelo sejam afetados por aspetos não relacionados com o fenómeno de interesse, mas com os dados (número de variáveis em cada bloco, unidades utilizadas, etc.). A estratégia proposta nesta tese abrange desde o tratamento dos efeitos intra-bloco relativos à qualidade dos dados (Nível I), passando pelo balanceamento da escala das variáveis (Nível II) até à equalização e afinação da variabilidade inter-blocos (Nível III). Novos métodos robustos de pré-processamento de Nível III são também propostos e comparados com as abordagens de escalonamento de blocos pertencentes ao estado-da-arte, em dois casos de estudos reais (Capítulo 5).

Esta tese fornece também uma revisão abrangente e crítica da literatura sobre abordagens multibloco, bem com uma avaliação comparativa extensiva de metodologias multibloco considerados como estado-da-arte em relação às suas capacidades de previsão e interpretabilidade (Capítulo 6). Os seguintes métodos multibloco são explorados nesta tese: método PLS concatenado, PLS hierárquico (HPLS) (Slama, 1991, Wold et al., 1987b, Wold et al., 1996), PLS multibloco (MBPLS) (Wangen e Kowalski, 1989, Wold et al., 1983), Indução de Rede para Aprendizagem Supervisionada (NI-SL) (Reis, 2013b) e PLS Ortogonalizado Sequencial (SO-PLS) (Jørgensen et al., 2004, Næs et al., 2011b, Jørgensen et al., 2007). Algumas limitações e oportunidades de melhoria desses métodos são destacadas.

Será ainda apresentada uma nova metodologia mais eficiente para realizar análises preditivas multibloco chamada SO-PLS passo a passo, que supera várias limitações encontradas nos métodos atuais de última geração, como o problema de estabelecer a melhor ordem para análise dos blocos, o problema de selecionar os blocos a analisar e a descartar, e o desafio de mitigar o impacto do escalonamento entre blocos. O novo método é testado em dados reais e os resultados são totalmente discutidos (Capítulo 7).

Em resumo, as contribuições desta tese representam um esforço para colocar a análise de dados multibloco na vanguarda dos métodos analíticos avançados a serem adotados no âmbito da ciência dos dados industriais, capacitando os profissionais com metodologias e ferramentas sistemáticas e eficientes para lidar com problemas em que as variáveis podem ser naturalmente organizadas em blocos.

Palavras-Chave: Modelos de múltiplos blocos de dados, métodos com base em variáveis latentes, pré-processamento inter-bloco e intra-bloco, método PLS concatenado, PLS hierárquico (HPLS), PLS multibloco (MBPLS), Indução de Rede para Aprendizagem Supervisionada (NI-SL) e PLS Ortogonalizado Sequencial (SO-PLS), SO-PLS passo a passo

Table of Contents

List of Acronyms

COMFA	Combinatorial molecular field analysis
CPCA	Consensus principal components analysis
CV	Cross validation
DISCO	Distinct and common simultaneous component analysis
GC-MS	Gas chromatography coupled to mass spectrometry
GCA	Generalized canonical correlation analysis
HBRS	Hard block rank scaling
HBS	Hard block scaling
HBVS	Hard block variance scaling
HPLS	Hierarchical partial least squares
IQR	Interquartile range
JIVE	Joint and individual variances explained
K-fold-CV	K-fold cross-validation method
KPI	Key performance indicator
LOOCV	Leave-one-out cross validation
LV	Latent variable
MBPLS	Multiblock partial least squares
MVP	Multiblock variance partitioning
NI-SL	Network induced supervised learning regression method
NIPALS	Non-linear iterative partial least squares
NIR	Near infra-red spectroscopy
O2PLS	Orthogonal two-block partial least squares
OA	Organic acids
OnPLS	Orthogonal n-block partial least squares
P-ComDim	Predictive common dimensions method
PC	Principal components
PCA	Principal components analysis
PCR	Principal components regression
PLS	Partial least squares
PLS-2H	Hierarchical two-block predictive partial least squares
PO-PLS	Parallel orthogonalized partial least squares
PPBS	Pseudorank penalization block scaling
QSAR	Quantitative structure activity relationships
RMSEP	Root mean square error of prediction
ROSA	Response-oriented sequential alternation
SBRS	Soft block rank scaling
SBS	Soft block scaling
SBVA	Soft block variance scaling
SCA	Simultaneous component analysis
SHBRS	Super hard block rank scaling
SHBS	Super hard block scaling
SHBVS	Super hard block variance scaling
SIMPLS	Straightforward implementation of a statistically inspired
SLIDE	Structure learning and integrative decomposition
SNV	Standard normal variate
SO-PLS	Sequential orthogonalized partial least squares
SPLS	Serial partial least squares
UV-Vis	Ultraviolet-visible spectroscopy

List of Figures

List of Tables

Chapter 1 - Introduction

Chapter 1. Introduction

The first chapter of this thesis provides an overview of the main themes discussed throughout the document and is divided in four sub-sections. In the first sub-section, the scope and motivations for this work are set, highlighting the importance of multiblock modelling in different disciplines. Considering this importance, the second sub-section presents the pursued objectives followed by the description of the main contributions of this work in the third sub-section. Finally, the last sub-section describes the organization of this document.

1.1 Scope and Motivation

The technological advances in the instrumentation employed in industry and life sciences have enabled the collection of an unprecedented amount of data from multiple sources. This data is valuable for process optimization, product quality prediction and real-time monitoring and control of industrial units. However, in practice, process operators and engineers are overwhelmed, not only because of the vast amount of data continuously being generated but also because of the complex structure of the data gathered and their heterogeneity. Very often, data can be naturally grouped into different blocks of descriptor variables according to an underlying rational (e.g., according to their location in a process or in a reactor or according to their similarity). The grouping of data into meaningful blocks can be known from *a priori* information or learned using data driven methods. This type of structured data is called multiblock data and is the focus of this research – see Figure 1-1 for some examples of such data structures. Examples of multiblock data can be found in analytical chemistry where different analytical methods (e.g. different spectroscopic and chromatography data such as GC-MS and HPLC, etc.) can be used to provide a detailed and complete description of the products under investigation (Campos et al., 2017). In fact, with such a comprehensive use of analytical techniques on the same set of samples, multiple predictor and response blocks are produced that carry potentially meaningful information for the analysis. Typical examples can also be found in food science, where the same products may be characterized by sensory attributes, instrumental measurements, and consumer acceptance (Næs et al., 2011a). In medical science, the blocks might correspond to different –omics platforms as well as clinical measurements and lifestyle variables of the same patients (Song et al., 2020). In quantitative structure activity relationships (QSARs) multiple blocks of 3D-descriptors are often produced to parameterize the variation in the chemical structure of a small number of compounds (Schmidt et al., 2021). These structure descriptors may then be used simultaneously to predict interesting biological activities. In combinatorial molecular field analysis (COMFA) several probes are used at every grid point to give extra information about the molecules. The grid points of each probe can also be divided into separate blocks (Westerhuis, 1997). Multiblock data sets are also abundant in systems biology: in metabolomics it is increasingly common to measure the same set of samples on different analytical platforms to obtain a comprehensive view of the metabolites in those samples (Smilde et al., 2005, Crockford et al., 2006) one can also study functional genomics measurements of the same type performed in different organisms (Alter et al., 2003), or in different compartments of the same organism, for example, in plasma and tissue (Noguchi et al., 2006). Finally, examples can also be found in modern industry in batch and in continuous processes. In continuous processes measurements are made at various points of the process (unit operations) and in different compartments of major pieces of equipment. In batch processes, the raw materials and the intermediate products both give information about the end-product, constituting different blocks, and different data sources can also arise from the different process unit steps.

In this perspective the information from the different data sources (i.e., blocks of data) must be combined by means of efficient data driven methods in such a way to achieve the desired goal of the analysis (being it prediction or exploratory analysis) as well as in order to extract maximum information from the system or process under study. Methods capable of modelling the different data blocks together while retaining the integrity of each data source are called multiblock methods.

Multiblock methodologies are known under different names in different disciplines, some important examples being data fusion, data integration, multiset analysis, multimode analysis, among other – for definitions, see the works of Van Mechelen and Smilde (2010), Lahat et al. (2015). Multiblock data sets can be denoted by different terms as well, such as: coupled, multi-modal, multi-source, linked, or multiset data (Mishra et al., 2021, Van Mechelen and Smilde, 2010). Throughout this thesis the words multiblock modelling and multiblock data sets will be used consistently.

Several multiblock methods have been published in the literature over the last decades (Westerhuis et al., 1998, Westerhuis, 1997, Rännar et al., 1998, Wold et al., 1998b, Frank et al., 1984, Frank and Kowalski, 1985) and are becoming increasingly popular in the current Big data era and in the context of Industry 4.0 (Lahat et al., 2015, Smilde et al., 2017, Mishra et al., 2021). However, these methods remain still quite underexplored, and their use is largely limited to chemometricians, while non-experts usually have little contact with them. Furthermore, most current data driven methods presented in the literature are primarily focused on the prediction capability of the models and much less on their interpretation capability. In the analysis of complex systems, however, one of the main interests is precisely the induction of relevant associations, to understand or clarify the way the system operates in order to optimize them. Multiblock methods incorporate interpretational-oriented analysis features right from the onset of the analysis, which constrain the predictive space thus improving the interpretability of the models (Reis, 2013a, Reis, 2013b). Interesting enough, constraining the predictive space do not usually compromise the methods' performance, when compared to their unconstrained versions (Campos et al., 2017). Understanding the inter-relationships among the different data blocks and the relative contribution of each block in the model can significantly improve the model interpretability allowing maximum information extraction from the data collected. In addition, most often the interest is in assessing the commonalities (i.e. the overlapping information) and differences (i.e. the uniqueness of each data block) between the different data sets, also taking into account their linking relations (Alinaghi et al., 2020, Måge et al., 2019, Smilde et al., 2017, Song et al., 2020, Rännar et al., 1998, Westerhuis et al., 1998).

Handling these types of data structures raises several challenges namely on how to efficiently fuse the information from the different blocks and explore their inter-relationships. In the development of multiblock modelling approaches the pre-processing strategy applied is of paramount importance because it constitutes the critical-to-quality driver to scrutinize and optimize. The best pre-processing method to be applied in each case must be selected in an efficient way, and in the context of multiblock methods with very complex and heterogeneous data blocks this is even of higher importance since it is highly dependent on the characteristics of the data sets at hand. In cases where the data structures have different underlying dimensionality and sizes, the pre-processing applied can significantly hinder the outcome of the modelling if not handled appropriately. In other words, when multiple blocks exist and are to be integrated together in the model there is the increase challenge of handling the inter-block (i.e., between blocks) variability in addition to the intra-block (i.e., within blocks) variability. Moreover, the possibility of excluding blocks of data that do not add value to the modelling scheme is of interest in multiblock modelling (even more so when the number of data blocks available is significantly high) since it will lead to models that are simpler, more parsimonious, and easier to interpret. Consequently, the selection and application of pre-processing within the multiblock contexts is a topic worth to explore and devote research efforts. This has motivated the development of a systematic workflow for the application of pre-processing in multiblock modelling in this thesis. In addition, as the current state-of-the-art pre-processing methods for multiblock modelling have some drawbacks, new efficient pre-processing methods to account for the inter-block variability were proposed and tested in this thesis.

Another motivation of this thesis is to increase the awareness and understanding of the potential of multiblock methods. Therefore, a thorough systematic and extensive comparison of a carefully selected pool of state-of-the-art multiblock methodologies in terms of their predictive and interpretative capabilities was performed thesis. In fact, there is currently a relative lack of information regarding the relative potential of the existing multiblock methods. Thus, the methods were compared using real data and under a robust statistical framework to rigorously establish the methods predictive relative performances and their ability to bring additional interpretation features to the analysis. Some identified limitations were also addressed and a proposal of a new methodology for multiblock modelling was put forward. This new methodology is more efficient (less computational steps, hence faster execution time and less computational effort) and overcomes some of the main drawbacks from the analysed state-of-the-art methods.

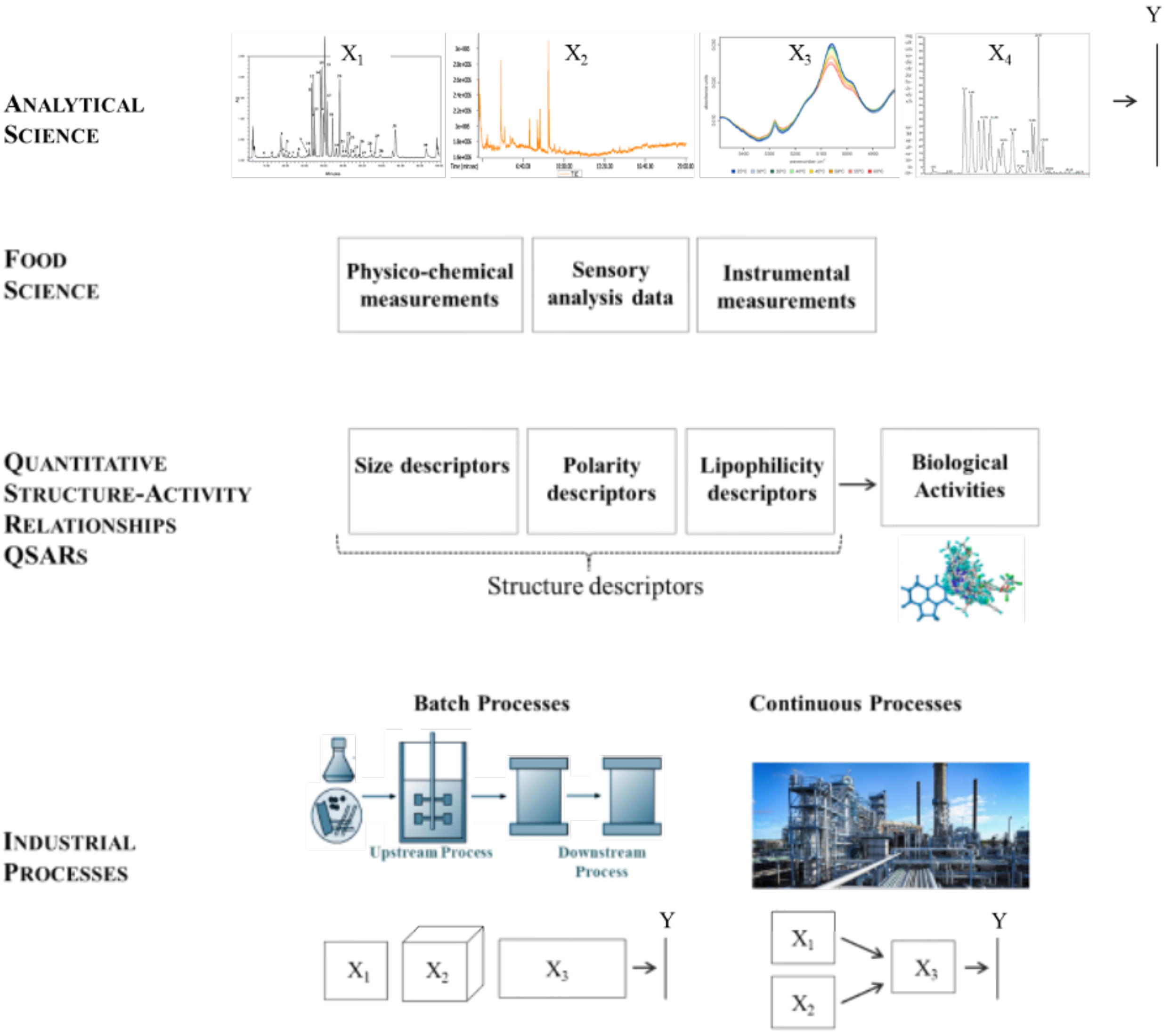


Figure 1-1 – Examples of multiblock data in food and analytical science, QSARs and industrial processes.

1.2 Thesis Goals

The research reported in this thesis aims at contributing to the progress of multiblock data analysis, including a thorough critical assessment of state-of-the-art multiblock methodologies in order to explore their prediction and interpretation capabilities and to address some of their limitations by proposing new and more efficient and effective solutions. Therefore, the following specific goals were defined for this research work:

- Generate critical information to assist practitioners on the selection of currently available data driven methodologies capable of handling data naturally organized as multiple blocks of variables (called multiblock data sets);
- Identify limitations and improvement opportunities found in state-of-the-art multiblock methods;

- Develop new and more efficient methodologies that address the limitations identified;
- Propose an integrated workflow for multiblock modelling and analysis with particular focus on the efficient selection of the best pre-processing strategy;
- Establish a robust comparison framework to compare and test the proposed solutions against current state-of-the-art methods.

The methodologies proposed in this thesis are envisioned to be effectively applied in real world scenarios and, therefore, should comply with the requirements of being robust and easy to implement. Moreover, the systematization of the workflow for the application of multiblock modelling approaches including pre-processing strategies should contribute to making these methods more accessible to practitioners so that they can take most out of the available data, but also exploiting the systems structure and existing *a priori* knowledge.

1.3 Thesis Contributions

In accordance with the general goals described above, the following developments can be considered as the main contributions from this thesis:

(I) A systematic pipeline for multiblock analysis, including a three-stage approach for defining the pre-processing for multiblock modelling;

(II) New block scaling methods: block variance scaling; block rank scaling and the pseudorank penalization block scaling;

(III) A thorough and systematic literature review focused on multiblock approaches followed by and extensive comparison assessment of the most promising methods to identify their relative strengths and weaknesses with regard to both prediction and interpretability capabilities;

(IV) A stepwise approach for a more parsimonious, efficient, and interpretable implementation of Sequential and Orthogonalized PLS (SO-PLS) that addresses the definition of the order of the blocks and the selection of the data blocks to be incorporated in the model. Furthermore, a new variant of the standard SO-PLS approach was also proposed capable of excluding non-informative blocks.

The scientific papers associated with each contribution are presented in Table 1-1.

Table 1-1 - Published papers associated with each of the contributions made in this thesis.

Contribution	Reference
(I), (II)	Maria P. Campos, Marco S. Reis, *Data Pre-processing for Multiblock Modelling – A Systematization with New Methods*, Chemometrics and Intelligent Laboratory Systems (2020)
(III)	Maria P. Campos, Ricardo Sousa, Ana C. Pereira, Marco S. Reis, *Advanced predictive methods for wine age prediction: Part II – A comparison study of multiblock regression approaches*, Talanta (2017)

Contribution	Reference
(IV)	Maria P. Campos, Ricardo Sousa, Ana C. Pereira, Marco S. Reis, *Establishing the Optimal Blocks' Order in SO-PLS: Stepwise SO-PLS and Alternative Formulations*, Journal of Chemometrics (2018)

1.4 Thesis Overview

The present thesis is organized in nine chapters that can be grouped as shown schematically in Figure 1-2.

Chapter 1 is the introductory chapter of this thesis setting the motivation and general scope of the work presented in this thesis, as well as the main goals and contributions.

Chapter 2 includes a description of the background knowledge on the data driven multivariate methods used throughout this research including the statistical comparison framework implemented to test the different methods and perform a rigorous comparison of their relative predictive performances.

A structured and thorough review of the state-of-the-art multiblock modelling approaches and respective applications published in the literature is presented in Chapter 3.

In Chapter 4 two real case studies are presented that were used throughout this thesis to illustrate the applicability of the proposed methodologies.

The work presented in Chapter 5 to Chapter 7 can be summarized in a workflow that covers the pre-processing step for multiblock modelling followed by an extensive modelling comparison of current state-of-the-art multiblock models by means of a robust statistical framework while exploring their predictive and interpretation capabilities, culminating in the last stage of the workflow, regarding methodologic improvement, with the proposal of a more efficient method called Stepwise Sequential and Orthogonalized PLS (Stepwise SO-PLS), that addresses some limitations found during the extensive critical assessment of multiblock methods.

Finally, Chapter 8 summarizes the main conclusions of this thesis and puts forward directions for future research.

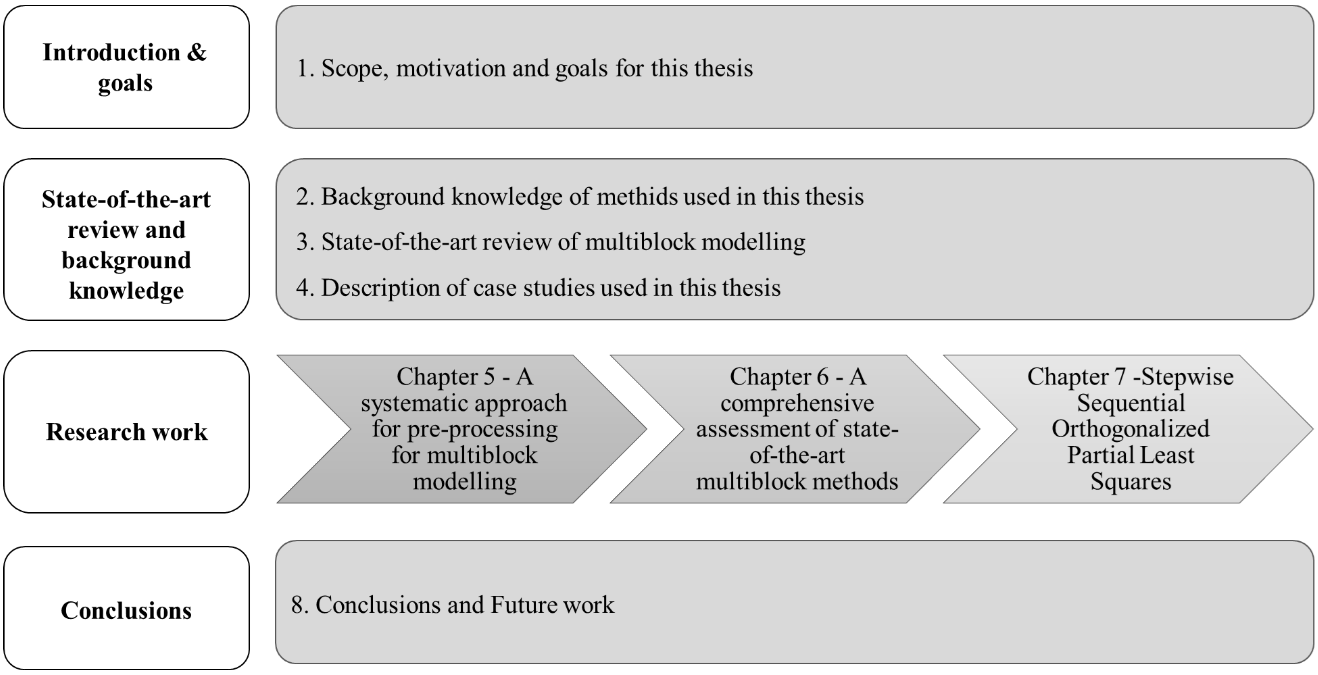

Introduction & goals
1. Scope, motivation and goals for this thesis
State-of-the-art review and background knowledge
2. Background knowledge of methids used in this thesis
3. State-of-the-art review of multiblock modelling
4. Description of case studies used in this thesis
Research work
Chapter 5 - A systematic approach for pre-processing for multiblock modelling
Chapter 6 - A comprehensive assessment of state-of-the-art multiblock methods
Chapter 7 -Stepwise Sequential Orthogonalized Partial Least Squares
Conclusions
8. Conclusions and Future work

Figure 1-2 - Overview of the organization of this thesis

Chapter 2 – Background Knowledge

Chapter 2. Background knowledge

This chapter presents a short description of the data driven methods used, as background knowledge for this thesis. The most common traditional chemometric methods such as principal components analysis (PCA), partial least squares (PLS) and their variants, are examples of single block techniques. Single block methods make no explicit use of the variables natural grouping. These methods are presented in Section 2.2. Multiblock methodologies, on the other hand, can deal with multiple blocks of data and are capable of retaining their integrity in the model. These methods keep track of the different data blocks similarly to what single block methods do with each single variable. The state-of-the-art multiblock methods used in this thesis are described in Section 2.3.

2.1 Notation

A consistent notation scheme is adopted for all methodologies referred in this thesis. Data matrices are denoted by bold uppercase characters (e.g., **X**, **Y**), vectors are represented as columns and denoted with bold lowercase characters (e.g., **p, t, c**) and scalars are represented as lowercase characters. The characters n, p, m, b and a, are relative to the number of objects, the number of explanatory variables (or predictors), the number of response variables, the number of predictor blocks and the number of retained latent variables, respectively.

2.2 Single block latent variable methods

The most commonly used chemometric data driven methods belong to the group of latent variables methods and are based on the assumption that a few underlying unobserved variables (called latent variables) are responsible for the observed variability on both **X** (input) and **Y** (output) variables (Burnham et al., 1996, Burnham et al., 1999, Burnham et al., 2001). The latent variables can be estimated through linear combinations of the measured original variables. This class of methods is appropriate to handle data sets with high levels of collinearity, since in this case the reduced set of latent variables is able to efficiently extract the main patterns of variation and association observed. The three most well-known methods from this class are: principal components analysis (PCA) (Wold et al., 1987a), principal components regression (PCR) (Jackson, 2005, Jolliffe, 2005) and partial least squares (PLS) (Geladi and Kowalski, 1986, Wold et al., 2001a, Wold et al., 2001b). PCA is an exploratory data analysis method, whereas PCR and PLS are both regression methods.

There are several other classes of data driven methodologies for single block data sets that can be adopted depending on the purpose (prediction or classification) and the specific characteristics of the data set under analysis, namely the degree of sparsity, linearity and collinearity of the regressors; for more information on these, please refer to Rendall et al. (2017). In the present thesis we focus on the class of linear regression methods based on latent variables.

2.2.1 Principal Components Analysis (PCA)

Principal components analysis (PCA) (Wold et al., 1987a, Jackson, 1991, Jolliffe, 2002) is the basic methodology for multivariate data analysis in high-dimensional settings, and the favourite tool of chemometricians for dimensionality compression and information extraction. PCA searches for directions (also known as principal components) of maximum variance in the space of the variable under analysis, say **X** (n×p). The principal components are the projections of the original variables onto such directions and consist of simple linear combinations of the raw variables (usually after some pre-processing). These projections are usually called *scores*, and the coefficients of the linear combinations, *loadings*. The scores are uncorrelated and may reflect an underlying phenomenon that is not directly measured but estimated from the available data. Thus, they provide a good summary of the predictors' space with the advantage that a relatively small number of principal components (a) is usually required to describe a relevant fraction of the total **X**-variation. Because of this feature, PCA is often used as a dimensionality reduction technique.

Mathematically, PCA relies on an eigen-decomposition of the covariance matrix of **X**, possibly after some adequate pre-processing. PCA provides an approximation of the data matrix, **X**, in terms of the product of two small matrices: the matrix of scores, **T**, and the matrix of loadings, **P**. These matrices, **T** and **P**, capture the essential data patterns of **X**, both in the observations mode (**T**), and in the variables' mode (**P**).

The PCA model is presented in equation (1):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where, **T** is a n×a score matrix and corresponds to the orthogonal projections of **X** into the subspace spanned by the principal components, **P** is a p×a matrix of loadings and contains the directions of maximum **X**-variability, and **E** is a n×p matrix of residuals. Note that the scores values are orthogonal, i.e., uncorrelated to each other. The residual matrix contains the variation in the data that could not be extracted by the principal components and is usually interpreted as representing unstructured variation or noise.

In PCA, the maximum number of components that can be extracted from **X** corresponds to the number of linearly independent columns or rows there are in **X**, i.e., the rank of **X**. As components are extracted in order of the amount of variation they explain, the extraction generally stops when the remaining components explain too little variation. The threshold for this is case-specific, nontrivial to determine, and often done by consideration of several criteria in parallel (Mosier, 1951, Wold, 1976, Wold, 1978, Bro et al., 2008).

2.2.2 Principal Components Regression (PCR)

Principal components regression (PCR) (Jackson, 2005, Jolliffe, 2005) extends PCA to regression problems. PCR is a latent variable regression method that consists of applying PCA to the **X** variables and then using the resulting components (called scores) as regressors to predict the target response **Y**. The number of principal components (a) is the method's hyper-parameter controlling model complexity and is usually selected through cross-validation.

PCR is a two-step method and thereby has the risk that some useful (predictive) information will end up being discarded in the principal components that were not selected and that some noise will remain in the components used for regression.

2.2.3 Partial Least Squares (PLS)

The pioneering work of partial least squares (PLS) was largely done by Wold et al. (1983); see also Martens, H., & Naes, T. (1989) and Geladi, P., & Kowalski, B. R. (1986). Similarly, to PCA, PLS also searches for directions in the **X**-space. However, instead of searching for directions that explain the **X**-variation, it looks for directions presenting maximum covariance with the response, **Y**. Therefore, in addition to minimizing the prediction residuals, PLS also gives a good approximation of the **X** space. PLS assumes that both **X** and **Y** are governed by the same latent structure and the available data are used for estimating it. The latent variables are related to the response variable (example for one singe value) according to equation (2).

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E}$$

$$\mathbf{y} = \mathbf{Tc} + \mathbf{f} \tag{2}$$

Where **T**, **P** and **E** are respectively the PLS **X**-scores, **X**-loadings and **X**-residuals, **c** is the **y**-loading vector and **f** represents the **y**-residuals, which are minimized (PLS can also handle multiple responses, **Y**, in which case **c** and **f** gives rise to **C** and **F**). Beyond minimizing the **f** residuals, PLS also provides a good approximation of the **X** space. As in PCA, the loadings can be analysed to determine which original variables in **X** are most strongly related to **Y**.

The number of latent variables (a) is the hyper-parameter controlling model complexity and must be appropriately chosen to capture the relevant variation without modelling noise, usually using cross-validation (Mosier, 1951, Wold, 1976, Wold, 1978, Bro et al., 2008).

The two most popular algorithms for estimating the PLS parameters is the non-linear iterative partial least squares, NIPALS (Wold, Martens & Wold, 1983) algorithm and the SIMPLS algorithm (De Jong, 1993). The pseudocode for the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm to estimate the PLS model, is presented in Table 2-1. Several other PLS algorithms have been published in the literature for the single response case (Andersson, 2009). The algorithms vary in terms of speed and numerical stability, and some have been developed for specific purposes, e.g., handling matrices of particular sizes (De Jong, 1993, Andersson, 2009, Rännar et al., 1994)

Table 2-1 - Pseudocode for computing PLS using the NIPALS algorithm (single response case).

```
(1)   Center and scale X and y
For each latent variable (LV= 1:a)
Loop until convergence of $t$
(2)   $w = X^T y$                      % X weights ($w$)
(3)   Scale w to $||w|| = 1$
(4)   $t = Xw$                         % X scores ($t$)
(5)   $c = y^T t/t^T t$                % y loading vector ($c$)
(6)   $p = X^T t/t^T t$                % X loadings ($p$)
(7)   $X=X-tp^T$
(8)   $y=y- tc^T$
```

2.3 Multiblock latent variable methods

This section presents a description of the state-of-the-art multiblock algorithms published in the literature that were used in the work reported in this thesis. The most common multiblock approaches are based on latent variable models as it was proven decades ago to be a very powerful paradigm to describe data arising from many practical applications. When considering multiple blocks of data, each block is summarized by its latent variables (or components) and the relationships between the blocks are then modelled by establishing the relationships between those latent variables (or components). Latent variables approaches have the following benefits:

- ✓ The number of sources of variability in data blocks is usually (much) smaller than the number of measured variables.
- ✓ Can handle data with high collinearity, i.e., variables that are not independent, but instead correlate strongly with each other.
- ✓ They can be used when the number of variables is larger than the number of observations (i.e., n > m).
- ✓ Latent variables-based methods are suitable for interpretation through the scores and loadings associated with the extracted components.
- ✓ Underlying latent variables are appropriate for mental abstractions and interpretation.
- ✓ Multivariate data analysis becomes numerically stable and statistically robust if the latent variables are chosen in a suitable way.
- ✓ The effect of measurement noise is reduced.

- ✓ Outliers can often be detected by visual inspection of the associated subspace projections provided by the extracted latent variables together with the associated residuals.
- ✓ Reducing complexity and simplifying analysis

2.3.1 Concatenated PLS

Concatenated PLS consists in concatenating all data blocks in a single augmented matrix and apply classical PLS method described in Section 2.2.3. It is important to emphasize that the different blocks should be properly weighted before being used in the model, in order to give equal importance to all or to increase or decrease the importance of a block in a model when there is some prior knowledge available. Therefore, the selection of the appropriate block scaling method is a critical aspect in this modelling approach. For instance, if the variance in one block is much larger than all others, this block will dominate the estimation process, and the conclusions can be biased or misleading.

2.3.2 Hierarchical PLS (HPLS)

In the Hierarchical PLS (HPLS) method, each data block is considered as a separate source of information and the task of the multiblock model is to represent the common structure for the objects. This common structure is formulated in a so-called super level (global information), an additional top layer, combining information from all predictor blocks on the lower data level (local information). This means that block scores, loadings and weights for each separate block are available for interpretation in the lower level and super scores, loadings and weights are available in the super level for the interpretation of the global model. The method used in this work is the one presented by Slama (1991), in which a normalization of the super scores is performed instead of the normalization of the super weights, as indicated previously by Wold et al. (1987b). The pseudocode for the HPLS method is presented in Table 2-2 and a schematic representation of the algorithm is shown in Figure 2-1.

Table 2-2 – Pseudocode for the Hierarchical PLS (HPLS) multiblock method.

```
For each latent variable (LV= 1:a):
    The super score $t_T$ is initiated as the largest eigenvalue of
(1) $\mathbf{X}^T\mathbf{X}.p_b = X_b^T t_T / t_t^T t_T$          % $X_b$ block variable loadings ($p_b$)
(2) $t_b = X_b p_b / m_{X_b}^{1/2}$                      % $X_b$ block variable scores ($t_b$)
(3) $T = [t_1 \dots t_B]$                               % Combine all $X_b$ block scores in T
(4) $q = Y^T t_T / t_t^T t_T$                            % Y weight ($q$)
(5) $u = Y q / q^T q$                                   % Y score ($u$)
(6) $w_T = T^T u / u^T u$                               % X super weights ($w_T$)
(7) $t_T = T w_T / w_t^T w_T$                           % X super scores ($t_T$)
        normalize $t_T$ to $\|t_T\| = 1$
    loop until convergence of $\mathbf{t}_T$
    $X_b = X_b - t_T p_b^T$
    $Y = Y - t_T q^T$
```

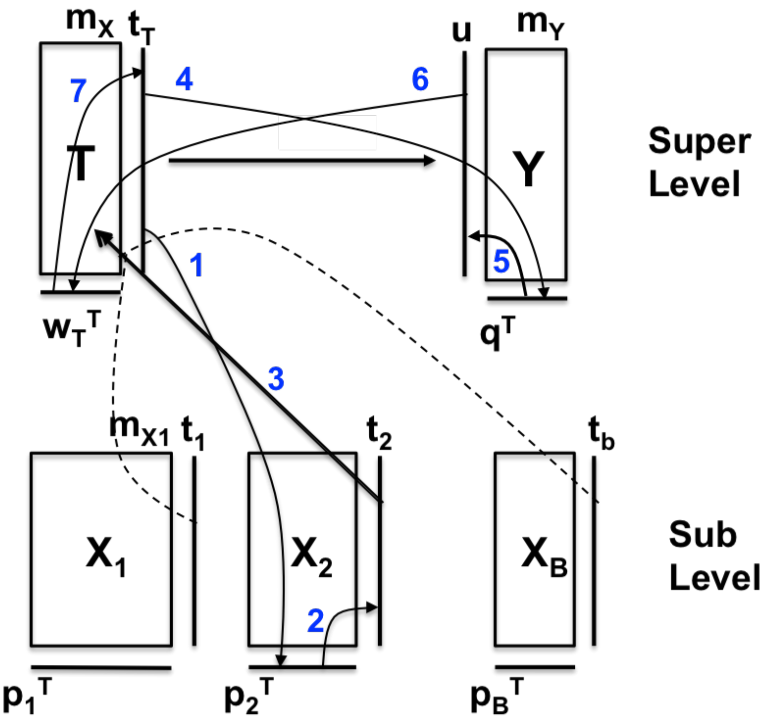


Figure 2-1 - Schematic representation of the HPLS algorithm. The numbers in each arrow represent a step in the pseudocode in Table 2-2.

2.3.3 Multiblock PLS (MBPLS)

Multiblock PLS (MBPLS) was proposed by Wold et al. (1984) and later on by Wangen and Kowalski (1989). Similarly to the HPLS method, this multiblock method also presents two modelling levels: the super level with global information and the lower level with information from each separate block. On each level, PLS **X**-scores and loadings are available for interpretation. The super-weights are useful for the interpretation of the models. They express how much each block contributes to the prediction of the response; high super-weights' values mean high contributions. On the other hand, the inspection of the block scores, block loadings and block weights, allow for the characterization of the individual blocks when the interpretation focus moves to the pertinent blocks and their dominant variables.

The main difference between this method and HPLS is that the **Y** block is regressed on all descriptor **X** blocks, whereas in HPLS the **Y** block is only regressed on the super block, which means that the block scores are calculated in an unsupervised way. This causes the block scores to be different in the two methods. Another difference lies in the normalization used by these two methods: in MBPLS the block variable weights, and super weights are normalized to length one, whereas in the HPLS model only the super scores are normalized. Normalizing block variable loadings (or weights for PLS) seems more appropriate than scores normalization, as this facilitates the comparison between the different blocks (Westerhuis and Coenegracht, 1997).

The multiblock PLS algorithm implemented in this thesis is the one proposed by Wangen and Kowalski (1989). The pseudocode for this algorithm can be found in Table 2-3 and a schematic representation of the algorithm is shown in Figure 2-2.

Table 2-3 - Pseudocode for the Multiblock PLS method (MBPLS)

```
      For each latent variable (LV= 1:a):
      The Y score u is initiated as a column of Y.
(1)   $w_b = X_b^T u/u^T u$                                % Xb block variable weights ($w_b$)
                  normalize $w_b$ to $\|w_b\| = 1$
(2)   $t_b = X_b w_b/m_{X_b}^{1/2}$                        % Xb block variable scores ($t_b$)
(3)   $T = [t_1 \dots t_B]$                                % Combine all Xb block scores in T
(4)   $w_T = T^T u/u^T$                                    % $X$ super weights ($w_T$)
                  normalize $w_T$ to $\|w_T\| = 1$
(5)   $t_T = T w_T/w_t^T w_T$                              % $X$ super score ($t_T$)
(6)   $q = Y^T t_T/t_T^T t_T$                              % Y weight ($q$)
(7)   $u = Y q/q^T q$                                      % Y score ($u$)
loop until convergence of $t_T$

Super Scores Deflation method
$p_{bT} = X_b^T t_T/t_T^T t_T$
$X_b = X_b - t_T p_{bT}^T$
$Y = Y - t_T q^T$

Block Scores Deflation method
$p_b = X_b^T t_b/t_b^T t_b$
$X_b = X_b - t_T p_b^T$
$Y = Y - t_T q^T$
```

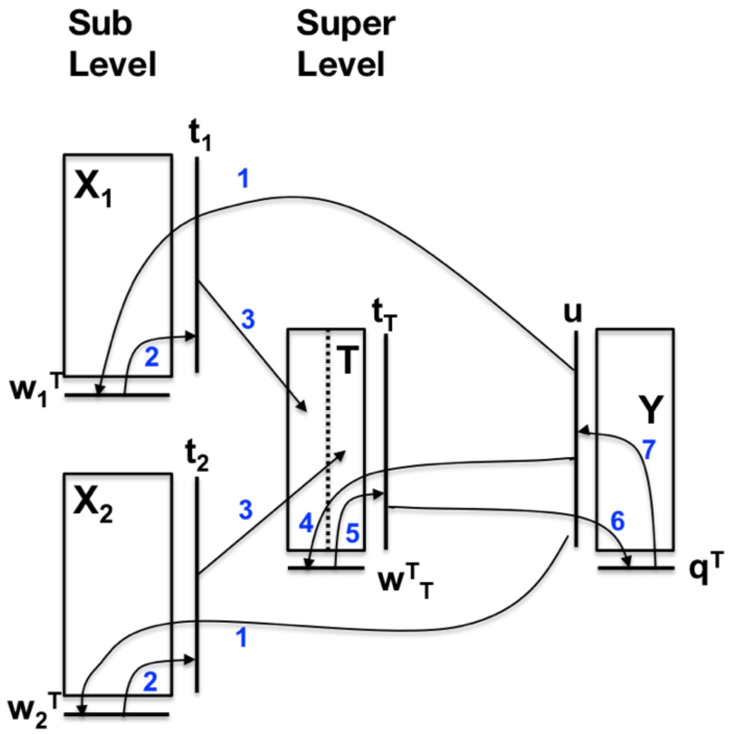


Figure 2-2 - Schematic representation of the MBPLS algorithm. The numbers in each arrow represent a step in the pseudocode in Table 2-3.

Two approaches can be used for the deflation of the blocks in multiblock PLS and both of them will be considered in this thesis. In the first approach, employed by Wangen and Kowalski (1989), the block scores are used for the calculation of the loadings and residuals, and will be referred as the block score deflation method. This approach produces orthogonal block scores, but the super scores are correlated. Westerhuis and Coenegracht (1997), showed that by using the block score deflation method, some of the information in the blocks may be lost in the deflation step, as more variation than that used for the prediction of **Y** is removed from each block. This undesirable effect may become worse as the number of blocks increases, because the individual contribution of each block to the prediction decreases and therefore the fraction of variation removed in each block in relation to the one used for prediction will increase (Westerhuis and Coenegracht, 1997). This loss of information when deflating with the block scores can finally lead to poor performance. Therefore, Westerhuis and Coenegracht (1997) suggested another deflation method, namely using the super scores for the deflation step. In this procedure, one deflates only the information from the blocks that were used for the prediction of the response **Y**. In this approach, the super scores become orthogonal, and the block scores are only slightly correlated. A detailed discussion of the problems related with the different deflation methods in the MBPLS can be found in Westerhuis and Smilde (2001).

Regarding the **Y** deflation, this task must be performed in the case of the block scores deflation method. The deflation prevents the same variation in **Y** from being predicted in different components from different blocks as different blocks may carry similar information. On the other hand, for the super scores deflation method, **Y** deflation makes no difference as one is deflating each block with the information used to predict **Y** (as this information was removed in each **X** block it does not make any difference if it is also removed in **Y** or not).

2.3.4 Network Induced Supervised Learning (NI-SL)

The method presented here is an adaptation of network induced supervised learning regression method (NI-SL) proposed by Reis (2013b), which is a supervised framework aiming at bringing interpretation features to the forefront of the analysis goals. Originally, this method was divided in two stages. Stage 1 consists of the network induced clustering algorithm with the aim of finding functionally related groups of variables (clusters), which will form meaningful **X** blocks. This step is required when no a priori information is available to cluster the variables into conceptually meaningful blocks. The second stage consists of developing a predictive model, based on the analysis of the blocks induced in the first stage. For such, classical PLS models are developed between each **X** block and the **Y** response, one-at-a-time, and a predefined number of latent variables are retrieved from each block (in the present study five latent variables were retrieved from each block). These latent variables are gathered in a super block. Then, forward stepwise regression is used to select the subgroup of latent variables that lead to the best fit. This is a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression (the p-value of the partial F-test) – please refer to Draper and Smith (1998). The pseudocode for this multiblock method is summarized in Table 2-4**Table *2-4*** and a schematic representation of the algorithm is shown in Figure 2-3.

Table 2-4 - Pseudocode for NI-SL

```
      For each latent variable (LV= 1:a):
      The Y score u is initiated as a column of Y.
(1)   $w_b = X_b^T u/u^T u$                          % $X_b$ block variable weights
      normalize $w_b$ to $\|w_b\| = 1$
(2)   $t_b = X_b w_b$                                % $X_b$ block variable scores
(3)   $q = Y^T t_b/t_b^T t_b$                        % Y weight
(4)   $u = Y q/q^T q$                                % Y score
  loop until convergence of $t_b$
(5)   $T_b = [t_1 \dots t_a]$                        % Combine all scores in a augmented score matrix
      $T = [T_1 \dots T_B]$
  Perform forward stepwise selection of T to predict Y
```

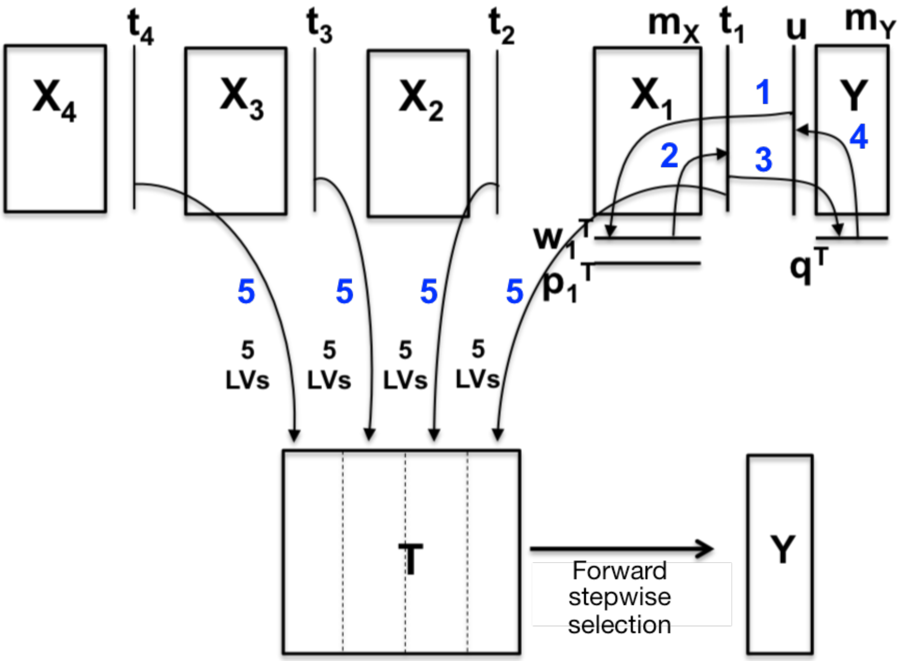


Figure 2-3 - Schematic representation of the NI-SL method. The numbers in each arrow represent a step in the pseudocode in Table 2-4.

2.3.5 Sequential Orthogonalized-Partial Least Squares (SO-PLS)

The version of the Sequential Orthogonalized Partial Least Squares (SO-PLS) method used in this thesis is based on the work of Naes et al. (2013). The pseudocode of this method is summarized in Table 2-5**Table *2-5*** and a schematic representation of the algorithm is shown in Figure 2-4.**Figure *2-3***

This method is focused on incorporating blocks of data, one at a time, and evaluating/interpreting the incremental or additional contribution of the different blocks for improving the model predictions. This is important when one wants to assess the gain of introducing an additional data source into the model. It can also be interesting in modelling industrial processes, for instance in the case where one wants to investigate how much raw material variation adds to the process settings in describing the final product quality. Or in analytical chemistry to understand if the additional information from another analytical method is relevant to fully characterize a product.

The SO-PLS regression method extracts information sequentially from each data block, which means that the chosen order of the blocks can influence the result. When there is no *a priori* information for ordering the blocks, one can try all possible combinations and use them as an additional interpretation tool. However, this combinatorial task can be challenging to perform and interpret if several blocks are available.

One of the main advantages of this method is that it is invariant to the relative weighting of the blocks, which is an interesting feature if the data blocks have widely different measurement units. Another advantage is that it can handle situations with different underlying dimensionality (pseudo rank) of the blocks. This makes it suitable for situations where one of the blocks is, for instance, a design matrix originated from an orthogonal design of experiments (independent variables) while the other ones are highly multivariate and collinear (e.g. spectroscopic data) (Jørgensen et al., 2007). A third advantage regards its ability to convey information about the additional variability of the successive input blocks and the incremental prediction capability they bring to the model.

Table 2-5 - Pseudocode for SO-PLS

```
I – Select order of the blocks to perform SO-PLS
II – Perform PLS between first predictor block and Y
              The Y score u is initiated as a column of Y.
(1)           $w_b = X_b^T u / u^T u$                       % Xb block variable weights ($w_b$)
              normalize $w_b$ to $\|w_b\| = 1$
(2)           $t_b = X_b w_b$                               % Xb block variable scores ($t_b$)
(3)           $q = Y^T t_b / t_b^T t_b$                     % Y weight ($q$)
(4)           $u = Y q / q^T q$                             % Y score ($u$)
loop until convergence of $t_b$. Steps (1)-(4) are repeated for each latent variable.

III – Perform Orthogonalization of all successive blocks (for all Xi , i>b) using previous ones
(5,11,13)     $p_i^{orth} = X_i^T t_b / t_b^T t_b$          % Xi loadings orthogonalized
(6,12,14)     $X_i^{orth} = X_i - t_b p_i^{Torth}$          % $X_i^{orth}$ orthogonalized
IV – Repeat PLS steps (1)-(4) between the successive orthogonalized predictor block $X_i^{orth}$ and Y, until convergence of $t_i$.
```

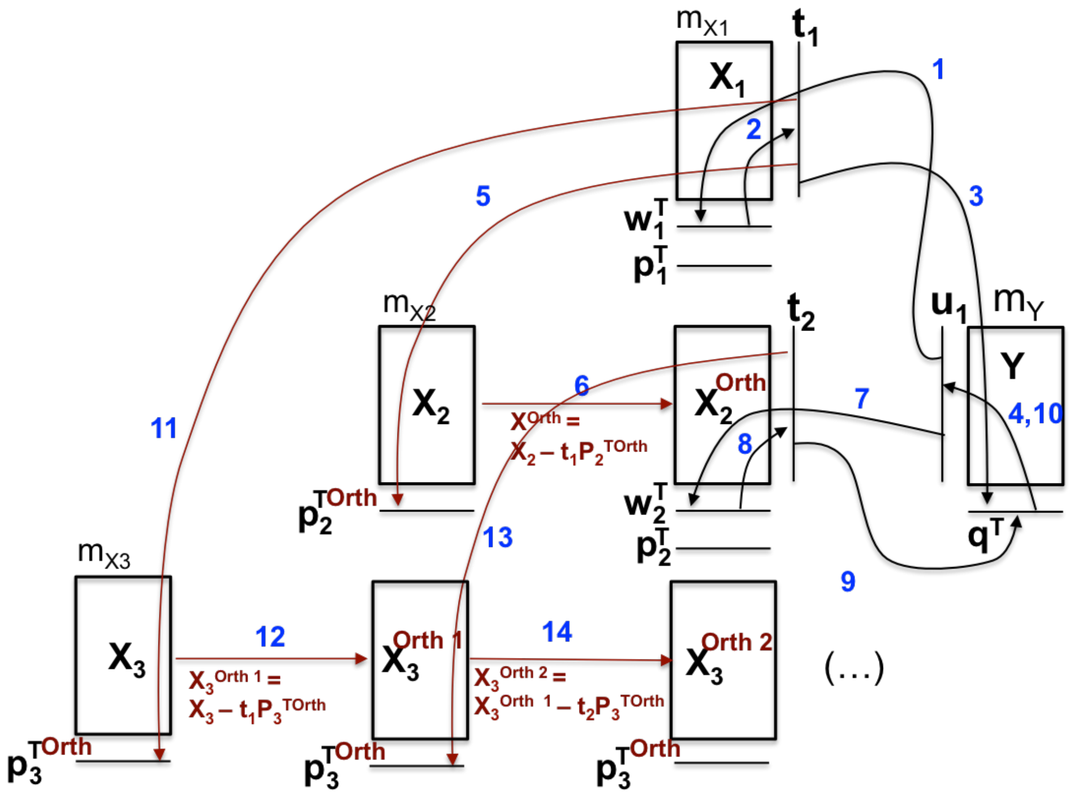


Figure 2-4- Schematic representation of the SO-PLS algorithm. The numbers in each arrow represent a step in the pseudocode in Table 2-5.

2.4 Monte Carlo framework for assessing model prediction performance

In order to critically assess the relative prediction performances of the methods under investigation, suitable performance metrics and robust and accurate comparison frameworks need to be defined and implemented. Therefore, a statistical framework consisting of a Monte Carlo double cross-validation scheme was adopted for this research, which is schematically represented in Figure 2-5. In this framework the data set is randomly divided into a training set (80%) and a test set (20%) in each Monte Carlo iteration (outer circle). The training set is used to calibrate the model and to determine the respective hyper-parameters based on 10-fold cross-validation method (inner circle). K-fold cross-validation (K-F CV) (Breiman and Spector, 1992, Hastie et al., 2009, James et al., 2013, Kohavi, 1995) is a cross-validation strategy that consists of randomly splitting the entire data set into k groups, or folds, of approximately equal size. This strategy is less computationally intensive when compared to other common cross-validation approaches like leave-one-out (LOOCV), because the fitting step is conducted less times (Breiman and Spector, 1992, Hastie et al., 2009, James et al., 2013, Kohavi, 1995). The test set is then used for prediction and to compute the root mean square error of prediction (RMSEP), for each Monte Carlo run and each method, using equation (3). A total of fifty Monte Carlo iterations were carried out originating fifty models for each method tested. The lower the *RMSEP* value, the better the prediction performance of the corresponding method.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{pred,i} - y_{obs,i}\right)^2}{n}} \tag{3}$$

Where, $y_{obs,i}$ is the i th observed response, $y_{pred,i}$ is the corresponding estimate, and n stands for the number of observations in the testing set.

The distribution of the fifty RMSEPs obtained characterizes the method performance in terms of prediction accuracy and consistency. Moreover, it is also used to compare different methods in a more rigorous and robust way, using a statistical framework based on formal hypothesis testing. Since the same data sets were used as training and test sets for each method, a paired t-test can be adopted to compare the prediction capability of all pairs of methods under investigation in a pairwise manner, as follows: the null hypothesis is that the mean difference between prediction errors of the two methods under comparison is zero; the null hypothesis is rejected when the p-valued obtained is less than 0.05. To facilitate the analysis of all the outcomes of the paired t-tests, a key performance indicator (KPI) is computed that summarizes the overall performance of each method: a score of 2 is given to the method with statistically significant inferior RMSEP; a score of 1 is given to both methods in case no significant difference in RMSEP is obtained and a score of 0 is given to the method with statistically significant superior RMSEP. The scores are summed up for each method for all pairwise comparisons and a final KPI is obtained characterizing the method´s overall relative prediction performance. Methods with higher KPI values are performing better in terms of relative prediction capabilities.

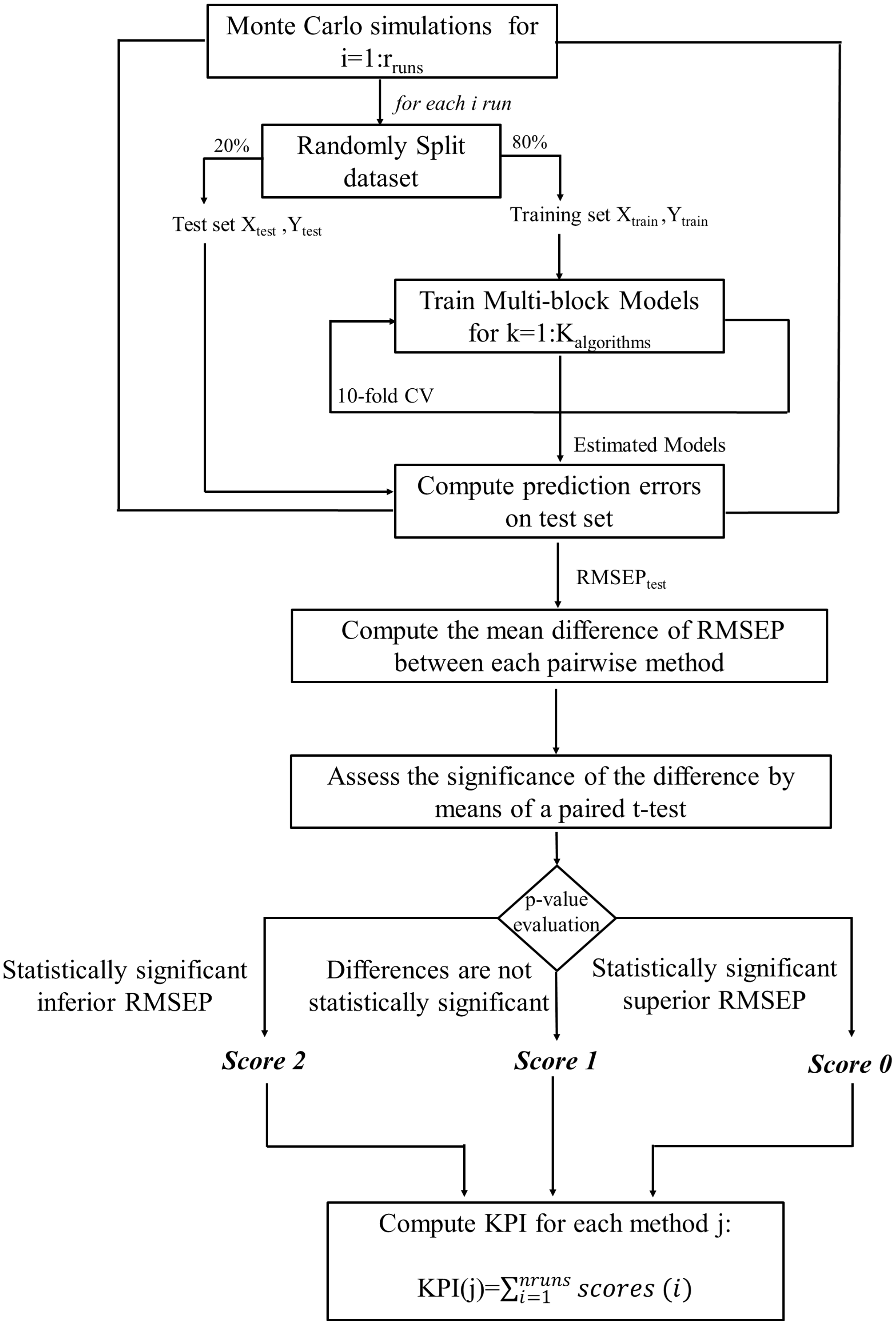

Monte Carlo simulations for i=1:r$_{runs}$
for each i run
Randomly Split dataset
20%
80%
Test set X$_{test}$,Y$_{test}$
Training set X$_{train}$,Y$_{train}$
Train Multi-block Models for k=1:K$_{algorithms}$
10-fold CV
Estimated Models
Compute prediction errors on test set
RMSEP$_{test}$
Compute the mean difference of RMSEP between each pairwise method
Assess the significance of the difference by means of a paired t-test
p-value evaluation
Statistically significant inferior RMSEP
Differences are not statistically significant
Statistically significant superior RMSEP
Score 2
Score 1
Score 0
Compute KPI for each method j:
KPI(j)=$\sum_{i=1}^{nruns} scores\ (i)$

Figure 2-5 - Statistical framework for the comparison of multiblock methods in terms of their prediction capabilities

Chapter 3 – A state-of-the-art review on multiblock modelling

Chapter 3. A state-of-the-art review on multiblock modelling

This chapter provides a critical overview of multiblock modelling and its application to a variety of problems faced nowadays in modern industry and analytical laboratories.

3.1 Multiblock Methods – Overview

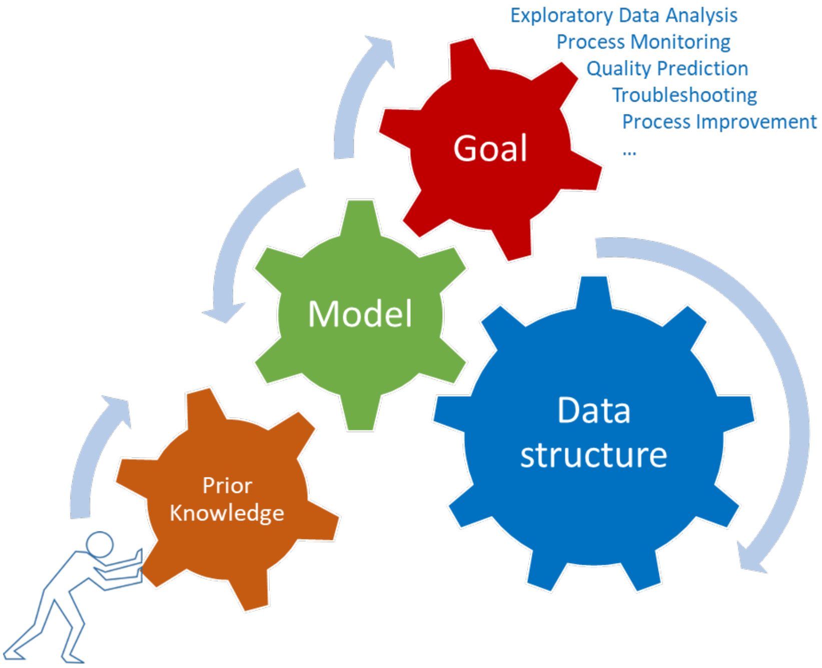


Figure 3-1 - Incorporating prior knowledge into data driven modelling to achieve the desired goal when data with a given structure is available.

We are witnessing the development of Industry 4.0 and the emergence of the Big Data era which is often characterized by the so called 5 V's: Volume, Velocity, Variety, Variability and Value. Organizations can collect large amounts of data (Volume) of different types (Variety) at high rates (Velocity), with varying levels of uncertainty and quality (Veracity), that have the potential to improve processes (Value). The goal is to achieve value by transforming data into knowledge by means of effective data driven methodologies.

Incorporating knowledge already known *a priori* into the data analysis can make the difference in the quality of the results – see Figure 3-1. This aspect reduces the number of degrees of freedom available to explore all the data available (possibly lowering the fitting ability) but potentiates model interpretability and robustness (with benefits in the prediction ability on future samples). One possibility to incorporate *a priori* knowledge into the modelling is the identification and integration in the model of natural blocks of data that comprise local information about parts of the process or specific phenomena. These data blocks (non-overlapping sets of variables) convey meaningful complementary pieces of information about the system as they are potentially actively involved in a variety of system's functions and sometimes cooperate and act together in certain phenomena. Therefore, it would be highly desirable that they maintain their integrity during data analysis (Reis, 2013b). Examples of multiblock data are abundant and can be found in a wide variety of areas: chemical engineering, analytical science, medical science, systems biology, food, and pharmaceutical, among other disciplines.

Casting all the different sources of data into an immense "data lake" and model them altogether is the most straightforward way to fuse data. Classical (single block) multivariate data driven methods such as PCA or PLS for instance can then be applied to the augmented matrix. The advantage of this approach is that it can easily handle any number of blocks. The drawback, however, is that the individual block contributions and their relationship to each other are hard to interpret. Another possibility, also limited from the standpoint of interpretation, is the isolated analysis of each single block by means of PCA or PLS. In this approach, interactions between blocks are impossible to establish and it is also more time consuming when many data blocks are available (as equal number of models need to be developed).

Multiblock methods are systematic approaches that take blocks of variables as their natural inputs, much like classical data analysis methods take isolated/individual variables as the entities whose effects they aim at unravelling. The aim of these methodologies is to preserve the variables' composition of each block and keep track of the information extracted from the different blocks. Modelling the data in this structured way enables inferences about the inter-relationships and overlapping information among the different data blocks and their relative contribution to the outcomes, thus significantly improving the interpretability of the model (Frank and Kowalski, 1985, Gerlach et al., 1979) and also, at the same time, potentially improving its prediction performance (Campos et al., 2017). They also allow for the identification of the most relevant blocks and to perform a "zoom in" of the relevant blocks, streamlining the analysis while enhancing model interpretation (Skov et al., 2008, Kourti and MacGregor, 1995).

Another major benefit of the application of multiblock methods is assessing the commonalities and differences between the different data sets and retrieving quantitative measures on the amount of distinct and redundant (common) information among the multiple blocks. Common variation (also called joint or overlapping) refers to underlying phenomena that are captured by several of the data blocks, while the distinct variation (also called individual or unique) correspond to phenomena that are only found in one block (Smilde et al., 2017). The common variation can be used to explain associations between different data sets such as same pattern of variability, while the distinct variation can provide a better understanding of specific sources of variability in each data set and their impact. These measures can be very advantageous when several blocks are available to see, e.g., if one or more blocks are sufficient or all are needed to describe the responses. The complexity increases with the increasing number of available data blocks, because the common variation can be either global (across all blocks) or local (across subsets of blocks) and the possible combination of local common variation increases exponentially with the number of available predictor blocks. These aspects are illustrated in Figure 3-2, where the three circles represent three data blocks, D is the distinct information and C is the common information between the blocks.

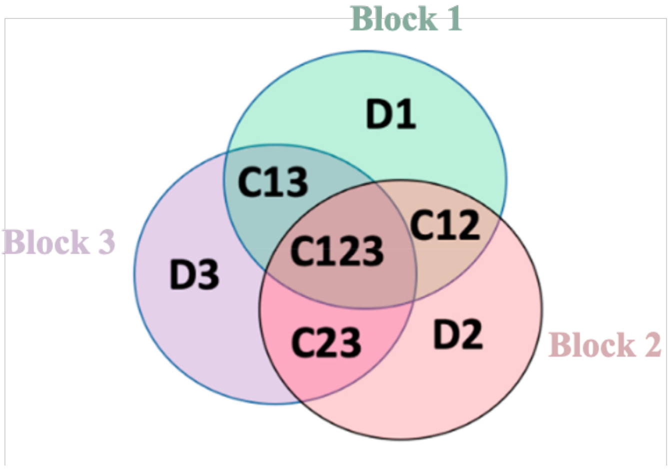


Figure 3-2 - Framework of common and distinct information extraction from multiblock data. Each circle represents one data source. Inside each circle, D is the distinct information and C s the common information. This figure is adapted from (Mishra et al., 2021).

All these benefits justify why multiblock models are becoming increasingly popular in the Industry 4.0 and Big data era. A summary of the benefits of using multiblock modelling approaches is summarized in Table 3-1.

Table 3-1- Summary of the benefits of adopting multiblock modelling approaches

Benefits of Multiblock Modelling
✓ Capable of handling multiple blocks of data (complex and heterogenous data structures) ✓ Incorporate additional information in the data modelling, thus reducing degrees of freedom during model fitting ✓ Provide information regarding the inter-relationships among the different data blocks and the relative contribution of each block into the model ✓ Keep track of the information in each data block enabling the visualization of the model results for each data block separately (“zoom in” feature) ✓ Improve model interpretability ✓ Capable of providing quantitative estimations of the common and unique information from each data structure

When discussing about multiblock modelling methods, a central notion is the concept of coupled data, which can be defined as a connected collection of data blocks with the connections between blocks consisting of shared modes such as the sampling/objects/observations/rows mode and the variables/columns mode (Van Mechelen and Smilde, 2010). Without any coupling of data, modelling the different blocks together is not meaningful or even possible. The first case of coupling pertains to data coupled in the sampling mode, i.e., same set of samples analysed by different techniques (e.g., NIR, HPLC) – see Figure 3-3 (A). Another possibility is coupling along the variable mode where the same variables are measured for different sets of samples (e.g., process sensors data collected at two different unit operations) – see Figure 3-3 (B). In chemistry, this class of problems is encountered e.g., when analysing different chemical systems with the same spectroscopic methods. Multivariate curve resolution (MCR) has been developed to analyse such multiblock data sets (Tauler et al., 1993).

Monitoring of batch processes can also be done in this way (Wold et al., 1998b, Rendall et al., 2019). Moreover, hybrid cases are also possible; see Figure 3-3 (C). A special case is when data sets share both modes. Structures sharing both modes are called multiway structures, and specific methods have been proposed for addressing multiblock and multiway analysis (Acar et al., 2014, Bro, 1997, Kroonenberg and De Leeuw, 1980, Smilde et al., 2000). The present thesis will focus on the class of multiblock problems falling under the scope of case (A) of Figure 3-3, where the blocks, formed by different groups of variables, are coupled in the sampling mode.

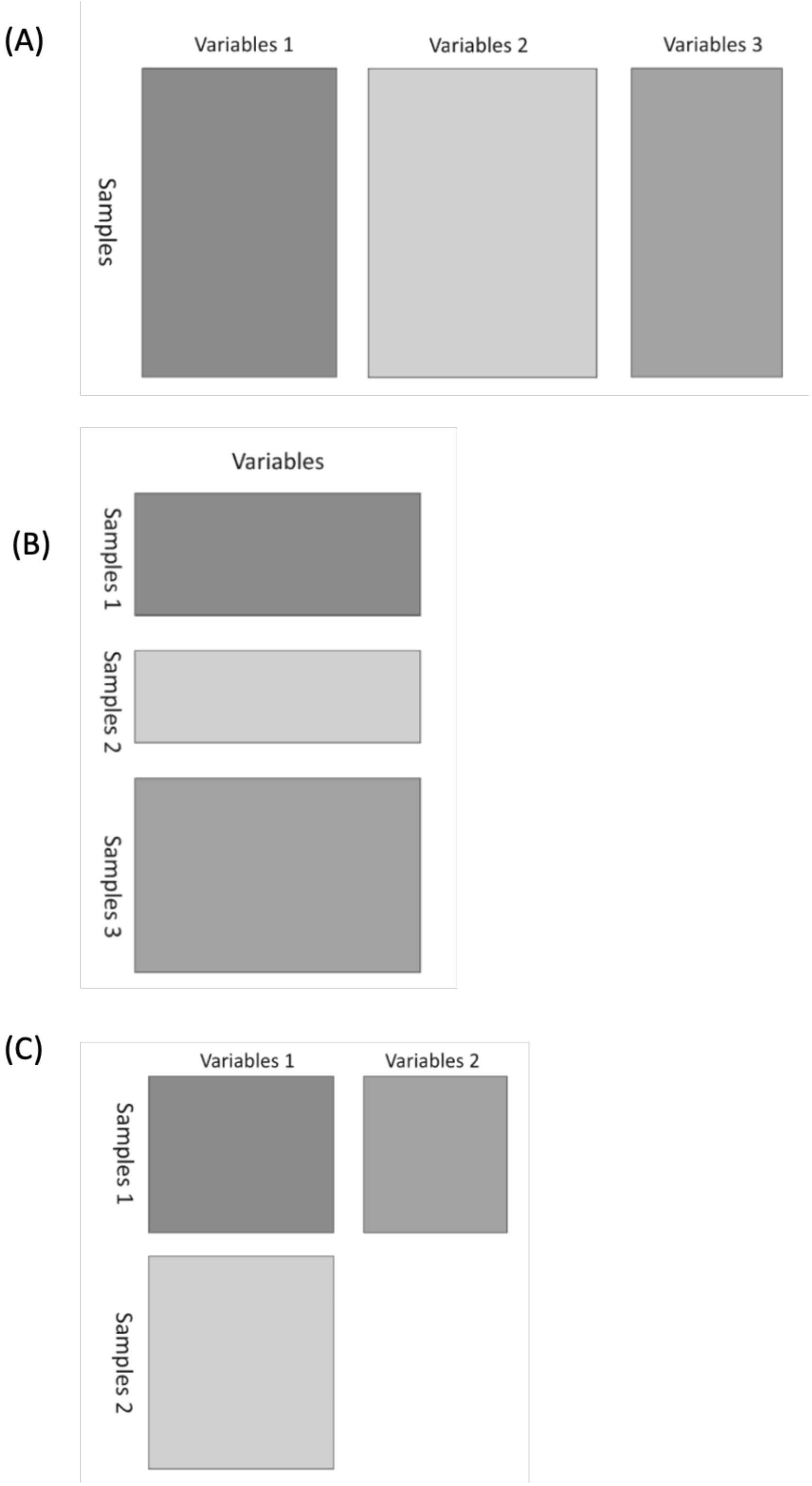


Figure 3-3 - Examples of different data block's couplings. (A) data blocks coupling along the sampling mode; (B) data blocks coupling along the variable mode; (C) hybrid case - coupled along the variables and samples modes.

Despite the clear potential benefits and the significant work that has already been done in the field of multiblock modelling approaches, their usage is still quite underexplored in industrial data science and there are several challenges that need to be addressed in order to become part of the mainstream of data analysis methods.

To start with, a unified and systematic workflow for the application of multiblock methods is not available. This is critical for practitioners, engineers, and data analysts, that are not experts on multiblock analysis, nor do they aim at being so, but are still interested in making good use of their data. Furthermore, a robust and easy to use and interpret framework for assessing the existing methodologies performance is also lacking, further preventing a wider understanding and application of these methodologies. Multiblock approaches are very much dependent on the characteristics of the data blocks under analysis. The data sets for analysis can be very complex and present different degrees of sparsity, collinearity and linearity which can raise relevant challenges on the way the different information sources should be fused and synergistically explored in order to improve the desired goal of the analysis.

Other challenges that arise when handling multiblock modelling approaches include how to properly pre-process the different blocks (e.g., a systematic approach is lacking) and how to perform efficient block selection (e.g., exclude non-relevant blocks to the analysis to develop more parsimonious and easier to interpret models).

Furthermore, hitherto more emphasis has been given to the prediction capabilities of the data driven methods. However, the enhanced interpretation capabilities provided by the multiblock models, which is one of the major benefits for applying these methods, has been relegated to a secondary concern. This aspect needs to be further explored in order to use these methods at their full potential.

In the multiblock data analysis field, methodologies have been developed covering both the unsupervised and the supervised contexts. This terminology has the same meaning as in classical single block methodologies. Unsupervised refers to exploratory analysis looking for structure and connections across data blocks, typically using dimension reduction and visualization methods. In this context, the blocks are exchangeable, meaning that the order of the blocks can be changed without changing the solution or the nature of the analysis. Supervised refers to predictive data analysis, where emphasis is on explaining the variation in the block **Y** (response block) which is connected to one or more blocks of data, **X** (predictors) through regression. In this case the role of the blocks is relevant: some blocks are regarded as inputs (**X** blocks), and others as outputs (**Y** block).

3.1.1 Unsupervised multiblock exploratory data analysis

Since multiblock data analysis has been the focus of several research works over the last three decades or so, several methods dedicated to exploring the structure of data blocks and investigating their underlying relationships have been proposed and compared in the literature. Unsupervised multiblock methods are also called multiblock component models and have in common the fact that they extract global components (or latent variables) that highlight the main dimensions underlying the data and the associated blocks (or local) components – this relationship is schematically depicted in Figure 3-4.

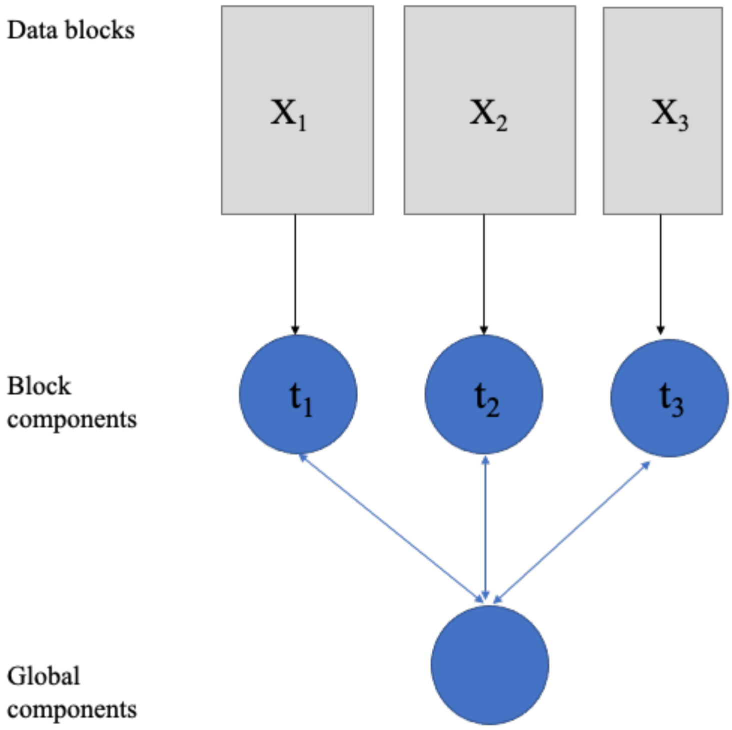


Figure 3-4 - Multiblock Component Problem. Relationship between global and block (local) components.

The consensus PCA (CPCA) was introduced by Wold et al. (1987b) at the Frankfurt PLS conference as a method for comparing several blocks of descriptor variables measured on the same objects. A consensus direction (super score) among all blocks is estimated and the individual blocks are then regressed on the super score to extract the weights for the individual blocks to have an insight into the contribution of each block to the super score. In 1996, Wold et al. (1996) presented a slightly different multiblock PCA method, called hierarchical PCA. The difference between CPCA and HPCA lies on the different normalizations used: in CPCA, super weights are normalized to unit length while in HPCA the normalization is on the super scores. Due to some convergence problems, some adaptive versions of these methods have been proposed with different normalizations (Rännar et al., 1998, Westerhuis et al., 1998).

Other similar methods have also been published in the literature, such as common components and specific weights analysis (CCSWA or ComDim) (Cariou et al., 2019, Hanafi et al., 2006, Qannari et al., 2000), that was shown to be equivalent to HPCA (Hanafi et al., 2010) but is more sophisticated in terms of the mathematical formulation, with several possibilities of expansion, namely to predictive applications (El Ghaziri et al., 2016) and path modelling (Cariou et al., 2018). The method consists of determining a common space for all the **X** data blocks, with each data block having a specific contribution to the determination of each dimension of this common space. This is achieved by extracting, in a sequential way, global and block components that recover the maximum of the total variance in all data blocks.

Methods such as CPCA, HPCA and ComDim are dedicated to the identification of the common information among the different data blocks and the relative contribution of each block to the common components. On the other hand, several methodologies have been developed for the simultaneous extraction of both unique and common information in the data blocks. Examples of these methods are distinct and common simultaneous component analysis, DISCO-SCA, or DISCO for short (Schouteden et al., 2013, Schouteden et al., 2014, Van Deun et al., 2012); joint and individual variances explained, JIVE (Hellton and Thoresen, 2016, Kuligowski et al., 2015, Lock et al., 2013); orthogonal n-block PLS, OnPLS, (Trygg, 2002); principal component analysis-generalized canonical correlation analysis, PCA-GCA (Smilde et al., 2017) and structure learning and integrative decomposition, SLIDE (Gaynanova and Li, 2019). These methods use different approaches for selecting the number of common and distinct components and also differ in the orthogonality properties of the scores (Måge et al., 2019, van der Kloet et al., 2016).

DISCO starts with the application of simultaneous component analysis (SCA) on the concatenated **X** matrix to determine the joint sub-space. In the next step, block loadings matrices are partitioned and orthogonally rotated to identify the common and distinct components. The rotation is orthogonal, meaning that all (both common and distinct) score vectors are orthogonal to each other. While the orthogonality definitely has some advantages regarding the mathematical treatment and the analysis of results, there are few reasons to justify the assumption that all distinct phenomena are orthogonal in real life. These constraints may become too strict and lead to a suboptimal representation of the common and distinct subspaces.

JIVE is also based on a SCA of the concatenated data sets (**X**) to estimate the common components. Then, the distinct components for each block are found by applying Singular Value Decomposition (SVD) on what remains after deflating the common part. The original **X** is then updated by deflating the distinct components, and the procedure is repeated until convergence of the residuals. By using an iterative and alternate optimization of the common and distinctive parts, the orthogonality between the two distinctive parts that does exist in DISCO is no longer enforced. In contrast to the other methods, the JIVE method only facilitates the decomposition into the global common and distinct subspaces and does not allow for the analysis of local common subspaces.

Multiblock methods based on orthogonal projections have received interest within life-sciences provided the model structure it can decompose the data blocks into. O2PLS (orthogonal two block PLS) is a multiblock extension of orthogonal PLS (OPLS) with the relevant difference that no asymmetric relation among the blocks is implied, so that the method can be used also for exploratory purposes (Trygg, 2002). O2PLS was later generalized into OnPLS (Löfstedt, 2012).

PCA-GCA, as the name suggests, is a combination of PCA and generalized canonical correlation analysis (GCA). The PCA-GCA method starts by decomposing each block individually by PCA, keeping a relevant number of scores from each block. Then, GCA is used to find canonical variates between these sets of PCA scores. The canonical variates with sufficiently high correlation coefficients define the common components. The common components are then removed from the original blocks by orthogonalization, and the distinct components are found by applying SVD on the remaining. A major difference between PCA-GCA and the other methods is that it operates on the individual data blocks, not on the concatenated data. This means that the common components are in the column spaces of each block, not of the concatenated **X**. Because of this, the method is invariant to between-block scaling,

Finally, the SLIDE method allows components to be partially shared (i.e., common only to some blocks). This is achieved by arranging the loadings in a block-dependent structure and imposing structure sparsity to reveal the common, distinct and the partially shared information.

3.1.2 Supervised multiblock regression

The multiblock regression problem is schematically represented in Figure 3-5 as an example in which three blocks $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are used to predict **Y**.

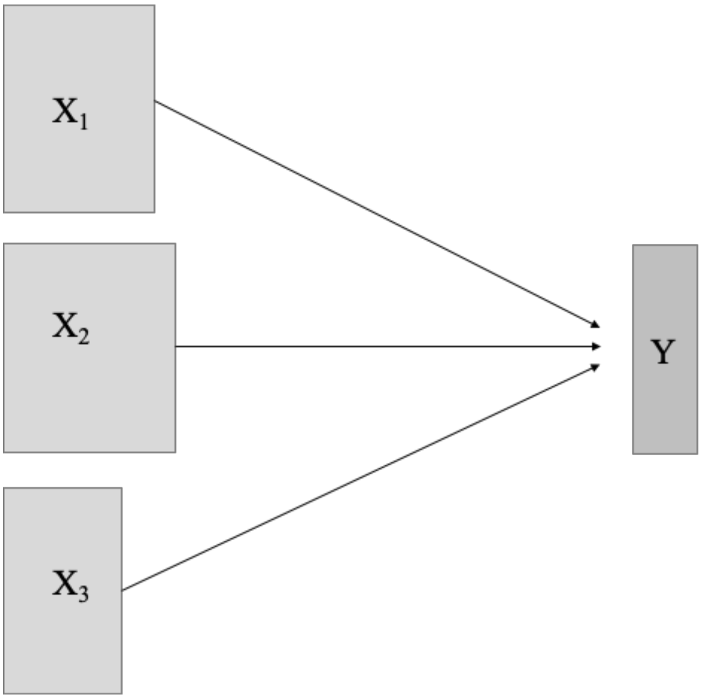


Figure 3-5 - Example of a multiblock regression model with three predictor blocks $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ and one response block **Y**.

Analysing the multiblock frameworks developed and published in the technical literature, as well as subsequent refinements and improvements over existing methods, it is possible to verify that the class of multiblock latent variable methods is currently dominating the application landscape, given their ability to accommodate large amounts of potentially collinear data. In chemometrics, most of the methods for multiblock predictive analysis are extensions of standard PLS regression to the multiblock scenario.

Multiblock modelling has its roots in path modelling in the fields of sociology and econometrics and the usage of multivariate projection methods to path modelling are largely due to the work of Wold (1982), later on adapted to chemistry-related applications by Wold (1987). The first predictive multiblock method application was presented by Frank and co-workers for the purpose of predicting the quality of adhesive tapes and wine quality (Frank et al., 1984, Frank and Kowalski, 1984). Moreover, Frank and Kowalski (1985) proposed two different algorithms for the multiblock method: the averaging method and the stepwise method. In 1987, during the Frankfurt PLS conference, Wold et al. (1987b) suggested a refined multiblock method called Hierarchical two-block predictive PLS or PLS-2H. After this paper, several variations of this hierarchical PLS (HPLS) method have been published in the literature, all leading to identical results (Slama, 1991, Wold et al., 1996, Wold et al., 1987b). For instance, Slama (1991) suggested a HPLS algorithm in which a normalization of the super scores $\mathbf{t}_T$ is performed instead of the normalization of the super weight $\mathbf{w}_T$, as done previously by Wold et al. (1987a). A different multiblock algorithm, called multiblock PLS (MBPLS), was described by Wold et al. (1983) and later by Wangen and Kowalski (1989). Both HPLS and MPLS model the common structure of the different predictor blocks in an upper level (called super level), while information from the separate blocks can be found in a lower modelling level (called block level). On each level, "standard" PLS or principal components (PC) scores and loading plots are available for model interpretation. This allows for an interpretation focused on pertinent blocks and on their dominant variables. The main differences between these two methods regard the way block scores are determined and how normalization is performed. In MBPLS there are two different approaches for the calculation of the residuals and loadings: the block scores deflation method (Frank and Kowalski, 1985) and the super score deflation method (Westerhuis and Coenegracht, 1997) – for more details on these methods please refer to Chapter 2.3. A similar method called P-ComDim (abbreviation of Predictive Common Dimensions), or P-ComDim (k+1) (El Ghaziri et al., 2016), extracts global scores that capture maximum covariance with the response variable(s). This is done by maximizing the covariances between the local scores of each block and the scores of the response block. The advantage of this method is that it is able to provide information on the specific block weights, which highlight the importance of the various blocks to the determination of the latent variables.

Serial PLS (SPLS) can be interpreted as an alternative multiblock PLS algorithm where the predictor blocks are modelled serially, i.e., the block models are calculated using the **Y** residuals from the previous block model (Berglund and Wold, 1999, Felício et al., 2005). Since the blocks are treated separately in SPLS, it is possible to determine if additional blocks have any significant modelling power. However, this algorithm is rather slow and no suggestion has been put forward to deal with more than two **X**-blocks.

The majority of the available state-of-the-art multiblock methods assume that the grouping of the data into conceptually meaningful blocks is known *a priori* due to some underlaying rational such as respective location in the process stream, analytical similarity, etc. However, there are other methods that are also able to identify natural clusters of highly connected variables. Network Induced Supervised Learning Approach, NI-SL, (Reis, 2013b) is a supervised framework aiming at bringing interpretation features to the forefront of the analysis goals and using this approach the blocks can be identified by means of a clustering algorithm that identifies groups of directly related variables, possibly constituting functional modules. Reis (2013b) demonstrated that it is indeed possible to obtain more interpretable models by modelling and keeping the integrity of the data blocks without compromising prediction ability (which in fact is often improved due to higher robustness of the more parsimonious models derived).

Some chemometric techniques have been put forward to provide a clear extraction of the common and distinct information from the distinct data blocks. Orthogonal n-PLS (OnPLS) was proposed by Löfstedt (2012) which is an extension of the two-block O2PLS (Trygg, 2002) to the multiblock scenario. As in O2PLS, the OnPLS does not introduce *a priori* any asymmetry between the blocks, however, if one block contains the response(s) to be predicted, by suitably combining the scores extracted from all the other matrices, a global regression model can be calculated. This model separates the variation in three parts: the globally joint variation found in all blocks, the locally joint variation found in at least two but not all blocks, and the unique variation found in only a single block. The different covariance matrices are concatenated and analysed similar to an SCA approach. In this way, a global common direction is estimated. Anything orthogonal to this direction is either distinct or partially common and is determined by repetitions of the OPLS procedure on these orthogonal parts. After the global and local variation is determined, they are removed from the original data and the global common variation is (re)calculated.

Sequential Orthogonalized Partial Least Squares (SO-PLS) (Jørgensen et al., 2004, Næs et al., 2011b, Jørgensen et al., 2007) approach involves a series of standard PLS regressions and matrix orthogonalization operations, to extract sequentially the complementary information from different data blocks. The major advantages of SO-PLS are linked to the orthogonalization, which removes redundant information, and to its sequential nature, which allows the interpretation of the incremental contributions provided by each data block. The SO-PLS approach is particularly advantageous when the aim is to identify possible additional benefits from the inclusion of more blocks of information into the model.

Parallel Orthogonalized Partial Least Squares (PO-PLS) approach (Måge et al., 2008, Måge et al., 2012), on the other hand, involves a combination of PLS regression, generalized canonical correlation analysis (GCA) and multiple orthogonalization steps. PO-PLS, unlike SO-PLS, does not explore the blocks sequentially, but aims at identifying the common and the distinct information in different blocks, to have a better understanding of how the combinations of blocks contribute to the improved predictive performances.

Response Oriented Sequential Alternation (ROSA) (Liland et al., 2016) is a multiblock method capable of handling a large number of blocks in an efficient way, being also invariant to block scaling and ordering. ROSA is also computationally faster than the two previous methods, as it does not require any deflation step to calculate orthogonal scores and loading weights.

Multiblock variance partitioning (MVP), originally proposed by Skov et al. (2008), presents some similarities with both SO-PLS and PO-PLS and it was one of the first methods to specifically focus on the separation of the common and unique variation in the prediction blocks. The total **Y**-related variation for each block is partitioned into a unique part (imputable to only that particular prediction block), a common part (also shared with the other predictor blocks) and an uninformative part (which is the percentage of information of **Y** that **X** cannot explain).

It works by establishing local PLS models between predictor blocks and a common response block. For each predictor block, the uninformative variance is associated to the residuals of that regression while the unique contribution is calculated after orthogonalizing the predicted responses (variable-wise) with respect to the corresponding predicted responses based on all the other predictor blocks. The common variation is obtained by subtracting the contribution of the unique and uninformative parts from the total variance. MVP works on individual predictor blocks and thus no scaling issues are raised.

3.3 Applications of multiblock regression

Applications of multiblock methods have been reported in the literature covering a wide range of problems in industrial processes and analytical chemistry. This section contains some examples of such applications to provide a more tangible picture of the types of questions addressed, data sets analysed and potential benefits of using multiblock methods. The examples are organized according to the application scope.

Process Analysis and Optimization

In process analysis and optimization, more than estimation accuracy, the focus is on extracting information from processes and the structure of the system that could lead to potential roots for improvement. This implies studying the relations between variables pertaining to different blocks and between blocks themselves. This can be useful for several purposes: understanding how variables relate through blocks (e.g., operation units) and understanding which variables matter more in determining the final product quality and how disturbances in materials inputs or process parameters propagate to the final product for instance.

Several practical examples can be found in the pharmaceutical industry where it is critical to understand the driving forces acting upon the complex network of interactions between materials, processes, and final product. The amount and quality of the final product depends on all previous stages (typically the production process of an active product ingredient encompasses several batch and fed-batch process steps) and it is therefore crucial to consider all sources of variability present. In Lopes et al. (2002) the performance of an industrial pharmaceutical process (production of an active pharmaceutical ingredient by fermentation, API) is modelled to determine the contribution of the main production stages (inoculum growth and fermentation) to the API productivity. In Brás et al. (2004) the API's isolation stages were also encompassed.

Another example from the pharmaceutical industry regards the improvement of tablet final quality (J. a Westerhuis & Coenegracht 1997). It is well understood that the final tablet characteristics rely not only on the operation conditions of the tablet press, but also on the granule properties and on the tablet ingredients (formulation). To improve tablet quality one needs to understand the effect of formulation, how powders perform, how to handle properly powder characteristics and the environment, how to set the process variables, and how to satisfy the demands of a tablet press (Tousey 2002). Many papers have been published addressing the effect of operation conditions of different stages of the process on tablet properties (de Jong 1991, Debunne et al. 2004). However, significantly fewer papers have been published considering all the factors together (raw material characteristics, formulation and process conditions) in a general model to predict and control tablet properties (J. a Westerhuis & Coenegracht 1997).

A typical problem that industry faces is how to optimize process parameters (block **X**) to achieve the desired product quality (block **Y**) in the presence of raw material variability (block **Z**). This is a complex optimization problem, for which multiblock methods may provide insights and tools. In this context, one block of variables can be originated from designed experiments (process variables) and a second block consists of highly collinear data (spectral data). This is for instance the case when raw material properties of a product are measured by multivariate spectroscopy, samples are processed according to an orthogonal experimental design and the interest lies in the final product properties. Several multiblock methods have been proposed in the literature to address this problem (Jorgensen et al. 2005; Måge & Næs 2005; Henriksen et al. 2005).

In the optimization of chemical processes, engineers need to identify the effect that process variables have on the final product quality through the different stages of the process. For example Duchesne and MacGregor (2000) used a multiblock multiway PLS method in a styrene-butadiene rubber emulsion copolymerization for obtaining the sensitivities of final product quality to changes in the shape of process variable trajectories and subsequently using those for improving final quality.

Quality Prediction

Product quality prediction is a very important problem in industrial processes. It is usually required for quality monitoring, as well as to set up a strategy for improving product quality.

Gaydou et al. (2011) investigates the potential of using simultaneously near infrared (NIR) and mid infrared (MIR) spectroscopy for the quantification of vegetable oil in diesel/biodiesel blends. Three different approaches are tested: a) concatenating the descriptor block into a single matrix and then applying Partial Least Squares (PLS), b) Hierarchical PLS (HPLS) and c) Serial PLS (SPLS). Felício et al. (2005) developed models to use MIR and NIR simultaneously for gasoline quality control.

Another example is the modelling of contaminants in a river when different sources of contaminants are considered as different blocks of measurements (Frank & Kowalski 1985). This was one of the first applications of multiblock modelling approaches.

Some examples can also be found in the pharmaceutical industry, in which quality control (QC) is a key activity in ensuring that medicines have the required quality, safety and efficacy for their intended use. QC departments at pharmaceutical companies are responsible for all release testing of final products, but also all incoming raw materials. Hertrampf et al. (2016) present an application of a multiblock approach for discriminating between two different active pharmaceutical ingredients (API) dosages (identity) as well as to predict their dosage (semi-quantitative) based on NIR and Raman spectroscopies. Brás et al. (2005) used multiblock approaches (MBPLS and SPLS) to combine NIR and MIR spectroscopies in fermentation media quality assessment. The application of a multiblock approach to this case provides extra interpretation features enabling to zoom in into separate blocks and analyse which block causes certain events in the response data. For instance Sarraguça et al. (2011) showed that using multiblock approaches to predict particle size distribution (PSD) based on NIR spectra, flowability properties, and the concentrations of the components present in the samples, lead to a deeper understanding of the relation between the data blocks and the PSD.

Process Monitoring and Control

Process monitoring of complex industrial processes has attracted much attention from engineers and researchers in recent years (Reis, 2019, Reis and Gao, 2021, Reis and Gins, 2017, Reis et al., 2019, Reis et al., 2021, Reis and Saraiva, 2019, Joe Qin, 2003, Li, 2014, Yin et al., 2014). Online monitoring of process operating performance and the rapid diagnosis of faults that occur are extremely important to ensure the consistency of final product quality and to improve process efficiency, safety, economy and environmental fingerprint.

Multivariate statistical process control (MSPC) has traditionally focused on the monitoring of single process units. However, when the number of variables is very large the univariate approaches face several limitations and become extremely difficult to implement and interpret (see e.g. Wise & Gallagher 1996; Kresta et al. 1991; Kosanovich et al. 1996; Nomikos & MacGregor 1995). Moreover, in practice, industrial operations comprise a sequence of processing units and equipment, and it is the combined operation of these units that determines the quality of the final product. Process variables can have different correlation structures for different blocks. For instance, variables within a section can be highly coupled but variables between sections are much less associated and contributing to different aspects of product quality. Furthermore, each section can encounter its own set of special events. Therefore, it is logical to break up the process into blocks and to develop monitoring and diagnosis charts for each section/equipment, as well as an overall chart. This means that, in addition to a monitoring space for the whole process; one also obtains several monitoring spaces for each process block. When a fault occurs in the process, this approach makes it much easier to detect, isolate and identify the causes for the fault (Smilde et al. 2000).

Diverse multiblock frameworks capable of modelling data in their natural structures and presenting enhanced interpretability features, have been proven useful to address the process monitoring and control problem in different fields (Wold et al. 1996; Chen & McAvoy 1998; Kourti et al. 1995).

Chen and McAvoy (1998) employed a multiblock modelling approach for the predictive on-line monitoring of a chemical continuous process (the Tennessee Eastman process). The process is divided into three major blocks according to their locations in the plant, namely the reactor block, the separator block, and the stripper block. The fact that only the separator block shows a deviation helps to focus on where the fault is occurring. Simulated results show that multiblock approaches are more appropriate for monitoring this large scale process in comparison with single block approaches and it was demonstrated to be more efficient in detecting process faults. The potential of applying multiblock modelling approaches in a low-density polyethylene (LDPE) process has also been illustrated (Kourti and MacGregor, 1995), where each block corresponds to one process zone. Scores plots and residuals obtained for each block of the process were utilized to detect an abnormal event in the zone it occurred; then contribution plots were successfully used to assign causes to it.

Qin et al. (2001) describes a case where data from a polyester film manufacturing process were divided into seven sections each describing a unit or specific physical or chemical operation. Then, multiblock algorithms were applied for decentralized process monitoring and diagnosis allowing to easily identify which operation unit contained the disturbance. Another example of applying a multiblock decentralized process monitoring scheme can be found in (Tong & Yan 2015). In this paper it was also discussed the situation where no process knowledge is available to group measured variables into conceptually meaningful blocks.

Kourti and MacGregor (1995) extended the monitoring problem of a batch process based on process variables using multiway principal components analysis (MPCA) to include final product quality and measured initial conditions. These initial conditions usually consist of feedstock properties, pre-processing and other conditions such as raw material properties and conditions, charges of each ingredient, holding times in charge tanks, and discrete operating conditions such as the operator shift on which batch is produced, raw material suppliers, etc.

Many batch processes have important multiphase and multistage characteristics that are not readily accommodated by multiway methods since process dynamics may differ considerably between different phases and stages (Yao & Gao 2009). In these multiphase and multistage processes there are different covariance structures for the different phases. The relationship between these different dynamics is not considered with MPCA/MPLS, where all the data is treated as a single object. This can severely compromise understanding of the process as well as monitoring thereof. Lee and Vanrolleghem (2003) describe the application of an adaptive multiblock MPCA to monitor a wastewater treatment that uses a sequencing batch reactor (SBR) process. The process has a cyclic nature, each cycle consisting of several phases — fill, anaerobic, aerobic, anoxic, and draw. Using a multiblock approach, the authors organized and modelled data in blocks corresponding to different batch phases.

Flores – Cerrillo and MacGregor (2004) proposed an extension of the multiblock MPCA/MPLS approach to explicitly incorporate batch-to-batch trajectory information summarized by the scores of previous batches while retaining the advantages and monitoring statistics of traditional MPCA/MPLS methods. The main advantage of this approach was that it could detect problems when monitoring new batches in the early stages of operation.

Ramaker et al. (2005) used a similar concept but based on adaptive Hierarchical PCA proposed by Rännar et al. (1998) to build local models for batch process monitoring. This method overcomes the need found in the approach of Nomikos and MacGregor (1994) for estimating or filling in the unknown part of the process variables trajectory deviations from the current time until the end of the batch.

Product design

The problem of process design has been also addressed using multiblock approaches. Interesting applications of L-shape PLS and T-shape PLS, can be found in García-Muñoz and Polizzi (2012), Polizzi and García-Muñoz (2011), Muteki et al. (2006). The letters reproduce the modes that are coupled in the different blocks of data. These data driven methodologies can reduce the development time of new industrial polymer blends while reducing costs, which is critical aspect for securing a competitive advantage for producers.

Analytical Chemistry

The portfolio of analytical chemistry technology found in research and industrial laboratories often leads to more than one type of measurements to characterize the samples under analysis. The information in the data from these different analytical instruments is frequently regarded as being independent - i.e., to describe different phenomena in the sample, thus adding supplementary information each time a new instrument is used. However, often these data blocks have redundant information and comparing the information/variation in different data blocks can be useful for instance to assess the techniques in terms of their relevance regarding the characterization of the quality of the product and to identify which blocks are sufficient to secure a thorough analysis of the samples

Chapter 4 – Case studies

Chapter 4. Case studies

The multiblock methods covered in this thesis will be applied to different real-world scenarios, which will be described in this section. The purpose is to contextualize the problems to facilitate the understanding of the applicability and appropriateness of the methodologies presented and discussed later on, in the following chapters.

4.1 Case study 1: Madeira wine ageing

The data set for the case study 1 consists of twenty-six wine samples covering an ageing range of 20 years, in intervals of 2 years. Two or three wine samples were considered per ageing year studied. All samples correspond to wines produced from the same grape variety, Malvasia, and were supplied from the same Madeira wine producer (Rendall et al., 2017, Pereira et al., 2016, Pereira et al., 2010a, Campo et al., 2006, Pereira et al., 2010b).

The chemical characterization of the samples includes the quantification of organic acids (1^{st} block), the volatile profile (2^{nd} block), the polyphenols and two furanic compounds (3^{rd} block) and the ultraviolet-visible spectra (4^{th} block). Seven organic acids were quantified by Liquid Chromatography combined with Photodiode Array Detection. The volatile profile was analysed by gas chromatography coupled to mass spectrometry (GC-MS), preceded by solid phase extraction, where 83 volatile compounds were identified in full scan mode. This data set includes esters, carbonyl compounds (ketones and aldehydes), higher alcohols and the fatty acids, typical of aged Madeira wines. The third data set was obtained by High-Performance Liquid Chromatography combined with Photodiode Array Detection (HPLC-DAD; direct injection). Overall, 23 phenolic compounds were identified (3 hydroxycinnamic acids, 5 hydroxybenzoic acids, 3 hydroxybenzldehydes, 3 flavan-3-ols, 2 flavonols, 1 stilbene and 6 unknown) and two furanic compounds. The acquisition of UV-Vis absorbance spectra was done in a Perkin-Elmer Lambda 2 spectrophotometer (Waltham, MA, USA), using 10mm path-length quartz cells at room temperature. Samples were filtered using 0.45µm PTFE syringe filters and scanned in the range of 245-785nm at 5 nm intervals. More details on sample preparation, chromatographic setup, acquisition protocol and quantification methodology can be found in Rendall et al. (2017), Pereira et al. (2016), Pereira et al. (2010a), Campo et al. (2006) and Pereira et al. (2010b).

The main goal is to predict the wine ageing time, from the analysis of these four blocks of chemical information. The motivation is not only to derive a model with good prediction ability, but, as importantly, to extract insights into the chemistry of wine ageing, and the evolution of the wine chemical profile over time. Figure 4-1 presents a schematic overview of the four predictor blocks available and respective sizes for the Madeira wine ageing time case study.

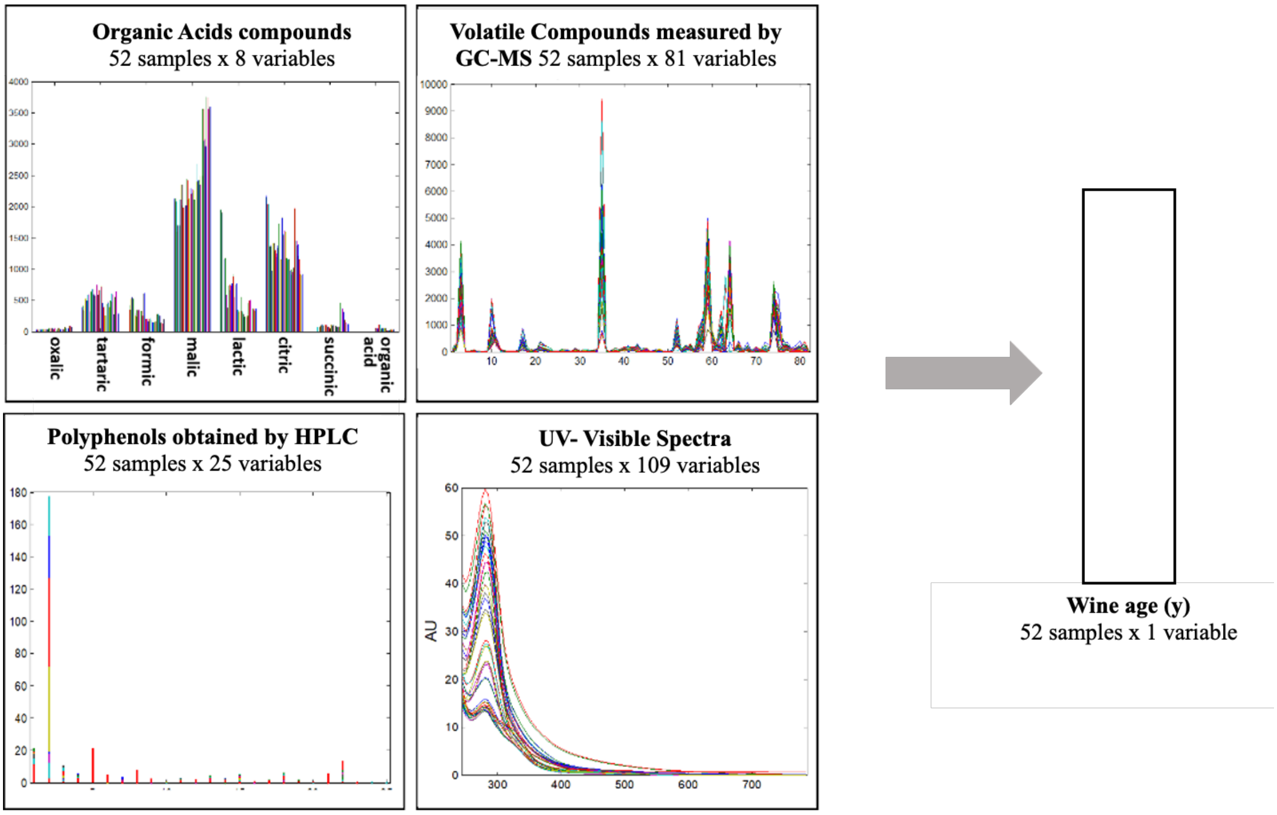


Figure 4-1 – Data blocks for case study 1: Madeira wine ageing.

4.2 Case study 2: Biodiesel production

The second case study used in this thesis concerns the prediction of the final quality of biodiesel, characterized by the percentage of Methyl Esters. Chemically, biodiesel is a mixture of fatty acid methyl esters (FAME), derived from vegetable oils or animal fats. Biodiesel is usually produced by a transesterification reaction, where the oils/fats react with an alcohol, in the presence of a catalyst – see Figure 4-2.

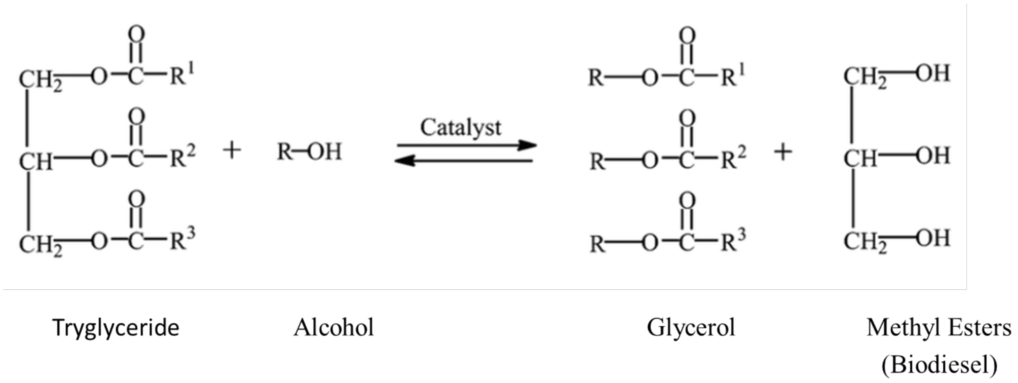


Figure 4-2 - The Reaction of Biodiesel: Transesterification

In this case study, a set of experiments were carried out using different molar ratios of methanol/oil (from 4.8 to 6.4) and different molar ratios of catalyst NaOCH3/oil (from 0.3% to 0.6% of the mass of oil) and the percentage of Methyl Esters was determined for each. The purpose of the case study is to derive a prediction model for the final quality of biodiesel (%Methyl Esters) based on the NIR spectra and two (2) operating parameters: the molar ratio of methanol/oil and the amount of catalyst used in each transesterification reaction. Twenty-one (21) soybean oil samples were used to produce the same amount of biodiesel. Figure 4-3 presents a schematic overview of the two predictor blocks available and respective sizes for the biodiesel production case study.

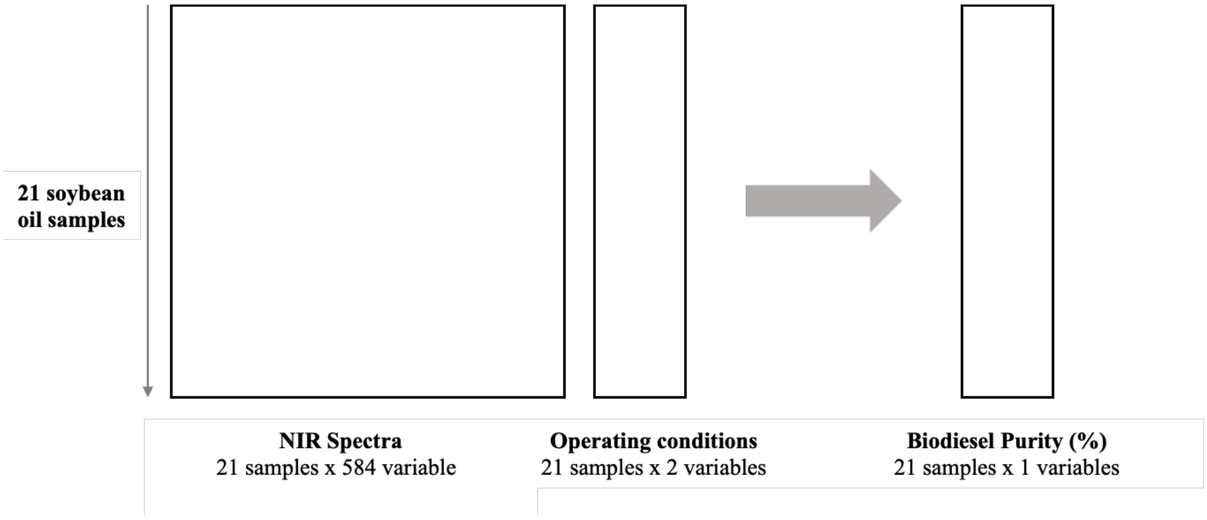


Figure 4-3 - Data blocks for case study 2: Biodiesel production.

Chapter 5 – A systematic approach for pre-processing in multiblock modelling

Chapter 5. A systematic approach for pre-processing in multiblock modelling

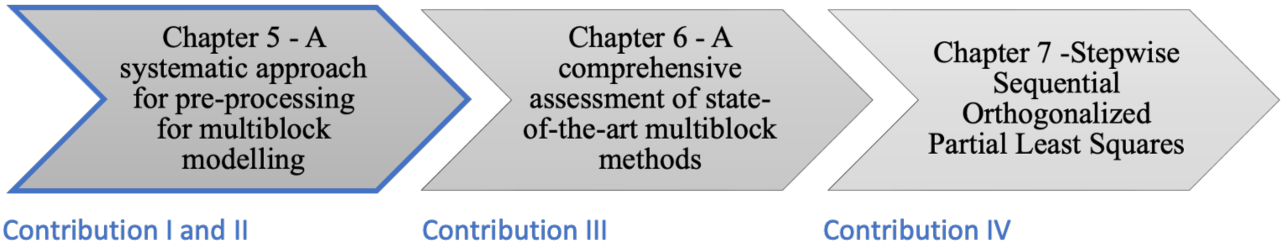

Chapter 5 - A systematic approach for pre-processing for multiblock modelling
Chapter 6 - A comprehensive assessment of state-of-the-art multiblock methods
Chapter 7 -Stepwise Sequential Orthogonalized Partial Least Squares
Contribution I and II
Contribution III
Contribution IV

Figure 5-1- Conceptual flow chart of the thesis organization. Chapter 5 is dedicated to the presentation of a systematic approach for selecting the adequate pre-processing methodology for multiblock modelling, as well as new pre-processing methods.

5.1 Introduction

The construction of multiblock models follows the same general workflow found in the development of data driven modelling approaches, schematically described in Figure 5-2. The data pre-processing step is an integral and critical part of any chemometric data analysis task, and the same applies for multiblock modelling. Pre-processing aims at maximizing the potential to extract useful information from data, which implies removing or mitigating undesirable effects such as artefacts (e.g. baseline and peak shifts, noise, systematic factors, etc.), while striving to bring out features that may be connected with the analysis goal (Engel et al., 2013, de Noord, 1994). Therefore, this step can greatly improve the Quality of Information extracted from data (Reis and Kenett, 2018) and has a strong influence in the success of any data analysis task, irrespectively of its goal (Famili et al., 1997). However, the choice of the adequate pre-processing method (or combination of methods) is definitely not a straightforward task. It depends on the structure of data, the purpose of the analysis and the modelling method selected. The order by which the pre-processing methods are applied is also critical and highly case dependent. Quite often, the definition of the pre-processing strategy entails extensive trial and error processes or is guided by the user accumulated experience on similar cases, or even by visual inspection. Recently, more sophisticated approaches have been proposed in the literature to support the selection of the best pre-processing method such as one based on statistical design of experiments (DoE) (Gerretzen et al., 2015) and others based on the use of quality parameters (such as the Pearson correlation coefficient (Wu et al., 2006) or the explained variance of the first principal component (Esquerre et al., 2012).

Typically, chemometric pre-processing methods are applied to a single block of predictors in order to handle what can be called the "intra-block" variability (i.e., within block variability). This is for instance the case of unit variance scaling and mean centering or the usual spectral pre-processing methods, such as standard normal variate, SNV, multiplicative scatter correction, MSC, and the Savitzky-Golay derivative filters. However, multiblock analysis requires the consideration of an additional degree of complexity arising from the "inter-block" variation (i.e., between blocks variability). This aspect should be carefully considered in order to avoid situations where some data blocks may dominate the analysis just because of size asymmetry effects.

Moreover, other steps also considered as pre-processing address distinct but important goals about the quality of data, such as the detection of outliers and the accommodation of missing values in the analysis. Data with missing and outlying entries are usually referred to as bad or contaminated data and violate the common statistical and data structure assumptions for data analysis, raising concerns on data accuracy and data integrity. The removal of outliers and/or the imputation of missing entries have been extensively reviewed in the literature and will not be covered in detail here (Zhu et al., 2018, Chiang et al., 2003, Arteaga and Ferrer, 2002, Nelson et al., 1996, Walczak and Massart, 2001, Møller et al., 2005).

In this chapter, the current state-of-the-art pre-processing methods for multiblock applications are collected and critically evaluated. The pre-processing strategy applied in multiblock methods is currently underdefined, and therefore another goal of the present work is to provide a systematic organization of the aspects that need to be considered when handling this important task to real problems. The applicability of such general systematic approach for defining the pre-processing strategy in multiblock applications is illustrated in two real case studies. The state-of-the-art block scaling methods are compared with more sophisticated methods and the prediction capability of the resulting models will be evaluated using a robust Monte Carlo cross-validation framework.

This chapter is organized as follows: in Section 5.2 the proposed pre-processing workflow based on a three-level approach for multiblock modelling is described; in Section 5.3 the results from two real word case studies are presented and in Section 5.4 the final chapter remarks are summarized.

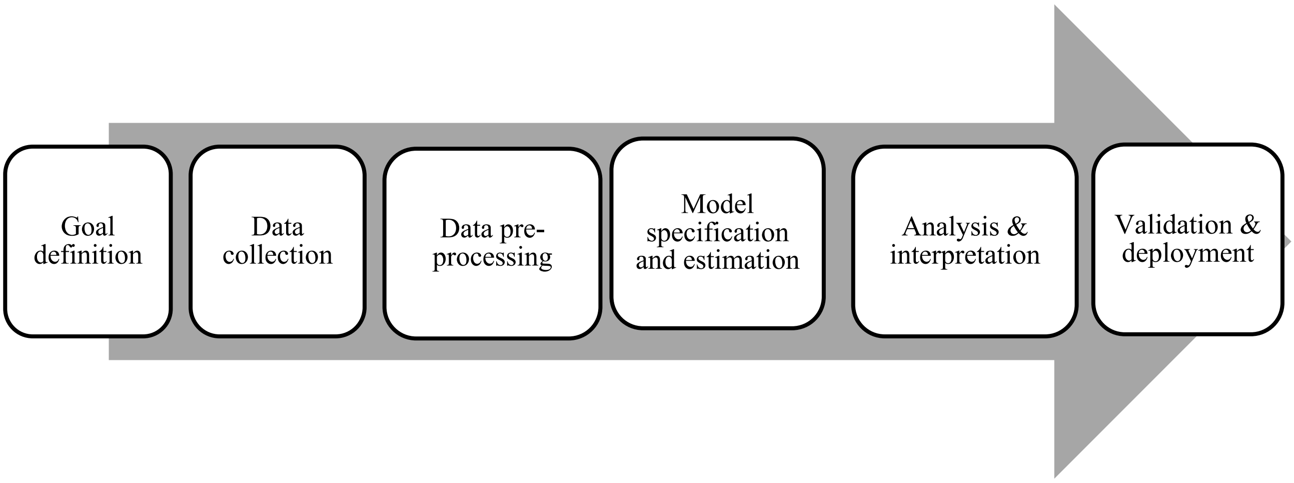


Figure 5-2 - General workflow for developing data driven modelling approaches, including multiblock modelling.

5.2 Pre-processing workflow for multiblock modelling – a three level approach

A pre-processing strategy for multiblock data analysis encompasses several stages where different aspects of the data structure must be considered in order to establish the best conditions for extracting the maximum information for achieving the analysis goal. Given the variety of aspects to consider and the alternative pre-processing methods available to handle them, a systematic approach that can guide users in multiblock analysis is highly desirable. Therefore, in this chapter such an approach is proposed based on past experience of addressing a variety of multiblock problems in different contexts and the results obtained.

The systematization macrostructure is composed by three fundamental stages that should be addressed sequentially – see Figure 5-3. In brief terms, the sequence of pre-processing steps proceed from the intra-block level to the inter-block level, as follows: Level I pre-processing is dedicated to the intra-block signal-to-noise ratio (SNR) improvement and feature enhancing; Level II performs intra-block scaling for equalizing the importance of the intra-block predictive components; finally, Level III is focused on balancing inter-block effects and making their contributions commensurate in the analysis. Not all the levels are required for every application and should therefore be evaluated case-by-case. However, their pertinence should always be sequentially considered before moving forward to the next stage. Each stage is discussed in detail in the next sections.

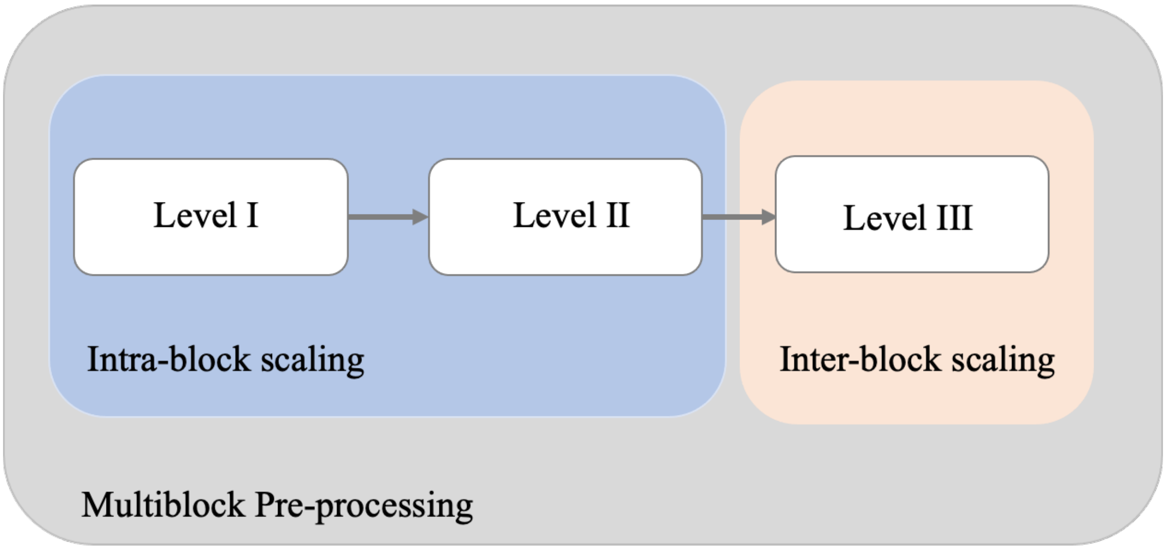


Figure 5-3- The three-level sequential approach for defining multiblock pre-processing

5.2.1 Level I pre-processing

The first pre-processing level aims at improving the quality of data by correcting for artefacts associated with each measurement system and/or by removing unwanted/uncorrelated variation with the desired phenomena. Fall in this category the variety of pre-processing methods for spectral and chromatographic data sets. Therefore, this level is particularly important for blocks composed by spectral data (such as NIR, MIR, NMR, RAMAN, fluorescence, UV, etc.) and data originated from chromatographic analytical methods. Table 5-1 presents an overview of the most common pre-processing methods available in the literature and the associated data artefacts they aim at addressing. These include, for instance, the presence of noise, which is a common feature in almost every analytical technique; baseline effects; light scattering artefacts, misalignments, and peak shifts. The Level I pre-processing strategy can be composed of more than one pre-processing method to address more than one type of data artefact. A large body of literature is currently available for addressing Level I pre-processing. Near-infrared spectroscopy is the spectroscopic technique that motivated the largest amount and diversity of Level I pre-processing techniques, because its spectra can be significantly influenced by non-linearity introduced by light scatter (Rinnan et al., 2009). All the artefacts referred above are intrinsically related to the analytical techniques employed to collect the data. However, artefacts can also be originated from sample-to-sample variation. In this case, they can be mitigated using row-wise normalisation approaches (Torgrip et al., 2008).

Level I methods are not restricted to data collected from process analytical technology devices. They can also be applied to industrial process data, namely, to increase the signal-to-noise ratio (Slišković et al., 2009), because unstructured variation sources can significantly blur the measurements and make the analysis less reliable.

Another group of pre-processing methods included in Level I is the one composed by supervised techniques, that require knowledge of the response variable, such as the Orthogonal Signal Correction (OSC) filters (Wold et al., 1998a). Soon after this method was published, a number of papers describing alternative OSC methods also became available (Fearn, 2000, Westerhuis et al., 2001, Andersson, 1999). These methods comprise primarily techniques that orthogonalize the data with respect to a reference response of interest.

The growing adoption of hyphenated techniques (such as combinations between chromatography and mass spectrometry based detection, e.g. GC-MS, or with IR detection, e.g. GC-IR, or with NMR detection, e.g. LC-NMR) is also bringing new additions to the Level I pre-processing toolkit, which are designed to handle more complex multiway data sets (Christin et al., 2008, Hendriks et al., 2011, Yi et al., 2016).

Table 5-1 - Overview of some well-known Level I pre-processing methods and the associated artefacts they aim at addressing.

Artefact		**Pre-processing method**	**Description**
Noise	Savitzky Golay filters	Local fitting of low-order polynomials and their subsequent use as an efficient filtering scheme.	(Savitzky and Golay, 1964)
	Wavelet based denoising	Data transformation that enables non-linear filtering and smoothing of signals with multiscale features.	(Reis et al., 2009, Rendall and Reis, 2014)
Baseline offset and slope	Detrending	Fitting of a polynomial of a fixed degree to the spectrum and subsequently subtraction of this polynomial from the spectrum.	(Barnes et al., 1989)
	Asymmetric Least Squares (AsLs)	A (Whittaker) smoother is used to estimate the baseline and asymmetric weighting of the deviations from smoothed signal is carried out such that the positive deviations with respect to baseline estimate are weighted (much) less than negative ones. The net effect is an automatic removal of background while avoiding the creation of highly negative peaks.	(Eilers, 2004, Eilers and Boelens, 2005)
	Iterative Polynomial Baseline Fitting (IPBF)	IPBF uses the same principle as de-trending and the basic difference is that in IPBF the fitting of the baseline is done iteratively. First a baseline of a chosen polynomial order is fitted to the sample spectrum. Then the measurement points lying above the estimated baseline are replaced by the predictions from the fitted baseline. This new (artificial) baseline spectrum is then fitted again with the same polynomial order, and this procedure is repeated until no new sample points are replaced. IPBF assumes that the baseline of the spectrum is given by the lowest points along the spectrum.	(Lieber and Mahadevan-Jansen, 2003)

Artefact		Pre-processing method	Description
	Derivatives	Consists of taking the derivative of the measured responses with respect to the variable number (index) or other relevant axis scale (wavelength, wavenumbers, etc.). Derivatives have the capability of removing both the additive and multiplicative effects in the data. A first-order derivative eliminates constant baseline (offsets) and a second-order derivative will also eliminate the baseline slope. Because derivatives de-emphasize lower frequencies and emphasize higher frequencies, they tend to accentuate noise (high frequency signal). For this reason, smoothing techniques such as Savitzky-Golay algorithm or the Norris Williams are used in order not to reduce the signal to noise ratio in the signal too much.	(Savitzky and Golay, 1964, Brown et al., 2000, Martens and Russwurm Jr, 1982, Norris and Jux, 1984)
Light scatter (additive or multiplicative perturbations)	Standard Normal Variate (SNV)	Row-wise scaling operation that removes the spectrum mean from all the spectrum variables and divide them by the spectrum standard deviation.	(Barnes et al., 1989)
	Robust Normal Variate (RNV)	Row-wise scaling operation that removes the spectrum median from all the spectrum variables and divides them by the spectrum robust standard deviation.	(Guo et al., 1999)
	Multiplicative Scatter Correction (MSC)	Also known as Multiplicative Signal Correction. The principle is to fit each spectrum to a reference spectrum and then to correct them. The reference spectrum must be representative thus generally the average calibration spectrum is used. MSC has given rise to other related methodologies such as the Extended Multiplicative Signal Correction (EMSC), which is based on a polynomial baseline correction depending on the wavelength.	(Geladi et al., 1985, Martens and Stark, 1991)
	Normalization	Divide each spectrum by an estimation of its spectral intensity. This can be done using the following properties: area (area normalisation), maximal peak (maximum normalisation), a specific spectral point (peak normalisation), length (unit vector normalisation), or the sum of the spectral values.	(Rinnan et al., 2009)

Artefact		**Pre-processing method**	**Description**
Temporal or spectral misalignment	Parametric Time Warping (PTW)	Polynomial transformation of the time axis that leads to maximal overlap between two samples. It is a global alignment method. The degree of the warping function can be chosen by the user: a zeroth-order warping signifies a constant shift, whereas a first-order function also introduces stretching or compression. Higher-order terms allow for even more complex behaviour.	(Eilers, 2004)
	Correlation Optimized Warping (COW)	Segment wise warping method consisting in splitting the signal into different segments and optimally align them to match the segments of the reference profile by linear stretching and shifting the points along the time axis.	(Skov et al., 2006)
	Dynamic Time Warping (DTW)	Point-wise warping method of the signal. It calculates the Euclidean distance between the target and the reference signal. An improved version of the method emerged later called Variable Penalty Dynamic Warping (VPdtw) in which a variable penalty is introduced into DTW process added to the distance metric when the signal is expanded or contracted.	(Ramaker et al., 2003, Clifford et al., 2009)
	Interval Correlation Optimised Shifting (icoshift)	Aligns each independent signal to a target signal by maximizing the cross-correlation between user-defined intervals	(Savorani et al., 2010)
	Binning	the signal is split into many segments (called bins). The integral of the signal, or, alternatively, the maximum intensity in each bin is used as a replacement for the original signal, thereby reducing the effect of small misalignment	(De Meyer et al., 2008)

Artefact		**Pre-processing method**	**Description**
Variation unrelated with response or variation from external factors	Orthogonal Signal Correction (OSC)	Removes variation from a data matrix **X** that is orthogonal to the response matrix **Y**.	(Fearn, 2000)
	Direct Orthogonal Signal Correction (DOSC)	Calculates directions in **X** that are orthogonal to **Y** and that account for the largest variance of **X**. These directions are obtained by least squares steps.	(Fearn, 2000, Westerhuis et al., 2001, Andersson, 1999).
	Direct orthogonalization (DO)	The principle of direct orthogonalization is to establish an orthogonal model with scores independent of the variables being modelled, and a conventional regression model on the data not extracted by the orthogonal model.	(Fearn, 2000, Westerhuis et al., 2001, Andersson, 1999).
	External parameter orthogonalization (EPO)	Estimates the space in which the influence of external factors occurs and removes it from the **X** matrix by orthogonal projection.	(Roger et al., 2003)

5.2.2 Level II pre-processing

Level II pre-processing methods equalize the contributions of all variables within each block. Methods described in Level I are, with higher incidence row-wise methods, i.e., the pre-processing is carried out sample by sample. Level II methods, on the other hand, are typically column wise treatments. They include the classical most widely used pre-processing methods of mean centering and scaling.

The mean centering pre-treatment (Bro and Smilde, 2003, van den Berg et al., 2006) consists of subtracting the average (mean) of each variable (column of the data matrix). This corresponds to repositioning the coordinate system in the origin (i.e., removal of the offset) leaving only the variation between samples (from the mean) for analysis. This step makes the modelling task simpler, avoiding the existence of components dedicated to the explanation of the mean levels of all the existing variables.

Scaling, on the other hand, is employed in order to make the different block variables comparable in importance before applying scale-dependent multivariate analysis methods, such as PCA, PCR or PLS. The most common technique is the unit variance scaling method (Bro and Smilde, 2003, van den Berg et al., 2006) where variables are divided by their respective standard deviations. The method is commonly applied to data sets containing variables with different units and scales (e.g., pH, volumes, flows, temperatures, etc.) in order to impose equal weights in the analysis. Other scaling techniques are also available, such as Pareto scaling (Eriksson, 1999), range scaling (Smilde et al., 2005), level scaling (van den Berg *et al.*, 2006), vast scaling (Keun et al., 2003) or scaling tailored to stress the importance of specific variables by giving them relatively higher weights. Typically, scaling methods are applied together with mean centering.

Variable transformations are also included in this group. An example is the column-wise application of nonlinear operations to correct for heteroscedasticity, such as the logarithmic transformation and the family of power transformations (Kvalheim et al., 1994).

5.2.3 Level III pre-processing

Level I and II are composed by single block techniques targeting the optimization of intra-block components of variation for analysis. Level III, on the other hand, addresses the equalization and tuning of inter-block systematic effects. The following important questions are addressed at Level III pre-processing: i) How to handle the existence of different scales in the blocks? ii) How to cope with their different sizes (number of predictors in each block)? iii) How to deal with different pseudo-ranks (underlying latent variable dimensionality)? Blocks may span different scales (for example a block composed by spectral data versus another with process data composed by, for instance, volume, pressure and temperature variables) and there may also be orders of magnitude differences between their sizes (a spectral block with hundreds or thousands of wavelength variables versus a process data block with dozens of variables). These effects can introduce significant bias in the analysis if not properly handled, as multivariate methods tend to favour higher variation, meaning larger blocks. Therefore, it is important to emphasize that the different blocks should be properly weighted before being subjected to multiblock analysis, in order to give them a priori the relative importance they should have or that we deem more appropriate for them to have. Usually this means equal importance, but sometimes it might also be the case where the intention is to increase or decrease the importance of certain blocks. This latter case is particularly important when there is some *a priori* knowledge available or when some blocks are especially noisy or affected by high levels of measurement uncertainty. Moreover, combining blocks with different underlying latent variable dimensionality can be challenging for example in situations where one of the blocks is a design matrix while the other one is a highly multivariate and collinear data set. More recently multiblock methods have been proposed in the literature for addressing these aspects such as the SO-PLS and PO-PLS. In fact, both SO-PLS and PO-PLS are potentially less sensitive to the relative weighting of the blocks and they can also handle situations with different underlying dimensionalities (ranks) of the blocks. The reason for this is that both algorithms incorporate one block at a time by sequentially performing PLS regression on matrices that are orthogonalized with respect to each other (Naes et al., 2013, P. Campos et al., 2018, Jørgensen et al., 2007, Måge et al., 2008, Jørgensen et al., 2004). However, the order by which the blocks are considered may result from a comparison of the quality of the models preliminarily derived when they are considered alternatively together with others already incorporated in the model. Therefore, the blocks inner scaling may end up still having some effect on the analysis, even though in a more subtle way.

It has been proven that, by optimizing inter-block pre-processing, it is possible to increase the prediction capability of the method and enhance the interpretation of the results as well (Campos et al., 2017). Therefore, proper scaling to account for the inter-block variability is imperative prior to modelling. Inter-block scaling approaches are also known as block scaling approaches and can be generalized by the following equation 4:

$$\boldsymbol{X}_{\boldsymbol{b}\,\textbf{Level III}\,pre-processed} = \frac{\boldsymbol{X}_{\boldsymbol{b}}}{k_b} \tag{4}$$

where $\boldsymbol{X}_{\boldsymbol{b}\,\textbf{Level III}\,pre-processed}$ stands for block indexed by b (b=1, … , B), after block scaling is applied; $\boldsymbol{X}_{\boldsymbol{b}}$ is the single block b before inter-block scaling; k_b is a block-scaling factor applied for collectively scaling the entire block b.

Level III pre-processing methods include the soft-scaling and hard-block scaling methods that are considered to be the state-of-the-art techniques for multiblock analysis (Eriksson et al., 2013) – see Table 5-2. These methods define the weighting factor depending on the number of variables in the block. In soft block scaling, each block of variables is scaled such that the sum of the variables variance after scaling equals the square root of the number of variables in that particular block. Hard block scaling introduces further down-weighting and with this approach all variables in a block are scaled such that the sum of their variances is unity. Another possibility is to simply scale each block by their number of variables. This is called super hard block scaling, in which the sum of the variable's standard deviations equals to unity. The main purpose of these methods is to account for the different block sizes and to avoid the case where larger blocks of variables dominate over smaller blocks of variables.

Table 5-2 - Classical multiblock scaling approaches (Eriksson et al., 2013)

Scaling method	**Goal**
Soft block scaling	For each block b: $\sum_{i=1}^{\mathrm{p}} \sigma_i^2 = \sqrt{\mathrm{p}_{block}}$
Hard block scaling	For each block b: $\sum_{i=1}^{\mathrm{p}} \sigma_i^2 = 1$
Super hard block scaling	For each block b: $\sum_{i=1}^{p} \sigma_i = 1$

Legend:

p_{block} – number of variables in the block X_b (b=1…B)

σ_i – standard deviation of each block X_b column (variable) (i=1…p)

A requirement (and possibly also a limitation) of the three above mentioned scaling methods is that data blocks have to be preliminarily scaled to unit variance during Level II pre-processing. Therefore, in cases where the blocks were previously pre-processed using mean centering only (which is typically the case with spectral or chromatographic data sets) the above-mentioned methods do not apply anymore. In reference (Campos et al., 2017), new block scaling approaches were presented and discussed, that can be used when the blocks are previously scaled to unit variance or in situations where all variables of a block have the same units and the block is mean centered. These approaches use as block scaling factor (see Equation 4) the block standard deviation rather than the number of variables in each block (as it is the case for soft, hard and super hard block scaling methods) and are here referred to as block variance scaling approaches. These block variance scaling approaches can be performed independently of the Level II (and Level I) pre-processing methods applied and lead to the same expected results in terms of the sum of the variables variance after scaling, while ensuring equal importance to all blocks. It is also important to mention that the standardization performed during the block variance scaling approaches is matrix-wise; thus, the standard deviation ratio between informative and noisy variables is maintained (opposite to unit variance scaling that is a column-wise type of standardization).

The state-of-the-art block scaling methods and the block variance scaling methods described above aim at providing a balanced importance to all blocks. However, the relative importance of each block can also be set according to iterative algorithms that alternate between model block-scaling and building stages. The underlying rational lies in the recognition that even when it is known that some blocks may be more important, there is no information about which weights should be used to reflect their relative importance.

In this chapter yet another group of block scaling methods is proposed in which the standardization is based on the underlying dimensionality of each block determined by PCA. These methods are here called block rank scaling methods and besides eliminating the systematic effects of scale and size, they increase the importance of blocks according to the relevant sources of uncorrelated variation they can bring to the analysis.

Finally, for the sake of completeness, a block scaling method described in Campos et al. (2017) that penalizes blocks with higher underlying latent variability will also be studied in this paper for comparison purposes. This method is here referred to as pseudorank penalization block scaling approach.

In this regard, the following groups of Level III inter-block scaling methods will be evaluated in this study: the state-of-the-art block scaling methods; variance block scaling methods; block rank scaling methods and the pseudorank penalization block scaling approach. Table 5-3 summarizes the different block scaling methods covered in this work.

Table 5-3 - Block scaling methods considered in the present work.

Method		**Scaling factor, $\boldsymbol{k_b}$**
NS	No Scaling	-
Block Scaling methods[(a)]		
SBS	Soft-Block Scaling	$\frac{1}{\sqrt[4]{p_{block\ j}}}$
HBS	Hard-Block Scaling	$\frac{1}{\sqrt{p_{block\ j}}}$
SHBS	Super Hard-Block Scaling	$\frac{1}{p_{block\ i}}$
Block Variance Scaling methods[(b)]		
SBVS	Soft-Block Variance Scaling	$\frac{\sqrt[4]{p_{block\ j}}}{\sqrt{\sum \sigma_i^2}}$
HBVS	Hard-Block Variance Scaling	$\frac{1}{\sqrt{\sum \sigma_i^2}}$
SHBVS	Super Hard-Block Variance Scaling	$\frac{1}{\sum \sigma_i}$
Block Rank Scaling methods [(b)]		
SBRS	Soft-Block Rank Scaling	$\frac{\sqrt[4]{p_{block\ j}}}{\sqrt{\sum \sigma_i^2}} \times pseudorank^{1/k}$
HBRS	Hard-Block Rank Scaling	$\frac{1}{\sqrt{\sum \sigma_i^2}} \times pseudorank^{1/k}$
SHBRS	Super Hard-Block Rank Scaling	$\frac{1}{\sum \sigma_i} \times pseudorank^{1/k}$
PPBS	Pseudorank Penalization Block Scaling	$\frac{1}{block\ pseudorank}$

Legend:

p_{block} – number of variables in the block X_b (b=1…B)

σ_i – standard deviation of each block X_b column (variable) (i=1…p)

[(a)] For these methods, data is assumed to be previously scaled to unit variance.

[(b)] For these methods, data is either scaled to unit variance or mean centered and all variables within the same block have the same units (e.g., spectroscopic data sets).

5.3 Results

In this section, the two real case studies described in detail in Chapter 4 will be used to demonstrate the applicability of the general systematic approach for multiblock scaling described in the previous sections. A critical evaluation of the different block scaling methods summarized in Table 5-3 is also performed in terms of the prediction capability achieved in the final models.

The first case study regards the problem of wine ageing prediction. In particular, the aim is to develop a multiblock approach to predict the ageing time of Madeira wine, making use of all analytical measurement sources available that can potentially bring different aspects to the modelling. Such a model finds interesting applications in quality prediction, process monitoring and fraud detection, besides increasing knowledge about the complex network of reactions and their evolution over time. The second case study refers to biodiesel production, where the final quality of the product is to be predicted based on raw materials information captured by NIR spectral data together with known operation conditions.

In both case studies the data blocks will be modelled by concatenating all blocks in a single augmented matrix, side by side, followed by the application of PLS method (see Chapter 2.3.1 for more details on the Concatenated PLS method). Prior to modelling, all the blocks are pre-processed according to the generalized three levels pre-processing strategy described in Chapter 5.1. The prepressing strategy Level I and Level II are fixed, as this study is mainly dedicated to the analysis of the Level III approaches. In this way, the study presented in the following sections evaluates the influence of the pre-processing strategy (particularly Level III) applied on the model performance, and not the modelling approach used (to avoid mixing the effects from these two aspects, only one was varied, the pre-processing). The different block scaling methods presented in Table 5-3 will be employed and the prediction capability of the resulting models will be evaluated based on the distribution of the root mean square errors of prediction (RMSEP) calculated via a Monte Carlo double cross-validation on 50 models developed for each method under study.

5.3.1 *Results for case study 1: Madeira wine ageing*

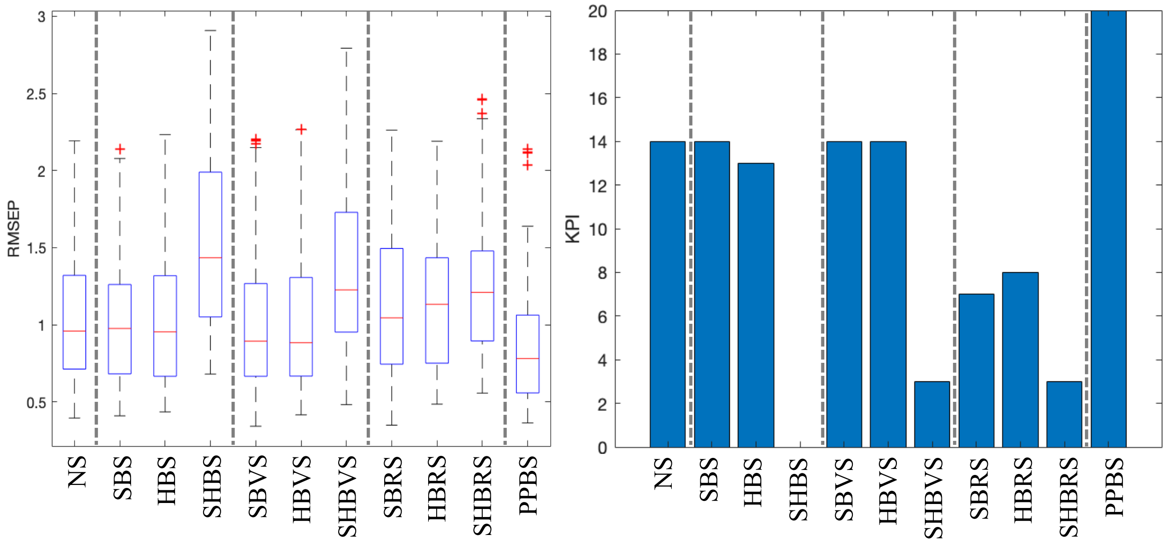
The pre-processing strategy applied (Level I, II and III) applied to each analytical data block is summarized in Table 5-4. The choice of pre-processing methods applied for Level I and II was based on prior knowledge and past experience on handling these types of matrices.

Table 5-4 - Pre-processing strategy applied to each analytical data bloc*k*

Block#	Description	Level I Pre-process	Level II Pre-process	Level III Pre-process
1	Organic Acids (52×8)	-	Unit variance scaling and Mean Center	See Table 5-3

Block#	Description	Level I Pre-process	Level II Pre-process	Level III Pre-process
2	Volatile Compounds (52×81)	-	Unit variance scaling and Mean Center	
3	Polyphenols (52×25)	-	Unit variance scaling and Mean Center	
4	UV-Vis spectra (52×109)	SNV	Mean Center	

Models were developed for the prediction of wine age differing only on the pre-processing strategy applied – see Table 5-4. The prediction results, namely the RMSEPs and the KPIs obtained from the models developed using the different pre-processing strategies studied in this work, are displayed on Figure 5-4 (the KPI is such that the higher its value, the better the corresponding method is). The graph on the left side of Figure 5-4 shows the distributions of the RMSEP results obtained from the 50 Monte Carlo runs and using the different block scaling approaches (Level III); the graph on the right shows the KPIs after the paired t-test comparison of the methods.



NS	No Scaling
SBS	Soft-Block Scaling
HBS	Hard-Block Scaling
SHBS	Super Hard-Block Scaling
SBVS	Soft-Block Variance Scaling
HBVS	Hard-Block Variance Scaling
SHBVS	Super Hard-Block Variance Scaling
SBRS	Soft-Block Rank Scaling
HBRS	Hard-Block Rank Scaling
SHBRS	Super Hard-Block Rank Scaling
PPBS	Pseudorank Penalization Block Scaling

Figure 5-4 - Impact of different block scaling approaches in the Concatenated PLS method prediction capability for case study 1: Madeira wine ageing. Graph on the left: comparison of the RMSEP distribution obtained for each block scaling method; Graph on the right: relative performance assessment using the Monte Carlo comparison framework

The method showing the best relative prediction performance in this case study is the PPBS (presenting the highest KPI score on the graph on the right side of Figure 5-4) in which the blocks are penalized by the potential analytical information that each block brings to the analysis (pseudorank). This pseudorank is determined in an unsupervised way, i.e., by principal components analysis (PCA), and consists of the number of principal components that can explain at least 90% of the block overall variability. On the other hand, the methods showing the worst relative prediction performances (with the lowest KPI values graph on the right side of Figure 5-4 are the SHBS, SHBVS and SHBRS (i.e., all the "super hard" versions of each group), which are also those where the factors most downweigh the blocks in terms of the corresponding scaling attributes (number of variables, standard deviation or rank).

Moreover, in general, when comparing the block variance scaling methods with the state-of-the-art block scaling methods, one can verify that they lead to better relative prediction performances with statistically significant difference existing between the block variance scaling methods HBVS and SHBVS and the corresponding state-of-the-art methods HBS and SHBS. As for the soft scaling methods, SVBS and SBS, it is observed that both the variance and the state-of-the-art method show similar prediction performance (same KPI value obtained). Since the block scaling variance methods use as scaling factor the block standard deviation, they can be applied in cases where the blocks were previously mean centered or scaled to unit variance. Note that, when the pre-processing method selected in Level II is unit variance scaling, then there will be no difference between the block variance scaling methods and the state-of-the-art block scaling methods. This may have a considerable practical value, as current practice consists in using scaling methods based on the number of variables in each block.

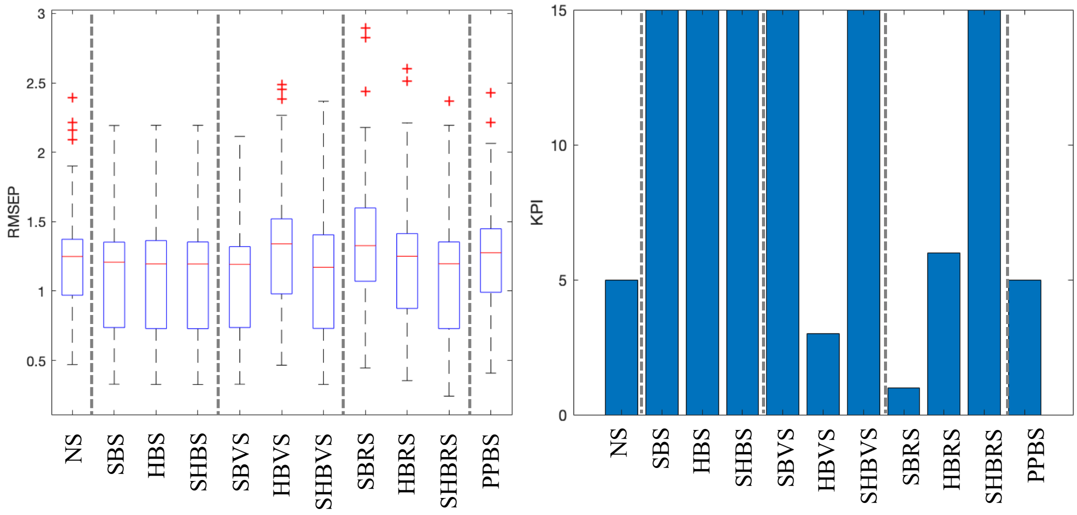
5.3.2 *Results from case study 2: Biodiesel production*

In this case study, the data block with the operation conditions was preliminary scaled to unit variance and mean centering, following standard practice (Level II; no Level I required) while the block with NIR spectra was pre-processed with SNV (Level I) and mean centering (Level II). Table 5-5 summarizes the pre-processing strategy applied for each data block. The choice of pre-processing methods applied for Level I and II was based on prior knowledge and past experience on handling these types of matrices.

Table 5-5 - Pre-processing strategy applied to each predictor data block

Block#	Description	Level I Preprocess	Level II Preprocess	Level II I Preprocess
1	NIR (21×584)	SNV	Mean Center	See Table 5-3
2	Operating conditions (21×2)	-	Unit variance scaling and Mean Center	

The results from the systematic comparison framework for the biodiesel case study are summarized in Figure 5-5. The graph on the right side of the figure shows the KPI results from the paired t-test comparison between the block scaling methods under study and the graph on the left shows the distributions from the RMSEPs obtained from the 50 outer Monte Carlo runs for each method. Several methods show equal performance in terms of prediction capabilities, namely SBS, HBS and SHBS (from the block scaling methods group); SBVS and SHBVS (from the block variance scaling group) and SHBRS (from the block rank scaling group). As can be observed from Figure 5-5, PPBS is not performing so well in this case study, when compared to the results obtained in the previous case study. In this case study it is observed that the variance block scaling methods have equal relative performance as the correspondent state-of-the-art block scaling methods except for the hard block variance scaling method (HBVS).



NS	No Scaling
SBS	Soft-Block Scaling
HBS	Hard-Block Scaling
SHBS	Super Hard-Block Scaling
SBVS	Soft-Block Variance Scaling
HBVS	Hard-Block Variance Scaling
SHBVS	Super Hard-Block Variance Scaling
SBRS	Soft-Block Rank Scaling
HBRS	Hard-Block Rank Scaling
SHBRS	Super Hard-Block Rank Scaling
PPBS	Pseudorank Penalization Block Scaling

Figure 5-5 - Impact of different block scaling approaches in the Concatenated PLS method prediction capability for case study 2: biodiesel production. Graph on the left: comparison of the RMSEP distribution obtained for each block scaling method; Graph on the right: relative performance assessment using the Monte Carlo comparison framework.

5.4 Chapter final remarks

A structured strategy for implementing pre-processing for multiblock modelling was presented and its application was illustrated in two real case studies. The pre-processing strategy applied is a critical step in the development of chemometric methods with significant impact on the analysis outcomes.

In the case of multiblock methods, one needs to address not only the intra-block variability as in single block methods but also the variability inter-blocks prior to the data analysis activity. It is imperative to reiterate that the pre-processing method applied aims at giving equal importance to each data block before modelling (unless there is *a priori* knowledge about relative importance of certain blocks) and if the pre-processing is not applied correctly this is not achieved and certain blocks can dominate the analysis over other blocks (e.g., due to significant different size, or underlying dimensionality).

A systematic approach for pre-processing in multiblock modelling was proposed where intra-block effects regarding the data quality (Level I) and variables balancing (Level II) are first handled followed by the equalization and tuning of the inter-block effects (Level III). In general, Level I methods are row-wise, Level II methods column-wise, and Level III methods matrix-wise.

Having a systematic workflow to guide the practitioners during the selection and efficient application of pre-processing for multiblock approaches can significantly improve the development of multiblock methods and also contribute to increase their usage and overreaching to practitioners.

Level III group of methods has been quite under explored in the literature and the current available state-of-the-art block scaling methods present some limitations. Therefore, in these thesis three new groups of inter-block methods were proposed: variance block scaling methods; block rank scaling methods and the pseudorank penalization block scaling methods. Block scaling methods already existing in the literature use as scaling factor the number of variables of each block whereas the new proposed methods are based on the blocks variance and the blocks underlying latent dimensionality. The new proposed methods can be used applied independently of Level I and Level II pre-processing, i.e., in cases where the blocks were previously mean centered or scaled to unit variance. This may have a considerable practical value, as current state-of-the-art methods can only be applied when blocks were previously scaled to unit variance. All Level III methods found in the literature were compared with these new proposed inter-block pre-processing methods and their relative merits explored, case by case, with resort to two case studies.

The results presented in this chapter demonstrate that in general it is not possible to anticipate what the best block scaling method will be, and this has to be determined case by case. The selection of the best block scaling method needs to be based on a systematic testing and comparison framework based on the data collected and the purpose of the data analysis. In other words, the pre-processing strategy to be applied is a function of the data in hand, data analysis method selected and purpose of the analysis (Reis and Kenett, 2018).

In addition, multiblock pre-processing must be carefully planned in accordance with the multiblock analysis to be performed as not all multiblock approaches require all levels of pre-processing. For example, the sequential and parallel approaches to partial least squares regression, SO-PLS and PO-PLS respectively, are less sensitive to the relative scaling of the blocks and can also deal with the differences in the ranks of multiblock data.

Chapter 6 – A comprehensive assessment of state-of-the-art multiblock methods

Chapter 6. A comprehensive assessment of state-of-the-art multiblock methods

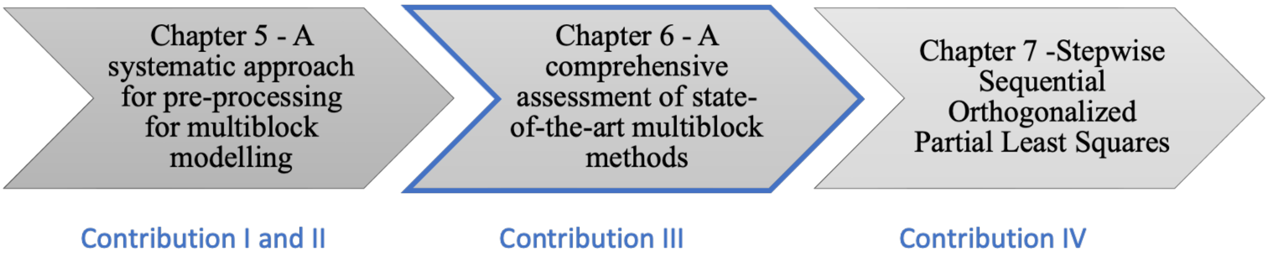

Chapter 5 - A systematic approach for pre-processing for multiblock modelling
Chapter 6 - A comprehensive assessment of state-of-the-art multiblock methods
Chapter 7 -Stepwise Sequential Orthogonalized Partial Least Squares
Contribution I and II
Contribution III
Contribution IV

Figure 6-1 - Conceptual flow chart of the thesis organization. Chapter 6 is dedicated to the presentation of a comprehensive assessment of state-of-the-art multiblock methods

6.1 Introduction

Multiblock methods have been object of renewed interest lately due to the increasing amounts of data generated in many fields, where these methodologies can add value to the analysis (Borràs E, Ferré J, Boqué R, Mestres M, Aceña L, Busto, 2015; Blanchet L, Smolinska, 2016). However, the potential of multiblock methods is still underexplored, both regarding the prediction capabilities and the value of interpretative information they are able to generate, which is another relevant aspect that this class of methodologies is capable to bring. In addition, in order to further disseminate the application of these methodologies a unified and systematic workflow to for their application as well as a robust and easy to use and interpret framework for assessing the existing methodologies performance is lacking.

The goal of this chapter is to provide a critical assessment of a rich variety of multiblock regression methods described in detail in Chapter 2: Concatenated PLS method, Hierarchical PLS (HPLS) (Slama, 1991, Wold et al., 1987b, Wold et al., 1996), Multiblock PLS (MBPLS) (Wangen and Kowalski, 1989, Wold et al., 1983), Network-Induced Supervised Learning (NI-SL) (Reis, 2013b), and Sequential Orthogonalized PLS (SO-PLS) (Jørgensen et al., 2004, Næs et al., 2011b, Jørgensen et al., 2007). This evaluation uses a robust Monte Carlo statistical framework described in Chapter 2.4.

The study presented in this Chapter 6 explores and reveals potential advantages of applying the current state-of-the-art multiblock methods for fusing data sets from different sources, both from the predictive and interpretability perspectives. In parallel this study also highlights relative weaknesses of these state-of-the-art methods and improvement opportunities.

For the purpose of this work, the problem of ageing time prediction of one of the finest Portuguese fortified wines, the Madeira wine, is considered for which data collected from 4 types of analytical sources (HPLC, GC-MS, UV-Vis and the content of organic acids) is explored simultaneously – please refer to Chapter 4 for a detailed description of the case study. The different sources of data have the potential of bringing complementary information about the phenomenon under analysis, but their use in simultaneous may significantly increase the associated cost, especially if the number of samples is high, or if the goal is to develop a new routine procedure to be implemented in a given industrial process. Furthermore, quite often, different techniques present a significant overlap in their information content and are affected to a great extent by the same structural features of the samples, leading to highly redundant information, with little added-value resulting from their combination. Problems like these are particularly frequent in modern analytical laboratories, and therefore are relevant and opportune to address.

In this context the information from different data sources have to be combined in such a way to give better predictions of the product quality as well as in order to extract maximum information from the system and explore the inter-relationships between the different data sources. One approach would be to analyse each block separately by means of classical single block multivariate data analysis methods. This approach was extensively addressed in Rendall et al. (2017), and the interested reader is remitted to this reference for more information. Other approach for handling multiple blocks of variables would be to analyse them altogether after side-by-side concatenation into a single augmented data block. However, in this approach the integrity of each data source is not retained, which can significantly blur the analysis of the final results as well as limit its predictive performance in test conditions. Moreover, the solution will depend heavily on how the different variable blocks are scaled relatively to each other and there may also be problems related to the different dimensionality of the blocks under analysis. Therefore, the more efficient way by far of modelling multiple blocks of data is by applying multiblock approaches that retain the integrity of each data block in the model. A comparison between the results obtained from combining all the data sources using multiblock approaches and the results from the application of single block approaches to each individual data blocks is also included in this study.

This chapter is organized as follows: Section 6.2 describes the assessment and comparison methodology applied in this study; Section 6.3 is dedicated to the analysis of the prediction accuracy, whereas Section 6.4 addresses the interpretation capabilities of the multiblock methods under investigation. The final remarks of this work are summarized in Section 6.5.

6.2 Assessment and comparison methodology

The critical assessment of the multiblock methods presented in this chapter is focused on two aspects: relative prediction performance of the methods and their interpretation capabilities as a result of maintaining the structural integrity of the blocks.

A robust Monte Carlo statistical framework (see Chapter 2.4 for detailed information) is employed to compare and critically analyse the selected state-of-the-art methods (Concatenated PLS method, HPLS, MBPLS, NI-SL and SO-PLS) regarding their predictive and interpretation capabilities.

In the implementation of this framework, care must be taken on how pre-processing is conducted as it can have a significant impact on the analysis outcomes. As described in detail in Chapter 5, in multiblock methods the sequence of pre-processing steps proceeds from the intra-block level to the inter-block level following a systematic three level approach. The pre-processing methods to account for the intra-variability of each data block (Level I and Level II) were selected based on past experience on handling these types of matrices as follows. For the polyphenols, volatile compounds and organic acids data sets, mean centering and unit variance scaling was used as the pre-processing approach in order to correct for the differences in the measurement units and because no prior information regarding the variables' importance was available. For the UV–Vis data set, standard normal variate (SNV) followed by mean centering were applied. Variable centering and scaling were employed in a consistent manner so that the results of the comparison are not biased: the training data is used to estimate the scaling parameters (e.g., mean and variance for auto-scaling), which are then used to scale the test data. The intra-block pre-processing was applied consistently for all tested multiblock approaches. The inter-block pre-processing (Level III pre-processing) was evaluated case-by-case since not all methods require inter-block pre-processing (e.g., SO-PLS). For the concatenated PLS method, different block scaling methods were evaluated to account for the inter-variability between blocks– current state-of-the-art block scaling methods and new proposed scaling methods were evaluated (see Chapter 5).

The distribution of the root mean square errors of prediction (RMSEP) obtained for the models developed using the robust Monte Carlo framework characterize the prediction performance and robustness of each method. In a more rigorous way, the relative prediction performance of each method is compared using the KPIs calculated from the paired hypothesis tests.

Moreover, a comparison with single block linear latent variables methodologies (viz. PLS and PCR) was also carried out and the results are summarized in Chapter 6.3.2. In Rendall et al. (2017) the same case study concerning wine age prediction was used to compare the relative prediction performance of a selected pool of single block methods by means of similar Monte Carlo double cross validation framework. The comparison between multiblock and single block methods was based on the results obtained for the coefficient of determination and RMSEP.

On the other hand, the interpretation capabilities of the state-of-the-art multiblock methods are assessed based on the ability of each method to bring more information to the analysis (e.g., quantitative measures on the amount of redundant information among the multiple predictor blocks).

6.3 Results - Prediction accuracy

6.3.1 Prediction capability of multiblock methods

This section presents the results of the extensive comparison study of the state-of-the-art multiblock algorithms (Concatenated PLS method, HPLS, MBPLS, NI-SL and SO-PLS) in terms of their ability to predict the ageing time of Portuguese Madeira wine. Figure 6-2 presents the comparison of the KPI values obtained for each multiblock method (computed as described in chapter 2.4). From the analysis of Figure 6-2, it can be observed that SO-PLS and the Concatenated PLS method tend to present superior results in terms of prediction accuracy with no significant differences observed between them.

As mentioned in Chapter 2.3, the SO-PLS method depends on the block order. In this study, all possible block orders were tested and the one leading to the best prediction results was selected for the comparison study. The resulting best order for the analytical data blocks was: polyphenols content → volatile profile → organic acids content → UV-Vis spectra. This order is also supported by prior knowledge accumulated during the application of single block methods in Rendall et al. (2017). More specifically, this result shows that maximum information is being retrieved from the most relevant blocks and then only residual additional information, that was not used before to predict wine age, is sequentially incorporated in the model in the subsequent stages. This might explain why this method is showing such a good relative performance. However, it should be mentioned that if no *a priori* information is available for the definition of the blocks order, it can be challenging and time consuming to select the best order when several blocks are available.

Regarding the Concatenated PLS method, an optimization of the block scaling method was performed and the best result in terms of prediction ability was obtained with the scaling based on the block pseudorank determined by PCA. As discussed in Chapter 5 the way each data block is scaled prior to modelling has a significant impact on model performance when the blocks are concatenated side by side and a PLS is applied to the augmented matrix. In this case the best pre-processing method to account for the inter-block variability was determined by applying different methods and evaluating the resulting the final model prediction accuracy (the distributions of the RMSEP obtained for 50 models developed during the Monte Carlo runs and using the different block scaling methods).

Even though SO-PLS and the Concatenated PLS method present similar relative performances, one can argue that SO-PLS has the advantage of not requiring block scaling and most importantly have the desirable feature of generating more interpretable information. The following best performing method is NI-SL followed by MBPLS with the two different deflation approaches showing identical relative performances in terms of prediction ability (no significant differences were observed between them in the present case). HPLS is the method showing the worst relative prediction performance. This is expected since the final regression with **Y** is based on block scores determined in an unsupervised way, being therefore expectable that they may miss some relevant components with predictive interest. In the methods MBPLS and HPLS, the block scaling was conducted by dividing the block scores by the square root of the number of variables in the correspondent block. Similarly, to the optimization carried out for the Concatenated PLS method, it is expected that the performance of these methods improves if the block scaling method applied is optimized.

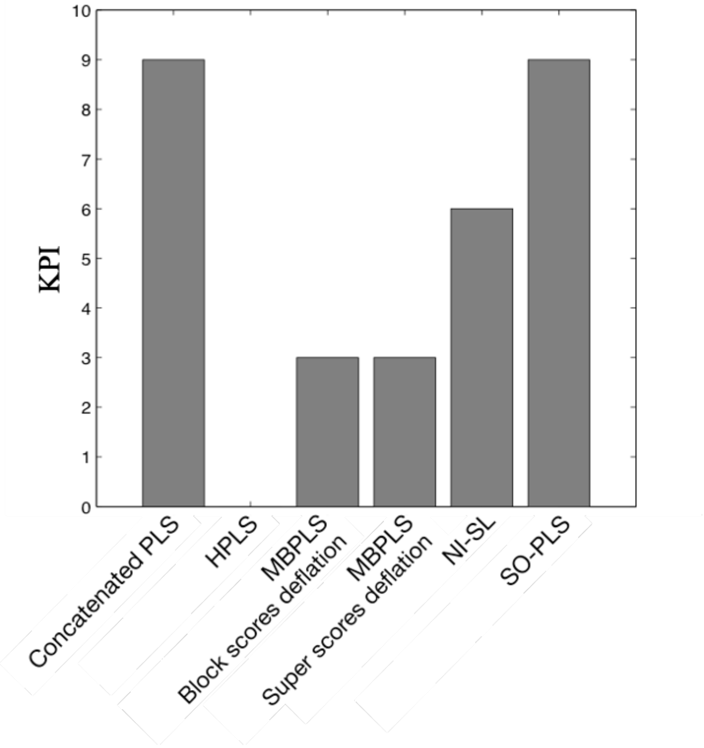


Figure 6-2 - Relative Performance of multiblock methods for predicting wine age.

A summary of the mean $\overline{RMSEP}$ for each multiblock method and the correspondent interquartile range (IQR) characterising its dispersion, is displayed in Table 6-1. To complement this comparison study, the values of the mean coefficient of determination ($\bar{R}^2_{Test}$) obtained from the Monte Carlo approach (50 models) for each analysed multiblock method in predicting the test set were also examined. The values of $\bar{R}^2_{Test}$ provide a good indication of the prediction accuracy of the different methods and are presented in Table 6-1 as well as the correspondent interquartile range (IQR). Table 6-1 supports the conclusions obtained from Figure 6-2. The Concatenated method has lower mean $\overline{RMSEP}$ than SO-PLS but slightly higher dispersion, as seen by the IQR. NI-SL has the third lowest mean $\overline{RMSEP}$ but the highest IQR. Both MBPLS deflation methods lead to very similar results in terms of mean $\overline{RMSEP}$ and IQR. HPLS has the lowest mean $\overline{RMSEP}$ and presents low dispersion of the results. In general, all methods perform rather well, exhibiting $\bar{R}^2_{Test}$ close to one, a good indicator of their prediction capabilities.

Table 6-1 – Mean coefficient of determination, $\overline{R}^2_{Test}$, and $\overline{RMSEP}$ obtained from the Monte Carlo approach for the multiblock state-of-the-art methods under study.

Method	$\overline{R}^2_{Test}$	**IQR (75%-25%)**	$\overline{RMSEP}$	**IQR (75%-25%)**
Concatenated PLS	0.98	0.03	0.93	0.61
Hierarchical PLS (HPLS)	0.95	0.03	1.48	0.44
Multiblock PLS (MBPLS) - Block Scores deflation	0.95	0.04	1.36	0.58
Multiblock PLS (MBPLS) - Super Scores deflation	0.96	0.04	1.34	0.54
Network Induced Supervised Learning (NI-SL)	0.97	0.04	1.17	0.85
Sequential Orthogonal-Partial Least Squares (SO-PLS)	0.99	0.03	0.97	0.49

6.3.2 Comparison between multiblock and single block approaches

In this section, the prediction results obtained with the single block approaches described in Rendall et al. (2017) are compared with those obtained with the multiblock approaches. Several single block prediction methods were analysed in Rendall et al. (2017), but the present study will focus only on the linear latent variable methods (viz. PCR and PLS) for a comparison with the prediction results obtained with the multiblock methods studied in this thesis. The rational for this option, is the interest of assessing the added-value of using simultaneously all the data blocks, for which one must adopt a similar modelling framework. Results from the single block methods are summarized in Table 6-2. Examining results in Table 6-2, one can verify that models based on the Polyphenol content and volatile profiles provided more accurate predictions of wine age, with lower test errors and coefficients of determination closer to 1, which are consistent with the conclusions obtained with the multiblock SO-PLS method. Comparing these results with the ones from the multiblock methods presented in Table 6-1, it is observed that SO-PLS, Concatenated PLS and NI-SL gave superior prediction results in terms of mean values of the coefficient of determination ($\overline{R}^2_{Test}$) and root mean square error of prediction $\overline{RMSEP}$. However, it should be mentioned that the best prediction results obtained with single block methods were the tree-based methods, and boosted regression trees for polyphenols, volatile and the organic acid data sets, suggesting a possible presence of a nonlinear relationship between each analytical predictor block and wine age (Rendall et al., 2017). This also suggests that the best performing multiblock methods discussed here could be improved if non-linear modelling elements are incorporated.

Table 6-2 – Mean coefficient of determination, $\overline{R}^2_{Test}$, and $\overline{RMSEP}$ results obtained for the prediction of wine age based on Monte Carlo approach using the single block approaches PCR and PLS.

Chemical Data	**Method**	$\boldsymbol{\overline{R}^2_{Test}}$	$\boldsymbol{\overline{RMSEP}}$
Polyphenol Content	PCR	0.95	1.18
	PLS	0.94	1.17
Volatile Composition	PCR	0.90	1.55
	PLS	0.91	1.43
UV-Vis	PCR	0.77	2.23
	PLS	0.78	2.86
Organic Acids	PCR	0.64	2.93
	PLS	0.65	2.86

6.4 Results - Interpretation features of multiblock methods

In comparison to the single block approaches, multiblock methods present additional interpretation features to be investigated in order to extract more insights from data (e.g. inter-relations between data sources). This important aspect is explored in the following sections.

6.4.1 MBPLS interpretation capabilities

In this section, the interpretability of the MBPLS methods is explored and both implementations of MBPLS, block score deflation and super score deflation, are considered and discussed. The super scores were calculated exactly in the same way for both methods but, as expected, results differ after the first latent variable due to the different deflation procedures. Figure 6-3 shows a comparison between the variance captured by each of the analytical data blocks used to predict wine age, namely: polyphenols, volatiles, UV-Vis spectra and organic acids content. These are the results obtained from the 50 Monte Carlo runs implemented with the comparison framework, with block scores and super scores deflation methods and using an average of 3.5 and 3.7 latent variables in each method, respectively.

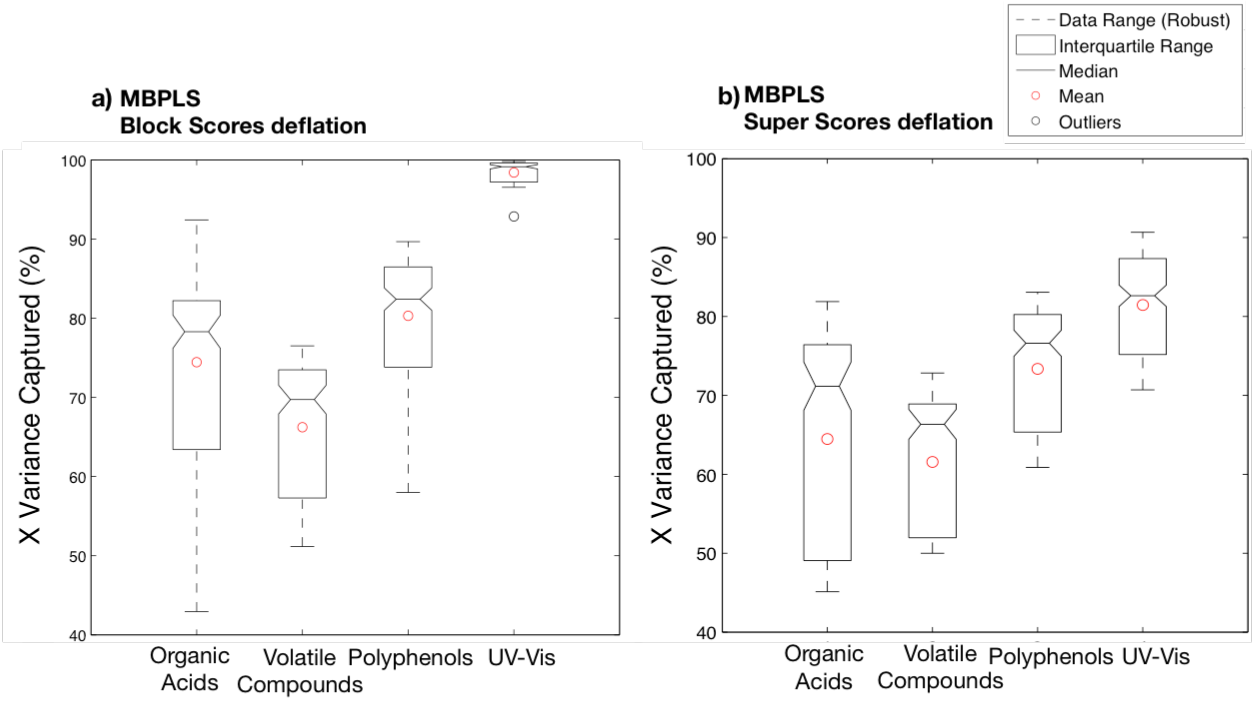


Figure 6-3 - Comparison of the **X**-variance captured (%) by each analytical data block in the two different MBPLS deflation methods: a) MBPLS with Block Scores deflation method; a) MBPLS with Super Scores deflation method.

In terms of the variance captured from each analytical block, it is observed from Figure 6-3 that, on average, the block scores deflation method describes more of the descriptor blocks than super scores deflation method, as expected as more variation is deflated for each latent variable. This is more evident for the organic acids data block and UV-Vis data block. The comparison between super scores and block scores deflation results showed in Figure 6-3 (namely the difference between variance captured by each block) indicate that the super scores deflation method is giving more importance to the volatile compounds and polyphenols blocks. Regarding the **Y**-variance captured the MBPLS algorithm with block scores deflation explains 98.2±1.5% (mean ± standard deviation) and the MBPLS algorithm with super scores deflation, 98.5±1.3%. This difference is not significant, and the response is predicted similarly by both methods. This is confirmed by the previous analysis showing that the difference between prediction accuracy of these two deflation methods is not significant in the present case.

6.4.2 NI-SL interpretation capabilities

The NI-SL multiblock method starts by retrieving five latent variables from PLS regressions between **Y** and each block **X**, followed by stepwise regression to select the most relevant variables. Therefore, it is interesting to investigate which latent variables were selected and from which block they arise, (x-axis on Figure 6-4), as well as the order by which they were selected by stepwise regression (y-axis on Figure 6-4**Figure *6-4***). This information is shown in Figure 6-4, where the size and colouring of each mark represents the frequency (%) by which each latent variable was selected in each stage of the stepwise regression method. This method selects an average of 9±1.8 variables. The analysis of Figure 6-4 reveals that all volatile compounds latent variables were selected with very high frequency and the polyphenols content block is selected with the second highest frequency. This suggests that volatile compounds and polyphenols content blocks carry very relevant information for the prediction of wine age. The organic acids and UV-Vis data blocks are, on average, contributing with one latent variable each. Regarding the order of the variables selected, we will focus on the most relevant variables selected during the first four stepwise regression iterations. It is observed that in the first step of the stewise regression, the algorithm selects the first latent variable (LV1) of the organic acids data block in 60% of the models followed by the LV1 of volatile compounds in 25% of the models developed. In the second step the method selects the LV1 of the UV-Vis data block in 50% of the models and the second latent valriable (LV2) of volatile compounds in 25% of the models. The third latent variable (LV3) of volatile compounds is selected in the third step in 65% of the models and finally during the fourth step the fourth latent variable (LV4) of polyphenols content is selected in 35% of the cases and LV4 from the volatile compounds data block in 25% of the cases.

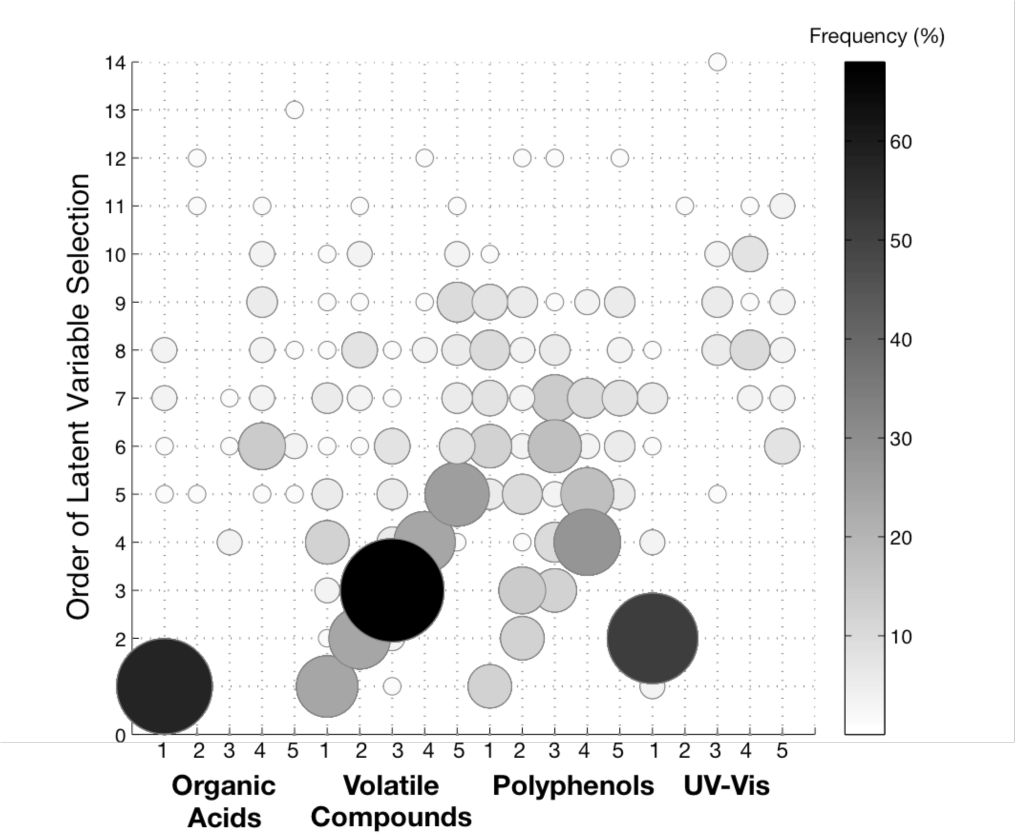


Figure 6-4 - Latent Variables order and frequency of selection by stepwise regression selected of each predictor block.

The analysis of the loadings for the latent variables selected with higher frequency in the models supports the interpretation of the results. The most important variables in the LV1 of the organic acids data set is malic, formic and lactic acids (see Figure 6-5). The malic acid is one of the organic acids already present in grapes while the remaining two appear during the winemaking process and increase their concentration during the ageing process (Rudnitskaya et al., 2010), namely when wines undergone higher temperatures during the ageing process (Pereira et al., 2010b). This is exactly what happens in the Madeira wine ageing process, during which wines in oak casks are stored in warmed lofts, with temperatures ranging between 30 and 35 °C. The analysis of the LV3 for the volatile compounds shows that there are several analytes playing an important role in prediction (see Figure 6-6). These compounds do not belong to the same chemical families (for example, aldehydes, higher alcohols, ethyl esters or other). This fact highlights the importance of performing the screening of volatile profile rather than a targeted analysis that is directed towards a specific family of volatiles compounds. The loadings of LV4 concerning the polyphenols content indicate the additional importance of mainly Gallic acid and Ellagic acid (probably having a great influence to distinct aged wines), trans-Resveratrol and Quercetin (with a more prominent role for young wines) for wine age prediction (see Figure 6-7). Finally, the loadings of LV1 of the UV–Vis spectra data revealed that the ultraviolet region is the most important (see Figure 6-8).

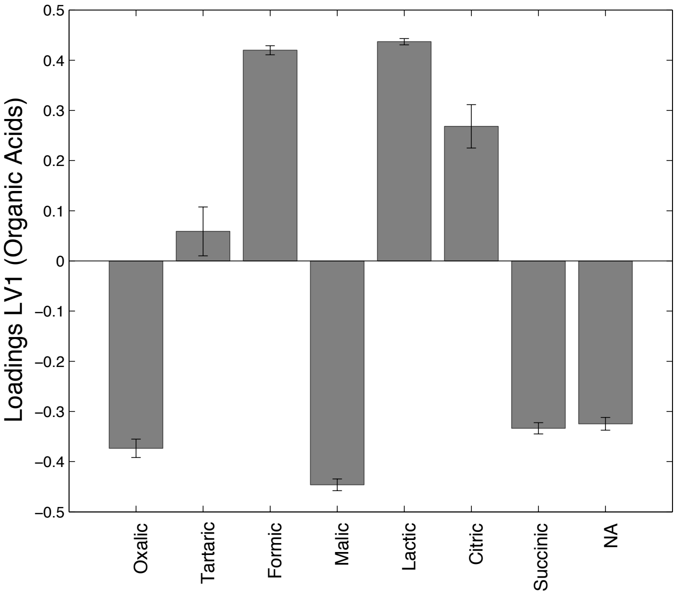


Figure 6-5 - Loadings of the first latent variable (LV1) for the Organic Acids block data.

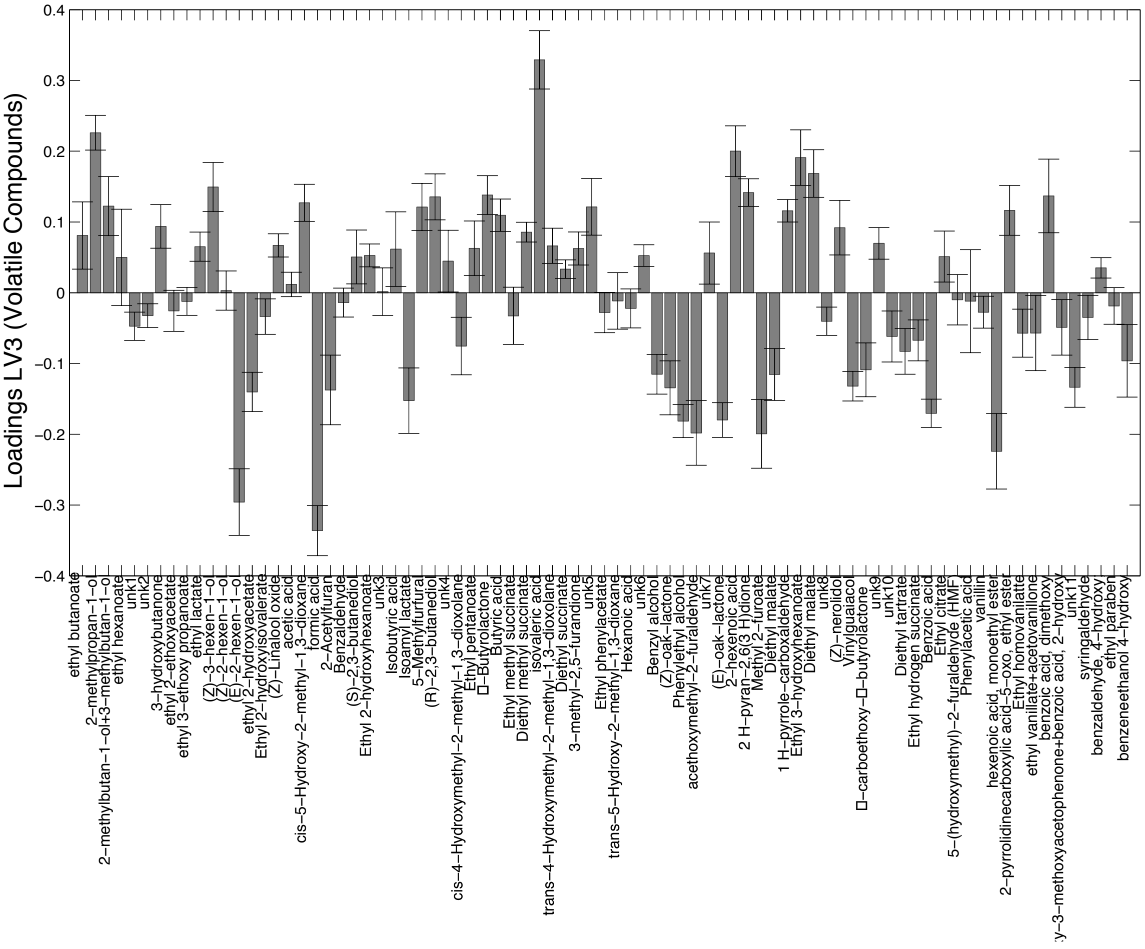


Figure 6-6 - Loadings of the third latent variable (LV3) for the volatile compounds data.

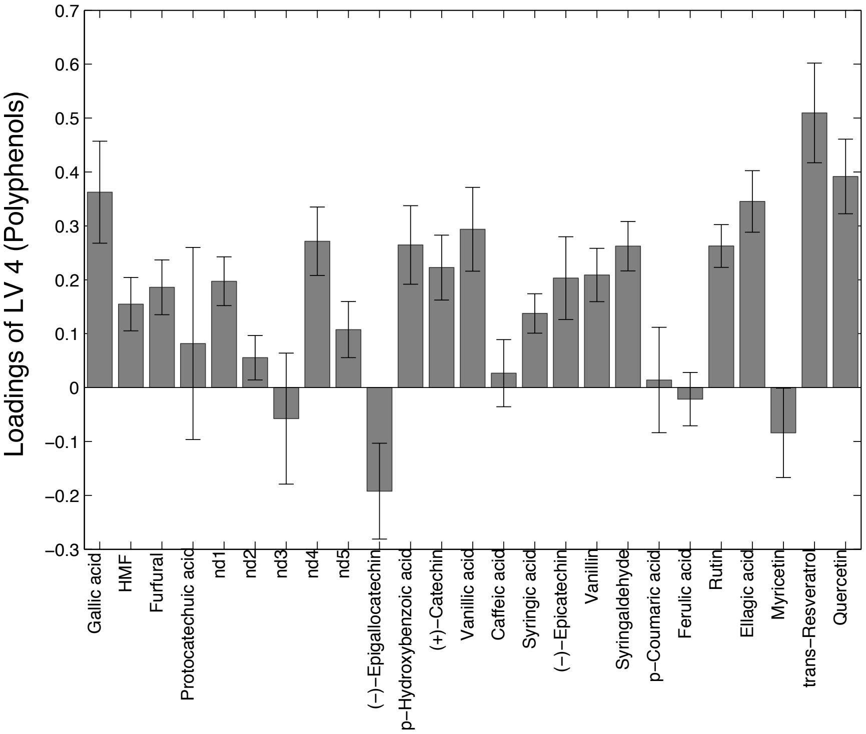


Figure 6-7 - Loadings of the fourth latent variable (LV4) for the polyphenols content data.

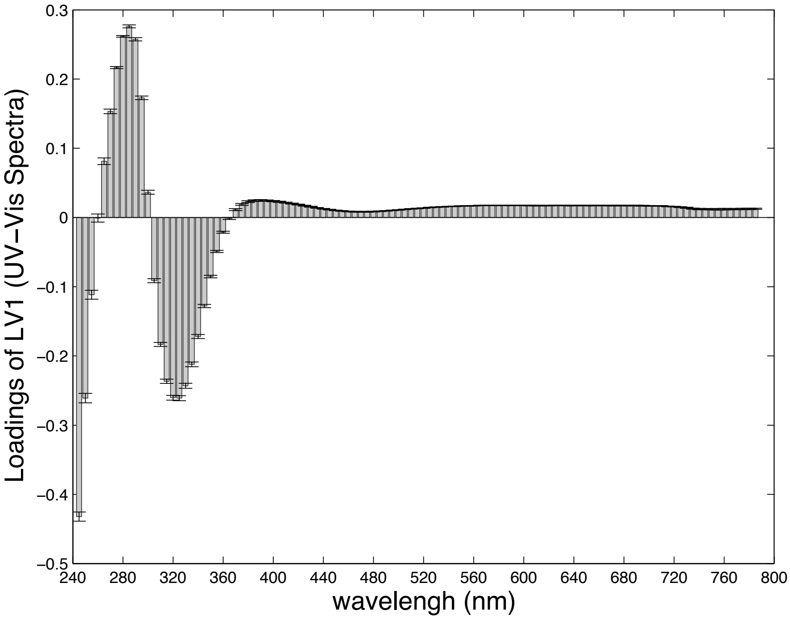


Figure 6-8 - Loadings of the first latent variable (LV1) for the UV-Vis data block.

In order to better interpret the results showed in Figure 6-4, the analysis of the correlation coefficients between each latent variable and wine age was also carried out (see Figure 6-9). In fact, all first latent variables are highly correlated with wine age (correlation coefficient > 0.85) and the organic acids LV1 has a slightly higher correlation, which explains why this was the first variable to be selected in stepwise regression. Moreover, organic acids LV1 is also correlated with the volatile compounds LV1 and polyphenols content LV1 but not as much with the UV–Vis LV1 as seen in Table 6-3. This suggests that this latent variable might have additional information for the prediction of wine age, which is consistent with the fact that it is being selected in the second step in some of the models. Regarding the other four latent variables, volatile compounds and polyphenols content have more relevant information that is being captured in these four remaining latent variables, as they exhibit higher correlation coefficients with wine age, especially in comparison with the last four latent variables of the organic acids block. This supports the fact that the remaining four latent variables of polyphenols and volatile compounds are being selected in latter steps of the method.

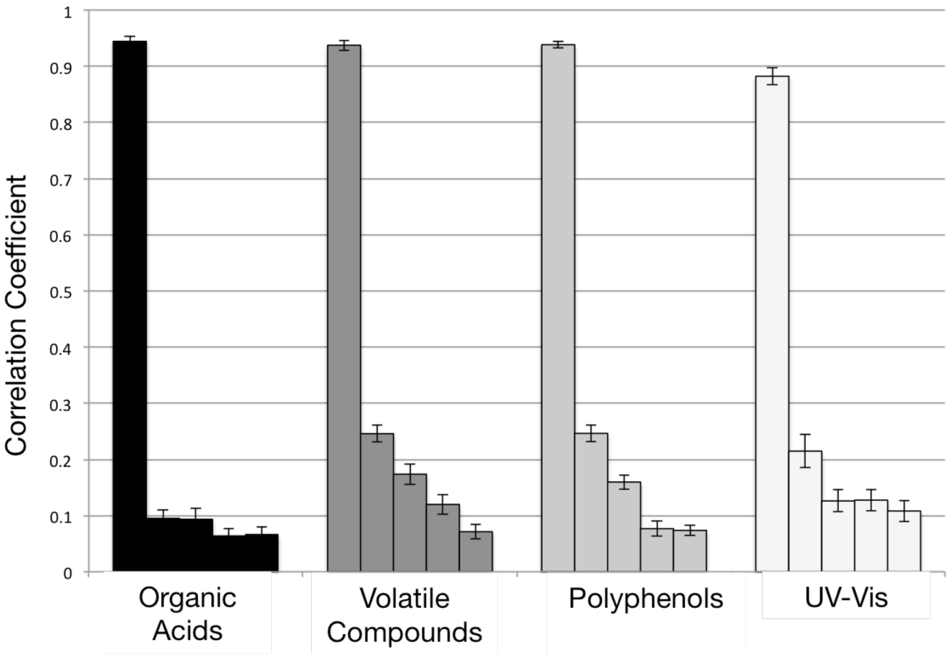


Figure 6-9 - Correlation Coefficients between each latent variable of each predictor block and wine age (5LVs were collected for each predictor block) and wine age.

Table 6-3 – Correlation Coefficient between the first latent variables (LV1) of each predictor block.

	LV1 Organic Acids	**LV1 Volatile Compounds**	**LV1 Polyphenols**	**LV1 UV-Vis**
LV1 Organic Acids	1.00			
LV1 Volatile Compounds	0.93	1.00		
LV1 Polyphenols	0.91	0.91	1.00	
LV1 UV-Vis	0.79	0.81	0.92	1.00

6.4.3 SO-PLS interpretation capabilities

The Sequential Orthogonalized PLS (SO-PLS) model was the method that led to the best results for wine age prediction, as can be verified by the RMSEP obtained and R^2_{Test}. Additionally, this method provides explicit quantitative measures about redundant information in the multiple **X**-blocks, which can be very advantageous when several blocks are available. This capability allows, for example, to assess if one or more blocks are sufficient or if all are really needed to adequately predict the response. As SO-PLS is dependent on the block order, a comparison between all possible block order combinations was performed (see Figure 6-10) to establish the best block ordering for modelling and further comparison with the other state-of-the-art multiblock methodologies.

The block orders that lead to the lowest mean RMSEP are 3-2-4-1 which corresponds to polyphenols content → volatile compounds → UV-Vis → Organic Acids; 3-2-1-4 which corresponds to polyphenols content → volatile compounds → Organic Acids → UV-Vis and finally 4-3-2-1 corresponding to UV-Vis→ polyphenols content → volatile compounds → Organic Acids. All these block orders led to similar prediction performances.

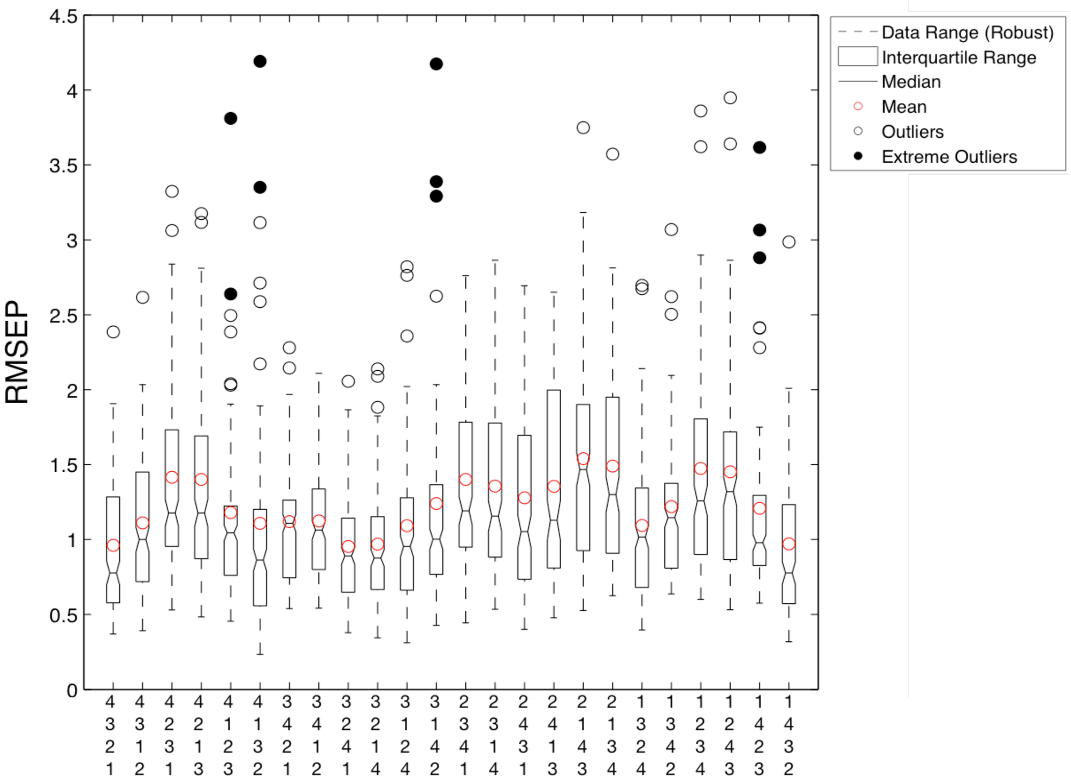


Figure 6-10 - SO-PLS results in terms of Root Mean Square Error of prediction (RMSEP) obtained from models using different block orders (24 permutations were performed).

In order to fully explore SO-PLS interpretation capabilities the model developed with the minimum RMSEP obtained was selected for a deeper analysis of the results. Figure 6-11 shows the results obtained for the best model using blocks in the order 3-2-1-4 (RMSEP=0.34 years). In this model, 4 LVs were used for polyphenols content and volatile compounds blocks and 2 LVs were used for Organic Acids data block and UV-Vis. Polyphenols content and volatile compounds are statistically significant to the model; both data blocks describe 99.7% of wine age. All analytical data blocks have significant amounts of redundant information. For instance, polyphenols content shares more than 50% of predictive information with the other data blocks for wine age prediction. In this Figure 6-11 one can also see that 85.4% of the variance captured by polyphenols content explains 97.6% of the information in the wine age block. All blocks are orthogonalized by the first block (Polyphenols) which means that the variability already explained by this first block is removed from the other blocks representing 58.8% of the variance in the second block (volatile compounds block), 53.6% of the third block (organic acids) and 84.8% of the UV-Vis block.

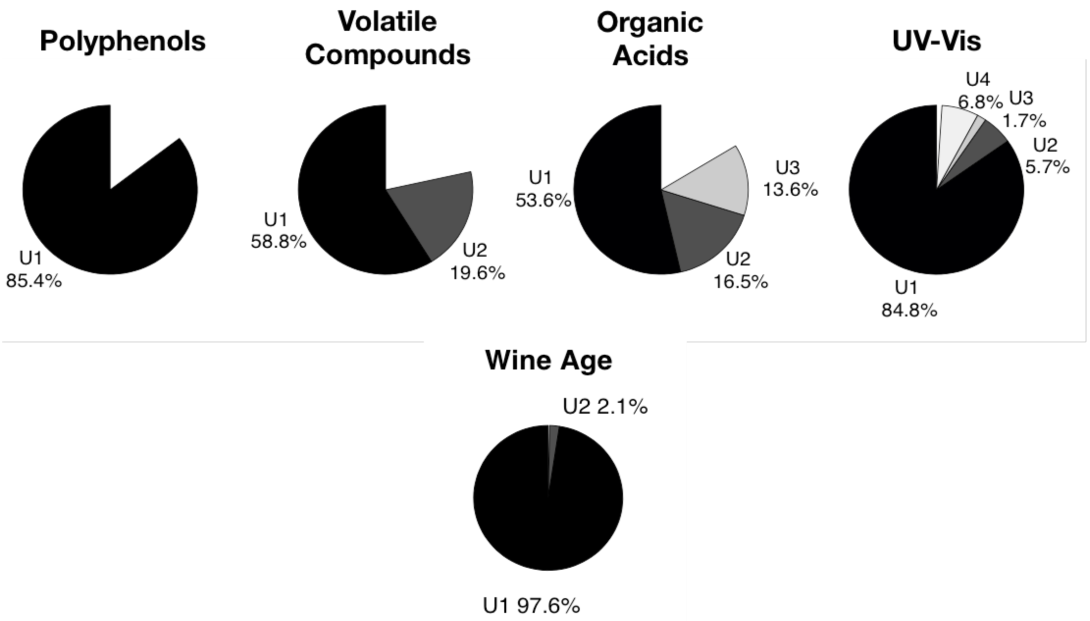


Figure 6-11 – X Blocks Variance Captured (%) and Y Variance captured for the SO-PLS model that gave the best result in terms of RMSEP.

According to the ANOVA table displayed in Table 6-4, only polyphenols content (block $\mathbf{X}_1$) and volatile compounds (block $\mathbf{X}_2$) are statistically significant (p-value<0.05). Which means that the other two blocks could be removed to make the model more robust. Nevertheless, SO-PLS was analysed with all block to make it more comparable with the other multiblock techniques.

Table 6-4 - SO-PLS model Results.

Source	Exp Var (%)	RMSEC	Exp Var (%)	RMSECV	p-value
Polyphenol Content	97.6	0.892	96.1	1.140	1.3x10^{-10}
Volatile Composition	99.8	0.274	97.9	0.833	0.02

Source	Exp Var (%)	RMSEC	Exp Var (%)	RMSECV	p-value
UV-Vis	99.8	0.249	98.0	0.821	0.12
Organic Acids	99.8	0.234	98.1	0.801	0.97

6.5 Chapter final remarks

In this chapter, the capabilities of five state-of-the-art multiblock methods for predicting the ageing time of the fine Portuguese Madeira wine were assessed, using the chemical information of polyphenol content, volatile composition, UV-Vis spectra and organic acids. The methods considered were: Concatenated PLS method, MBPLS, HPLS, NI-SL and SO-PLS.

The prediction relative performance of the state-of-the-art methods was evaluated by means of a robust statistical framework that implements a Monte Carlo double cross-validation scheme. The best results were obtained using the SO-PLS and the Concatenated PLS method, both showing equivalent relative performances, followed by NI-SL. SO-PLS method was optimized with respect to the blocks order so that the information is sequentially extracted from most relevant blocks and only additional information is incorporated from all blocks. If there is no prior knowledge to select the block order this can be challenging, in particular when the number of blocks is high (as all possible block sequences need to be evaluated). PLS Concatenated method was also optimized with respect to the block scaling method to be applied and results showed that weighting each data block by their pseudorank (number of principal components) leads to the best relative performance. How the different blocks should be pre-processed is also a challenging task and needs to be evaluated case-by-case thus, when no a priori knowledge exists this is not a straightforward task and can be time consuming (see Chapter 5).

A comparison between the best single block linear latent variable methods (viz. PLS and PCR) and multiblock methods in terms of prediction capability of wine age using the coefficient of determination and RMSEP, shows that the multiblock methods SO-PLS, Concatenated PLS and NI-SL lead to superior prediction results. However, results obtained in (Rendall et al., 2017) with the same data sets suggest that non-linear methods lead to even better prediction results suggesting the possible existence of a non-linear relationship between the predictor blocks and response block. This is an aspect to be explored in the scope of multiblock methods in the future.

The interpretational capabilities of the multiblock methods were also assessed and critically analysed. These methods are able to deal with large volumes of data and to include structure in the analysis. This makes the analysis more consistent with existing priori knowledge, leading to more informative insights in the end. All methods studied in this thesis have different particularities and bring different interpretational features into the analysis and should be selected depending on the data sets at hand, objective of the analysis, and also the experience of the user with the methods.

The analysis of the loadings of the (more frequently) selected latent variables for the NI-SL approach provide useful insights into the product and system under study. Moreover, this method provides the added-value that allows the identification of the different blocks when this is not known *a priori* (this feature was not explored here as the blocks were clearly defined).

Among the multiblock methods studied in this work SO-PLS has the particularity of providing an estimate of the contribution of each additional block to the **X-** and **Y-** variations explained by the model. The model interpretation of SO-PLS revealed that all four analytical data blocks have a significant amount of redundant information and that the polyphenols and volatile compounds data sources alone would explain more than 99% of the variability of the response block.

Chapter 7 – Stepwise sequential orthogonalized partial least squares

Chapter 7. Stepwise sequential orthogonalized partial least squares

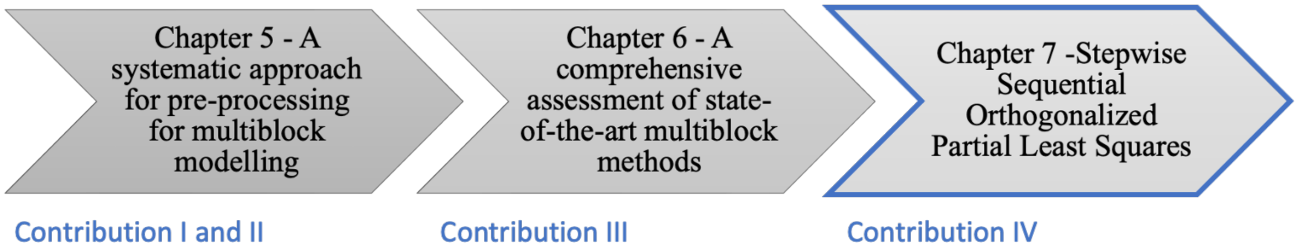

Chapter 5 - A systematic approach for pre-processing for multiblock modelling
Chapter 6 - A comprehensive assessment of state-of-the-art multiblock methods
Chapter 7 -Stepwise Sequential Orthogonalized Partial Least Squares
Contribution I and II
Contribution III
Contribution IV

Figure 7-1 - Conceptual flow chart of the thesis organization. Chapter 7 presents a new methodology called stepwise sequential orthogonalized PLS.

7.1 Introduction

Among the multiblock algorithms available in the technical literature, the Sequential Orthogonalized Partial Least Squares (SO-PLS) method proposed by Naes et al. (2013) and Jørgensen et al. (2007) has been attracting interest due to several distinctive and opportune features that address common concerns of practitioners.

One advantage of this method is its reduced sensitivity to the relative scales of the blocks, which is an interesting feature if the data blocks have widely different measurement units (this is because blocks are incorporated one at a time and orthogonalized). Another advantage is the ability to handle situations with different underlying dimensionality (pseudo-ranks) of the blocks. This makes SO-PLS rather flexible and suitable for situations where one of the blocks is, for instance, a design matrix originated from an orthogonal design of experiments (independent variables) while the other blocks consist of multivariate and highly collinear data (e.g. spectroscopic data) (Naes et al., 2013, Jørgensen et al., 2007). These advantages coincide with some of the main drawbacks of current state-of-the-art multiblock algorithms (Campos et al., 2017). Finally, and perhaps most importantly, an additional benefit of SO-PLS is associated with its interpretation capabilities, namely regarding the analysis of the relationship between blocks (degree of overlapping with respect to the prediction of the response) and the incremental contribution of each block to the improvement of prediction ability. SO-PLS provides quantitative measures of the additional contribution of each predictor block to the model, and of their overlap.

The implementation of SO-PLS (after an appropriate selection of the block order) is also not as challenging as some current multiblock methodologies, where the scaling and deflation operation are not easy to follow and reproduce, leading to the existence of several variants whose relative performance for each case cannot be anticipated. In addition to interpretation capabilities, the SO-PLS method (after an appropriate selection of the block order) also leads to good prediction performances when compared to the current state-of-the-art multiblock algorithms (Campos et al., 2017).

Examples of applications of SO-PLS for prediction purposes can be found in Naes et al. (2013), Jørgensen et al. (2007), Hertrampf et al. (2016) . In Hertrampf et al. (2016), SO-PLS is used to predict the end product release attributes of solid dosage forms based on NIR and Raman information from raw materials. This paper concludes that the use of complementary information from different analytical sources by means of SO-PLS can be of benefit in a lifecycle analysis perspective as the additional contribution of the second block will improve the model performance by handling, for example, changes in the supplier or lot-to-lot variability. In Jørgensen et al. (2007) an industrial application of the method is described. In this paper, SO-PLS offers a convenient solution for combining raw material information captured by spectroscopic methods with spectroscopic measurements taken at a later point in the process together with experimental design data on process variables, for modelling the end product quality. The applications of SO-PLS have also been extended to situations where classification is the main purpose and also to multiway arrays (Biancolillo et al., 2015).

Due to the sequential nature of SO-PLS and its successive orthogonalization steps, it is critically dependent on the particular order by which the blocks are incorporated into the model. The order of the blocks can be naturally defined by the causal linking between the blocks of variables, such as in continuous production units or in multistage batch processes. However, if the causal connectivity between blocks is unknown or does not even exist, the lack of any rational to guide de selection of the appropriate blocks' order constitutes a major drawback of this method. There are indeed cases where no order can be *a priori* postulated as being superior, given the symmetric role of the measurements available, as will be illustrated in the case study addressed. Another common situation is when the blocks refer to parallel processing plant streams in an industrial plant. The problem becomes even more complex when the number of blocks involved increases (≥3), as the number of putative orders to explore quickly grows in a combinatorial progression. As these conditions tend to be frequent in Industry 4.0 settings given the growth in the amount and variety of available measurement sources and the enlargement of the analysis scope to the entire value chain (characteristics constituting distinctive traces of the new industrial paradigm), effective solutions need to be found to address this problem in order to benefit from the good features of SO-PLS. More details on the background of SO-PLS and respective algorithm are given in Section 2.3.

Therefore, in this chapter a new variant of SO-PLS is presented, called Stepwise SO-PLS, that not only solves the order selection problem, but brings forward additional analytical features, such as the capability of blocks selection – either because they do not contain relevant predictive information for the response, i.e., the data set is sparse with regards to prediction, or because the additional contribution is not relevant given the information extracted from the precedent blocks, i.e., the natural evolution of variable selection methods, to multiblock contexts. Furthermore, the relative importance of the blocks' can also be appreciated, analysing the sequence by which they are selected by the algorithm. The proposed method implements an efficient stepwise approach for selecting the best ordering of data blocks. In brief terms, it starts by selecting the first data block, which is the one leading to the minimum root mean square error of cross-validation (RMSECV). Then, it evaluates the effect on RMSECV of incorporating in the model one block at a time, selecting at each step the block most contributing to improve the prediction ability (by decreasing the overall RMSECV). The process stops when no further improvement is verified in terms of prediction capability or no more blocks remain to be selected. This is a very important feature in Big Data multiblock applications, which is not shared by other current approaches: the capability of excluding blocks that are not relevant for the prediction of the response under analysis leads to more robust, parsimonious and interpretable models, which are critical requirements for most end users. Through the selective incorporation of informative blocks in the model, the method can simultaneously improve its prediction accuracy and interpretation insights.

A new variant of the combinatorial approach for establishing the order of the blocks for performing SO-PLS is also proposed here to allow for the exclusion of non-relevant blocks. This presents an advantage over the current state-of-the-art methodology but is still time and effort consuming.

To provide the context for presenting and illustrating the proposed methodologies, a real data set is adopted, reflecting the complexity of real processes and the multitude of variability sources likely to be the found in real working environments. The case study consists of the use of several analytical measurement sources in order to develop a predictive model for the ageing time of one of the finest Portuguese fortified wines: the Madeira wine. Please refer to Chapter 4 for more details on the case study. This is an example of a real situation where no particular block order can be postulated *a priori;* situations like this are rather common in analytical laboratories in both academia and industry, and therefore relevant to consider, as they have associated important issues regarding the complex balance between the quality of information generated in empirical studies and the cost of acquiring it. Furthermore, it is also relevant to analyse which measurements most contribute to improve the prediction ability of the model and whether all measurement blocks are really necessary or if some of them can be discarded, reducing the costs of implementation of the soft sensor in the future.

The new stepwise SO-PLS method is compared with the current benchmark approach for establishing the block order for SO-PLS (combinatorial approach) and the new proposed variant that allows for blocks exclusion. The comparison is made on the basis of the prediction accuracy achieved using a robust Monte Carlo Cross-validation framework and the computational effort required by each method.

This chapter is organized as follows: Section 7.2 describes systematic approaches for establishing the order of the blocks in SO-PLS: the standard approach currently in use based on blocks permutation (Section 7.2.1), a new variant proposed that allows selection of blocks (Section 7.2.2) and the new proposed stepwise SO-PLS approach (Section 7.2.3). Section 7.3 presents the application of stepwise SO-PLS and benchmark methods to a real case study and the discussion of the results obtained. The final remarks of this work are summarized in Section 7.4.

7.2 Methods

The outputs of SO-PLS are critically dependent on the order used to process the predictor blocks. This will have an influence on the prediction capability of the method and in the interpretational outcomes it provides. The impact of the blocks' order is more significant in the case where several blocks are available and not all of them contribute with additional valuable information to predict the response. When several blocks are available, it is important to determine which blocks contribute the most for explaining the response variability and carry complementary information, so that predictive insights can be sequentially extracted starting from the information rich blocks toward the information poor and noisy blocks. Therefore, efficient systematic approaches are required for the selection of the best order of the blocks when implementing SO-PLS. In the next subsections systematic approaches for the selection of the order of the blocks for SO-PLS are presented including the new methodologies.

7.2.1 SO-PLS with blocks permutation

The easiest way to select the blocks order occurs when *a priori* knowledge is available (e.g., the natural sequence of stages in a batch process or the sequence of equipment crossed by the product stream, or the unit operations applied to the product). When this is not the case, the current solution is to test and evaluate all possible combinations of sequences for ordering the blocks (a total of $B!$ for B blocks) and then determine the best ordering by looking to robust metrics of prediction ability. The models developed with all possible blocks permutations can also be used as an additional interpretation tool, using the adequate visualization tools.

This combinatorial approach does not raise any relevant problem when only a few number of predictor blocks are available. However, it can be a very time consuming task when a large number of blocks need to be analysed, a situation that is likely to become common in Industry 4.0 chain-wide applications – see Figure 7-2.

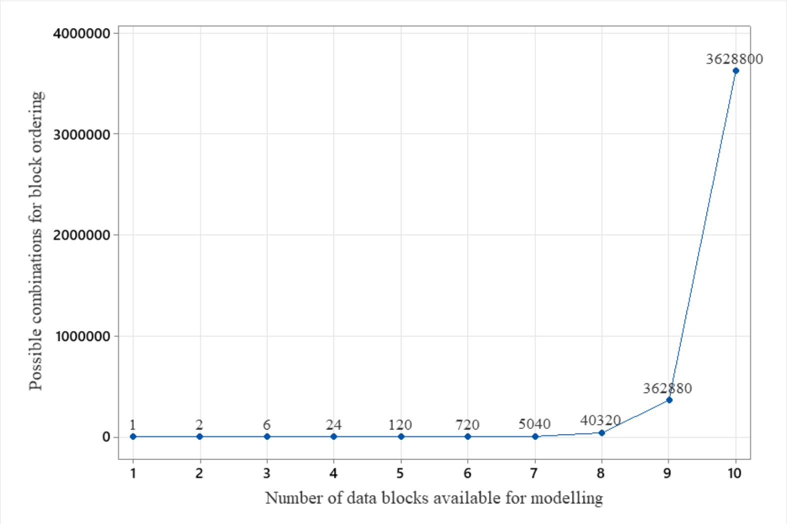


Figure 7-2 – Possible combinations for block ordering with increasing number of data blocks available for the analysis

For each possible combination a standard SO-PLS method is developed by testing all latent variables from 1 to the maximum predefined number of latent variables (*LVmax*). In this case five maximum latent variables were tested.

The pseudocode for this method is presented in Table 7-1. The prediction accuracy of the models developed by performing all possible combinations is evaluated by the root mean square error of prediction (RMSEP) obtained.

Table 7-1 – Pseudocode for SO-PLS with Permutations

```
REQUIRED: $\{\mathbf{X}_1,\mathbf{X}_2,\ldots\ \mathbf{X}_b,\ldots,\mathbf{X}_B\}$, Y
I. FOR: $i$=1:B! (For each permutation of the blocks)
    (1) Perform SO-PLS for permutation $i$ following the procedure in Table 2-5
    (2) Compute and save $median$(RMSEP) for $i$
II. Find permutation $i$: Min $\{RMSEP_i\}$ and extract the corresponding ordering of the blocks
```

7.2.2 SO-PLS with blocks permutation and allowing blocks selection

A new variant of the block's permutations approach was also explored in this work, that allow for the possibility of excluding less meaningful blocks among the available ones.

Similarly to the standard block permutations approach, all possible combinations of sequences for ordering the blocks (a total of *B*! for *B* blocks) are evaluated. For each possible combination a standard SO-PLS method is developed by testing all latent variables up to the maximum predefined number of latent variables (*LVmax*) - five maximum latent variables were tested for this approach. In order to be able to exclude blocks in this algorithm the possibility of testing "0 latent variables" from each block is also included, meaning that the block is effectively discarded. Therefore, this evolution in the standard approach implies some additional computational load but can lead to significant improvements in the prediction and interpretation outcomes of the final model (more parsimonious models).

7.2.3 Stepwise SO-PLS

The new algorithm proposed here is computationally faster and works efficiently even when a large number of blocks are to be included in the model. Since it is an extension of the original SO-PLS, it inherits the fundamental attributes of SO-PLS. In this algorithm, the order by which each block is added to the model is selected based on the analysis of the root mean square error of cross-validation (RMSECV), using the leave-one-out method (LOOCV). The main idea behind this procedure is to make sure that the information is extracted from the more informative blocks to the less informative ones, towards a minimum RMSECV solution. Other alternative cross-validation approaches could have been used instead of LOOCV, but its simplicity, widespread use and the fact that only one value is obtained in the end (contrary to K-fold cross validation, whose output depends on the partitioning), are arguments to use it for illustrating the mechanics of stepwise SO-PLS. However, one must also realize its limitations, especially for large data sets. At each step of the method, a block is only added if the estimated error decreases. The final solution is reached when the RMSECV stops decreasing or no more predictor blocks are left to be incorporated in the model.

Even though the implementation of a forward addition protocol, such as the one proposed, does not secure, in general, the achievement of the optimal solution, this is quite often the case. In the other cases, the solution achieved is nearly optimal, and obtained very rapidly. The successive addition of blocks in this protocol is also, by itself, an interesting source of information of Stepwise SO-PLS, further justifying the use of this approach, besides its high computational efficiency. The pseudocode for Stepwise SO-PLS is displayed in Table 7-2 and a schematic representation of the algorithm is showed in Figure 7-3.

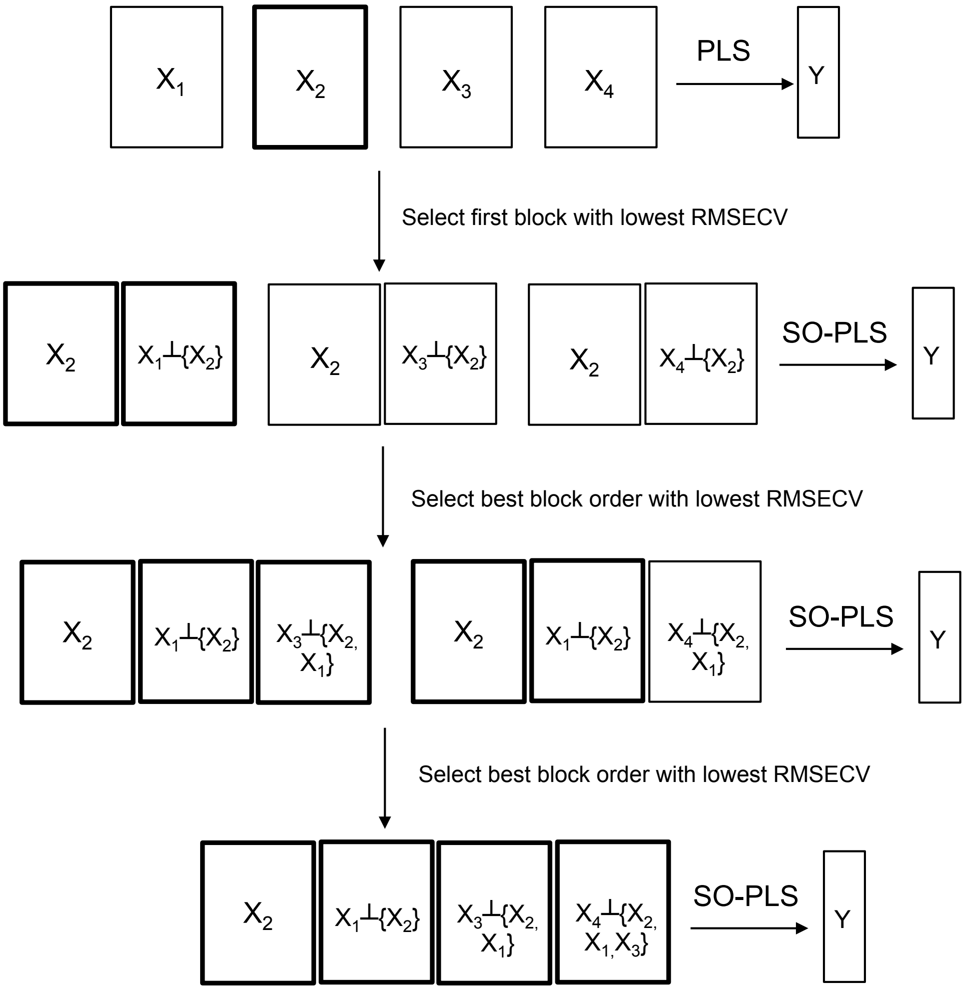


Figure 7-3 – Schematic representation of the Stepwise SO-PLS method. In the first iteration block X_2 is selected for the first position due to presenting lower RMSECV. All remaining blocks X_1, X_3 and X_4 are orthogonalized in respect to X_2. In the next step SO-PLS models are built using the first selected block and each one of the remaining orthogonalized blocks. At each iteration step blocks with relevant new information are sequentially added to the model and the best block order leading to the lowest RMSECV is determined. In this illustrative example the best block order is: X_2-X_1-X_3-X_4.

Table 7-2 – Pseudocode for Stepwise SO-PLS.

```
REQUIRED: $\mathbf{X}\equiv\{\mathbf{X}_1,\mathbf{X}_2,\ldots \mathbf{X}_b,\ldots,\mathbf{X}_B\}$, Y
SET:
  $BS = \emptyset$                                  % Set of selected X-blocks
  $NS = \mathbf{X}$                                  % Set of X-blocks left out
I – FOR $b$=1:#$NS$
Perform SO-PLS between each individual block in NS together with all blocks from BS and Y (if BS is empty, apply PLS)
(3)  Calculate and save RMSECV$_b$
II –Select the block minimizing {RMSECV$_b$}$_{b=1:\#NS}$: $\mathbf{X}_{b^*}$
IF RMSECV improves:
    REMOVE $\mathbf{X}_{b^*}$ from $NS$
    ADD $\mathbf{X}_{b^*}$ to $BS$
    GO TO I and keep the order of selected blocks
ELSE:
    STOP
```

Note: #NS represents the cardinality of the set of blocks left out, NS.

7.3 Results

In this section, we illustrate the application of Stepwise SO-PLS in a real case study, and compare the results obtained with the current benchmark that consists of using the computationally more expensive block permutation approach described in Section 7.2.1, as well as its new alternative presented in Section 7.2.2 that allows for blocks exclusion. The case study regards the problem of predicting the ageing time of Madeira wine, using for such several types of measurements arising from different analytical sources. In this situation, as in other scenarios where multiple measurement sources are available (such as in the domain of industrial soft sensors), no a priori ordering can be established. Furthermore, it is also relevant to analyse which measurements most contribute to improve the prediction ability of the model and whether all measurement blocks are really necessary or if some of them can be discarded, reducing the costs of implementation of the soft sensor in the future.

The comparison between the new proposed Stepwise SO-PLS and the standard blocks permutation approach for establishing the order of the blocks takes into consideration the prediction capability of the methods and the computational effort necessary to implement them. In terms of interpretation capabilities, once the best order for the blocks can be found by all methods under this study, no difference between the approaches is expected. However, the order by which the methods are selected in the stepwise SO-PLS can also provide further insights for the user, namely regarding the relative importance of the blocks under analysis.

The methods' computational effort (number of estimated models) and respective implementation time are correlated with each other. Computational effort for the different approaches is here assessed in terms of the major computational steps necessary to implement the method. Table 7-3 presents the number of models expected to be explored during the implementation of the different systematic approaches. Even through in the present case study the computational complexity might not be a critical factor, it becomes more relevant when more blocks are available for analysis.

The definition of the number of estimated models takes into consideration the number of blocks available for the analysis and the number of latent variables allowed for modelling.

The standard SO-PLS method is developed by testing all latent variables from 1 to the maximum predefined number of latent variables (*LVmax*). In the present case, since 4 predictor blocks are available and 5 latent variables were used as the maximum number of latent variables to be chosen by the successive PLS models (*LVmax*=5), a single ordering of the SO-PLS model requires the computation of 5^4=625 alternatives. The current systematic approach for the selection of the best block order contemplates all possible permutations of the 4 blocks (i.e., 4! permutations) followed by the application of SO-PLS to each possible combination.

In case one also wants to include the possibility of excluding less informative blocks, extra possibilities should be allowed in the modelling. This was implemented by allowing the model to choose "0 latent variables" meaning that the block would be excluded – in comparison with standard approach this would mean one additional latent variable to be included in the modelling. In spite of the additional computational burden, the capability to exclude blocks from the model is very interesting from both the predictive and interpretational perspectives: on one hand more parsimonious models can be constructed, which are more robust and therefore more stable and accurate; on the other hand, simpler models are easier to interpret and the removal of certain blocks from the model has an obvious meaning regarding their relative relevance. This is a feature not shared by other current multiblock approaches, and therefore it was also considered here, leading to a new variant of the standard SO-PLS approach.

In the case of stepwise SO-PLS, the sequence of blocks is selected in a stepwise manner: in the first step all blocks can be selected and all latent variables are tested from 1 to *LVmax*; in the second step only the remaining 3 blocks can be selected and so on and so forth, until all blocks have been included in the model and the final order of the blocks defined, or until the process terminates (RMSECV increases). Examining the number of required modelling steps, it can be observed that stepwise SO-PLS is of the same order of magnitude of a single run of the benchmark approach. Therefore, it is much more efficient, because with a considerable smaller computational effort compared with the blocks permutation approach and its new variant, the stepwise method is able to select the best, or close to best order of the blocks, leading to a better solution in terms of prediction and interpretation. Please note that the stepwise SO-PLS can be even more efficient than what is referred in Table 7-3, as the algorithm can terminate if the RMSECV stops decreasing before reaching the final block.

Table 7-3 – Comparison of the number of estimated models in the implementation of each method

Method	**Number of Steps (Estimated Models)**
SO-PLS with blocks permutations (benchmark)	$4! \times 5^4 = 15\ 000$ (1)
SO-PLS with blocks permutations and exclusion capabilities (new variant of the standard approach)	$4! \times 6^4 = 31\ 104$ (1)
Stepwise SO-PLS	$5 \times 4 + 5^2 \times 3 + 5^3 \times 2 + 5^4 = 970$ (2)

Legend:
(1) #Blocks! × #LVs#Blocks
(2) Maximum number of modelling steps. The actual number may be less than 970 if the algorithm terminates before reaching the final block (in case the RMSECV stops decreasing).

In terms of prediction accuracy, the distribution of the RMSEP obtained for the 50 models developed in the Monte Carlo cross-validation procedure with stepwise SO-PLS was compared with the distribution obtained for the SO-PLS with the optimal block order obtained from the benchmark method (blocks permutations) and its new variant (blocks permutations with exclusion capabilities). Figure 7-4 reveals that the prediction performances of the three approaches are equivalent. However, stepwise SO-PLS requires much less computational effort and implementation time. In the present case, all **X**-blocks bring some incremental predictive information to the model, and therefore no difference is observed between the standard approach and its new variant with blocks exclusion capabilities. However, the ability of blocks selection can be very interesting in other situations where blocks are irrelevant or redundant and simpler models would be sufficient to achieve the optimal performance.

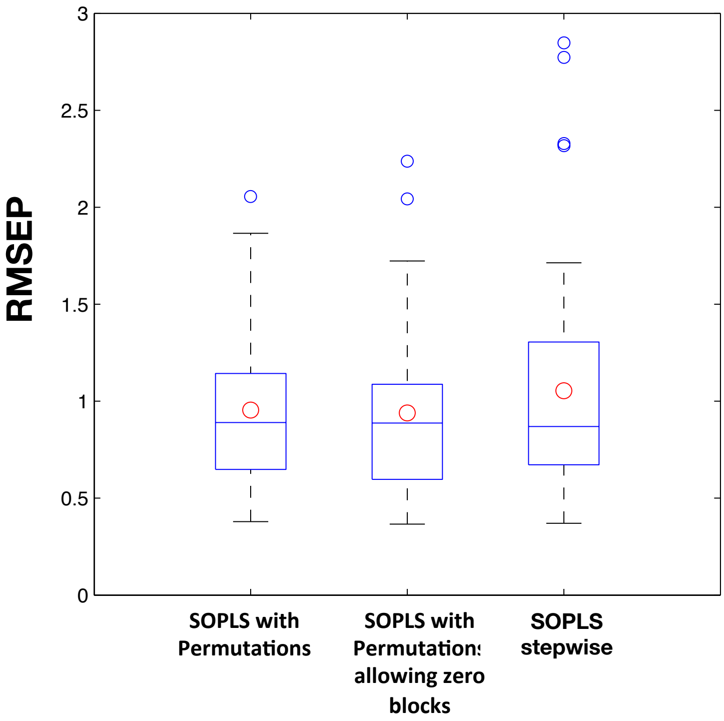


Figure 7-4 – Comparison of the RMSEP distributions obtained for the systematic approaches to select the best order of the blocks.

Figure 7-5**Figure *7-5*** shows a comparison between the RMSEP distributions obtained from the 50 Monte Carlo models developed with each block permutation (24 permutations in total). This figure **Figure *7-5***confirms that the order of the blocks does have an impact on the prediction accuracy of the method. The best order established by stepwise SO-PLS method in the 50 Monte Carlo models developed was: 3-2-1-4. This solution is highlighted in Figure 7-4 showing that the stepwise method is indeed performing well, leading to an optimum solution in terms of prediction accuracy, as this is also the best solution obtained with all possible block combinations (from all 24 permutations).

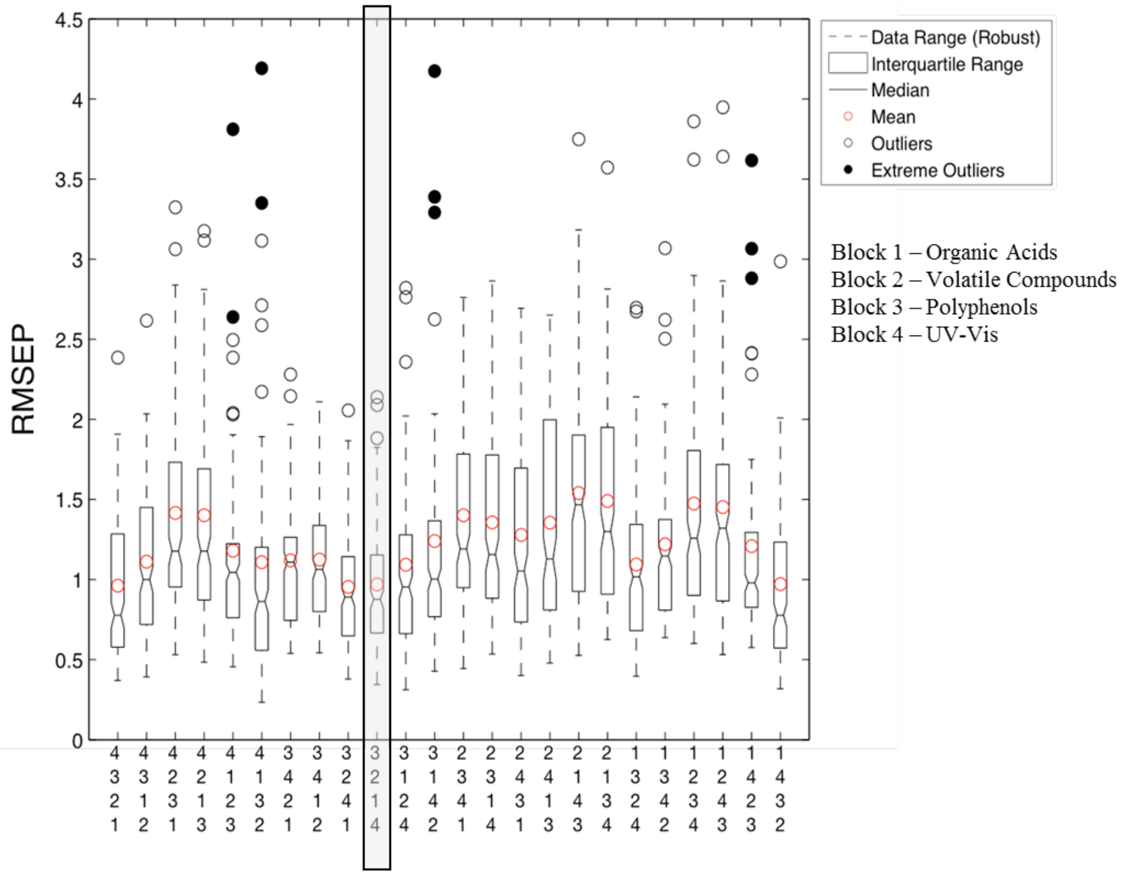


Figure 7-5 – Impact of different block orders in the SO-PLS algorithm prediction capability (as seen by Monte Carlo, RMSEP results). All the 24 possible blocks permutations orders were tested. The highlighted solution (order of the blocks 3-2-1-4) corresponds to be block order selected most frequently by the stepwise SO-PLS approach.

In analytical chemistry applications, such as the one addressed here, information regarding the additional contribution of each analytical method to the analysis and which are the more informative sources of information is very important. This also happens in soft sensor development in industrial settings, where the operational cost is directly related with the number or measurement sources used to build the predictive model. Therefore, a suitable platform is required to appreciate the relevance of using several measurement sources to predict a given phenomenon, as well as the degree of redundancy between these measurement sources for that end, based on which it is possible to determine which measurement sources are really necessary and what are the expected benefits from using each one of them.

Determining the best order of the blocks prior to the application of SO-PLS method is crucial for the utilization of the full prediction and interpretation capabilities of the method, since model outcomes is critically dependent on the order of the blocks. Blocks that are not important for the prediction when included in the model will not necessarily add any significant information to predict the response if added at the end (as currently done). However, if such less informative blocks are used first, they may seem important when analysing the model outcomes, which may be misleading for interpretation purposes. In other words, blocks capturing less information from the response but that can also be explained by other more important blocks, will not have significant importance if included in the model after the more important ones. But if these blocks are used in the first positions, one may end up using all the blocks even though a simpler model without these blocks would lead to the same prediction ability.

Therefore, the efficient application of SO-PLS (particularly in cases where the number of available blocks is larger than 3) requires methods capable of systematically selecting the best order of predictor blocks in an efficient way. The comparison of the new proposed stepwise SO-PLS with the current methodologies available to establish the order of the blocks in a real case example shows that stepwise SO-PLS offers a computationally very efficient solution, leading to one of the best orderings in terms of prediction accuracy and a possibly more parsimonious and robust model. The procedure is simple to programme, fast to execute and easy to interpret, and therefore it is a recommendable approach for implementing SO-PLS, especially in the presence of a large number of blocks with no prior knowledge about the blocks order.

Furthermore, the stepwise logic can also be implemented in situations where some prior knowledge on the order of the blocks exists, that should be enforced in the analysis. This is for example the case of a sequence of unit operations and raw materials where not all blocks are equally important for prediction and hence the method should be able to select the blocks that have significant contributions for the response. In this way, it is possible to find the unit operations that are most contributing to the response and how early in the process can one dispose of good predictions about the product end quality.

7.4 Chapter final remarks

SO-PLS is a state-of-the-art multiblock method, with good predictive and interpretational capabilities (Campos et al., 2017). Its major drawback arises when handling a significant number of **X**-blocks, under the lack of any *a priori* knowledge about their natural order, or when there is not even a natural order to consider. In these cases, the standard approach requires the analysis of all possible block sequence combinations, which is a very time consuming and complex activity. SO-PLS does not provide an efficient way for selecting the relevant blocks for building the model, which is an important limitation for practitioners. Therefore, the stepwise SO-PLS was proposed as an efficient solution to define the order of the blocks to adopt in a SO-PLS application. This method is computationally fast, easy to implement and includes a block selection/exclusion capability, a feature that is not shared by other current multiblock approaches. Being able to perform blocks selection, this method can potentially lead to more robust and parsimonious solutions. Stepwise SO-PLS was applied to a case study consisting of a real data set, and its outcomes compared with those for the standard blocks permutation approach and its new variant, using a robust comparison double cross-validation framework. The results obtained confirm that stepwise SO-PLS is indeed effective in finding an optimal, or close to optimal solution, in spite of being much more efficient and simpler to implement than the current benchmark methodology.

Recently a new multiblock method, related to SO-PLS, was proposed. It is called Response-Oriented Sequential

Alternation (ROSA) and is advocated to be capable of handling a large number of blocks in an efficient way (Liland et al., 2016). ROSA is a fast extension of PLS for multiblock data analysis. This method performs a forward selection of one latent variable of each block at a time, and the blocks can be used several times. This extra flexibility may turn however the interpretation of results more difficult. Stepwise SO-PLS handles each data block separately keeping their natural structure but including only the incremental contributions to the model due to the orthogonalization operations. The sequence of the blocks in ROSA is determined based on the minimum root mean square error of calibration (RMSEC) obtained, while the stepwise SO-PLS selection is based on the RMSECV which is more robust but also more time consuming.

Chapter 8 – Conclusions and future work

Chapter 8. Conclusions and Future Work

Multiblock analysis has the potential to play an important role in many fields of science and modern industry as more and more data are collected that can be naturally grouped into meaningful data blocks according to a valid underlying rational (e.g., different locations in a plant, different phases of a process, different measuring systems, etc.,). Even though these data blocks have their own structure and characteristics they often contribute together by providing complementary information to the system/product under study. Multiblock methods have the ability to integrate into the model *a priori* knowledge from the process/system under study, by including and keeping track of the naturally formed blocks of variables and exploring their inter-linkage.

However, despite the amount of work already done in this field over the last decades and the clear potential benefit of multiblock approaches, they are still underexplored in practice and there are still challenges to be overcame.

The present study represents an effort to push forward the area of multiblock analysis by proposing new and more efficient methods and also by establishing systematic frameworks to help practitioners developing multiblock predictive models in complex Manufacturing 4.0 scenarios (and not only), in an efficient, rigorous and robust way.

Particularly, one major contribution of this thesis is the establishment of a pipeline for the development of multiblock approaches including a three-level approach for pre-processing to guide the users in the application of multiblock modelling (Chapter 5). In the big data era, efficient data cleaning and pre-processing methods are important to extract value and knowledge from the process/system under study. The three-level pre-processing strategy proposed in this thesis consists of addressing intra-block signal-to-noise ratio improvement and feature enhancing (level I), intra-block scaling for equalizing predictive components importance (Level II) and finally moving on to balancing the inter-block effects and making their contributions commensurate in the analysis.

The current available state-of-the-art block scaling methods (grouped in Level III) have the requirement that blocks are previously scaled to unit variance since the scaling factor applied is based on number of predictors. The novel approaches developed in this thesis are able to provide a balanced importance to all blocks independently of the Level I and II pre-processing methods applied. Three groups of block scaling methods were proposed (variance block scaling methods; block rank scaling methods and the pseudorank penalization block scaling method), tested with real case studies, and discussed. These new approaches prevent the introduction of bias into the analysis (which can significantly impact the model outcomes) independently of the pre-processing method previously applied. In other words, they prevent that certain blocks dominate the analysis over other blocks due to size asymmetry effects, different scales or their underlying dimensionality. The work presented in this thesis showed that the selection of the appropriate pre-processing methodology has to be determined case-by-case depending on the data sets at hand, purpose of the analysis and data driven method selected.

A thorough and systematic literature review focused on multiblock modelling approaches and applications is presented followed by an extensive comparison between selected state-of-the-art methods in respect to their relative prediction performances and interpretation capabilities in Chapter 6. Five state-of-the-art methods were included in the study: Concatenated PLS method, Hierarchical PLS (HPLS) (Slama, 1991, Wold et al., 1987b, Wold et al., 1996), Multiblock PLS (MBPLS) (Wangen and Kowalski, 1989, Wold et al., 1983), Network-Induced Supervised Learning (NI-SL) (Reis, 2013b), and Sequential Orthogonalized PLS (SO-PLS) (Jørgensen et al., 2004, Næs et al., 2011b, Jørgensen et al., 2007). Their prediction performances were evaluated by means of a robust statistical framework based on Monte Carlo double cross-validation that was established for this purpose. The study also covers aspects of the multiblock arena which have not been widely explored such as the interpretability capabilities of the models, i.e., the ability of these methods to bring additional interpretational features into the analysis. This framework, based on statistical hypothesis tests, was able to quickly identify the best performing approaches, facilitating a quick comprehensive analysis for a large number of methods. It is therefore a useful tool to support the rational decision making about the methods worthwhile exploring in practice, given a set of goals to be achieved and the data available.

All methods studied in this thesis have different particularities and bring different interpretational features into the analysis and should be selected depending on the data sets at hand, objective of the analysis, and the experience of the user with the methods. **Table *8-1*** below summarizes the main findings from the analysis conducted.

Table 8-1 – Summary of the comparison study of state-of-the-art multiblock methods performed in this thesis.

Method	Prediction	Interpretation	Comments / key features
Concatenated PLS	+++	-	• Easy to implement, independently of number of blocks available. • Block scaling has significant impact the on the performance of the method. • No additional interpretation features, namely regarding inter-blocks relationships.
HPLS	-	+	• Optimization of block scaling method applied required . • Provides tools to explore the global information in a super level and also zooming in into the information from each separate block on the lower data level (local information).
MPLS	+	+	• Optimization of block scaling method applied required.

Method	Prediction	Interpretation	Comments / key features
			• Provides tools to explore the global information in a super level and also zooming in into the information from each separate block on the lower data level (local information). • Two deflation methods are available. In the present case, the two MBPLS deflation approaches show identical relative performance in terms of prediction capability.
NI-SL	++	++	• Analysis of the loadings of the selected latent variables provide useful insights into the system under study. • Allows the identification of the different blocks when this is not known *a priori.*
SO-PLS	+++	+++	• SO-PLS is not affected by the blocks having different variances (scale invariance). • The number of components for each PLS in the model can be defined for each block (independently on the others). • Brings additional interpretation features into the analysis: common and distinct information can be estimated. • It is strongly dependent on the order of the blocks and establishing the sequence of the blocks can be challenging in particular if number of blocks is high.

Finally, in Chapter 7 a new multiblock methodology called Stepwise SO-PLS was also developed, tested, and discussed in this thesis. This method is based on the SO-PLS method and introduces some novelties (e.g., capability of blocks selection), also bringing solutions to some of its current issues (e.g., selection of the order of the blocks). The prediction benefits of this method are similar to the standard SO-PLS but offering the added-value of being more efficient, by adopting a stepwise approach for selecting the best ordering of data blocks (leading to less computational steps hence faster execution time and less computational effort) was well as the possibility of leaving aside blocks that do not bring additional explanatory power.

Future work

A number of challenges still lie ahead to be addressed in future research. The list includes copying with heterogeneous data structures (such as sensors, spectra, images, unstructured data, etc.), non-linearity, non-stationarity, multiscale dynamics, time-delays in the supply-chain, the complex causal network structure, among others. These, and other challenges, may be addressed in future work on multiblock methods in order to make them more capable to help data scientists and engineers taking full advantage of the vast data resources currently available.

The fusion of different types of block sources, other than those originated in analytical instrumentation, is an interesting area for future research. Industrial scenarios, in particular, should be targeted, where complex and heterogeneous data, or multimodal data, is increasingly present, for which multiblock methods show great potential to conduct advanced data analysis and to provide more insights about the underlying processes. Developments in this direction would broaden the applicability of multiblock methods to a wider range of problems.

Furthermore, the integration of multiblock analysis and modern non-linear methods should be researched, to improve the interpretability of deep learning, in the spirit of XAI (eXplainable Artifical Intelligence). As a continuation of the work presented in this thesis, one obvious future perspective work would be in fact to extend the scope of multiblock methods to the non-linear multiblock methods already available in the literature in an extensive comparison framework in order to explore their applicability, key features and weaknesses. In this context, the systematic workflows developed and discussed in this thesis can also be applied for the non-linear multiblock methods.

Another topic to be further investigated is variable selection in the scope of multiblock modeling. This topic was started to be discussed in this thesis with the development of the Stepwise SO-PLS that enables the selection of blocks in the model. The capability of excluding blocks that are not relevant for the prediction of the response under analysis leads to more robust, parsimonious, and interpretable models, which are critical requirements for most end users. In this sense the selection of variables within each block can also bring further added value and therefore is worth to further explore in the future.

Finally, it is also important that, at the same pace as the multiblock research field evolves and new methods are becoming available, new software platforms are developed that implement them, including interactive data visualization tools, where multiblock methods can be implemented and tested, allowing even non-experts to have better comprehension of their data.

References

ACAR, E., PAPALEXAKIS, E. E., GÜRDENIZ, G., RASMUSSEN, M. A., LAWAETZ, A. J., NILSSON, M. & BRO, R. 2014. Structure-revealing data fusion. *BMC bioinformatics,* 15**,** 1-17.

ALINAGHI, M., BERTRAM, H. C., BRUNSE, A., SMILDE, A. K. & WESTERHUIS, J. A. 2020. Common and distinct variation in data fusion of designed experimental data. *Metabolomics,* 16**,** 1-11.

ALTER, O., BROWN, P. O. & BOTSTEIN, D. 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences,* 100**,** 3351-3356.

ANDERSSON, C. A. 1999. Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems,* 47**,** 51-63.

ANDERSSON, M. 2009. A comparison of nine PLS1 algorithms. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 23**,** 518-529.

ARTEAGA, F. & FERRER, A. 2002. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 16**,** 408-418.

BARNES, R., DHANOA, M. S. & LISTER, S. J. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy,* 43**,** 772-777.

BERGLUND, A. & WOLD, S. 1999. A serial extension of multiblock PLS. *Journal of Chemometrics,* 13**,** 461-471.

BIANCOLILLO, A., MÅGE, I. & NÆS, T. 2015. Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemometrics and Intelligent Laboratory Systems,* 141**,** 58-67.

BRÁS, L. G. P., BERNARDINO, S. A., LOPES, J. A. & MENEZES, J. C. 2005. Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour. *Chemometrics and Intelligent Laboratory Systems,* 75**,** 91-99.

BRÁS, L. P., LOPES, J. A., SANTOS, C. R., CARDOSO, J. P. & MENEZES, J. C. 2004. Modelling and identification of individual stage contributions in an industrial pharmaceutical process by multiblock PLS. *Computer Aided Chemical Engineering.* Elsevier.

BREIMAN, L. & SPECTOR, P. 1992. Submodel selection and evaluation in regression. The X-random case. *International statistical review/revue internationale de Statistique***,** 291-319.

BRO, R. 1997. PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems,* 38**,** 149-171.

BRO, R., KJELDAHL, K., SMILDE, A. K. & KIERS, H. 2008. Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry,* 390**,** 1241-1251.

BRO, R. & SMILDE, A. K. 2003. Centering and scaling in component analysis. *Journal of Chemometrics,* 17**,** 16-33.

BROWN, C. D., VEGA-MONTOTO, L. & WENTZELL, P. D. 2000. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Applied Spectroscopy,* 54**,** 1055-1068.

BURNHAM, A. J., MACGREGOR, J. F. & VIVEROS, R. 1999. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems,* 48**,** 167-180.

BURNHAM, A. J., MACGREGOR, J. F. & VIVEROS, R. 2001. Interpretation of regression coefficients under a latent variable regression model. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 15**,** 265-284.

BURNHAM, A. J., VIVEROS, R. & MACGREGOR, J. F. 1996. Frameworks for latent variable multivariate regression. *Journal of chemometrics,* 10**,** 31-45.

CAMPO, E., FERREIRA, V., ESCUDERO, A., MARQUÉS, J. C. & CACHO, J. 2006. Quantitative gas chromatography–olfactometry and chemical quantitative study of the aroma of four Madeira wines. *Analytica Chimica Acta,* 563**,** 180-187.

CAMPOS, M. P., SOUSA, R., PEREIRA, A. C. & REIS, M. S. 2017. Advanced predictive methods for wine age prediction: Part II–A comparison study of multiblock regression approaches. *Talanta,* 171**,** 132-142.

CARIOU, V., BOUVERESSE, D. J.-R., QANNARI, E. M. & RUTLEDGE, D. 2019. ComDim methods for the analysis of multiblock data in a data fusion perspective. *Data Handling in Science and Technology.* Elsevier.

CARIOU, V., QANNARI, E. M., RUTLEDGE, D. N. & VIGNEAU, E. 2018. ComDim: from multiblock data analysis to path modeling. *Food Quality and Preference,* 67**,** 27-34.

CHEN, G. & MCAVOY, T. J. 1998. Predictive on-line monitoring of continuous processes. *Journal of Process Control,* 8**,** 409-420.

CHIANG, L. H., PELL, R. J. & SEASHOLTZ, M. B. 2003. Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control,* 13**,** 437-449.

CHRISTIN, C., SMILDE, A. K., HOEFSLOOT, H. C., SUITS, F., BISCHOFF, R. & HORVATOVICH, P. L. 2008. Optimized time alignment algorithm for LC− MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Analytical chemistry,* 80**,** 7012-7021.

CLIFFORD, D., STONE, G., MONTOLIU, I., REZZI, S., MARTIN, F.-P., GUY, P., BRUCE, S. & KOCHHAR, S. 2009. Alignment using variable penalty dynamic time warping. *Analytical chemistry,* 81**,** 1000-1007.

CROCKFORD, D. J., HOLMES, E., LINDON, J. C., PLUMB, R. S., ZIRAH, S., BRUCE, S. J., RAINVILLE, P., STUMPF, C. L. & NICHOLSON, J. K. 2006. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. *Analytical Chemistry,* 78**,** 363-371.

DE JONG, S. 1993. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems,* 18**,** 251-263.

DE MEYER, T., SINNAEVE, D., VAN GASSE, B., TSIPORKOVA, E., RIETZSCHEL, E. R., DE BUYZERE, M. L., GILLEBERT, T. C., BEKAERT, S., MARTINS, J. C. & VAN CRIEKINGE, W. 2008. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical chemistry,* 80**,** 3783-3790.

DE NOORD, O. E. 1994. The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and intelligent laboratory systems,* 23**,** 65-70.

DRAPER, N. R. & SMITH, H. 1998. *Applied regression analysis*, John Wiley & Sons.

DUCHESNE, C. & MACGREGOR, J. F. 2000. Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemometrics and Intelligent Laboratory Systems,* 51**,** 125-137.

EILERS, P. & BOELENS, H. 2005. Leiden University Medical Centre Report. *Leiden University Medical Centre.*

EILERS, P. H. 2004. Parametric time warping. *Analytical chemistry,* 76**,** 404-411.

EL GHAZIRI, A., CARIOU, V., RUTLEDGE, D. N. & QANNARI, E. M. 2016. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K+ 1) datasets. *Journal of Chemometrics,* 30**,** 420-429.

ENGEL, J., GERRETZEN, J., SZYMAŃSKA, E., JANSEN, J. J., DOWNEY, G., BLANCHET, L. & BUYDENS, L. M. 2013. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry,* 50**,** 96-106.

ERIKSSON, L. 1999. *Introduction to multi-and megavariate data analysis using projection methods (PCA & PLS)*, Umetrics AB.

ERIKSSON, L., BYRNE, T., JOHANSSON, E., TRYGG, J. & VIKSTRÖM, C. 2013. *Multi-and megavariate data analysis basic principles and applications*, Umetrics Academy.

ERIKSSON, S. W. & TRYG, J. n.d. O2PLS® for improved analysis and visualization of complex data.

ESQUERRE, C., GOWEN, A., BURGER, J., DOWNEY, G. & O'DONNELL, C. 2012. Suppressing sample morphology effects in near infrared spectral imaging using chemometric data pre-treatments. *Chemometrics and Intelligent Laboratory Systems,* 117**,** 129-137.

FAMILI, A., SHEN, W.-M., WEBER, R. & SIMOUDIS, E. 1997. Data preprocessing and intelligent data analysis. *Intelligent data analysis,* 1**,** 3-23.

FEARN, T. 2000. On orthogonal signal correction. *Chemometrics and intelligent laboratory systems,* 50**,** 47-52.

FELÍCIO, C. C., BRÁS, L. P., LOPES, J. A., CABRITA, L. & MENEZES, J. C. 2005. Comparison of PLS algorithms in gasoline and gas oil parameter monitoring with MIR and NIR. *Chemometrics and Intelligent Laboratory Systems,* 78**,** 74-80.

FLORES-CERRILLO, J. & MACGREGOR, J. F. 2004. Multivariate monitoring of batch processes using batch-to-batch information. *AIChE Journal,* 50**,** 1219-1228.

FRANK, I., FEIKEMA, J., CONSTANTINE, N. & KOWALSKI, B. 1984. Prediction of product quality from spectral data using the partial least-squares method. *Journal of Chemical Information and Computer Sciences,* 24**,** 20-24.

FRANK, I. & KOWALSKI, B. 1985. A multivariate method for relating groups of measurements connected by a causal pathway. *Analytica Chimica Acta,* 167**,** 51-63.

FRANK, I. & KOWALSKI, B. R. 1984. Prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling. *Analytica Chimica Acta,* 162**,** 241-251.

GARCÍA-MUÑOZ, S. & POLIZZI, M. 2012. WSPLS—A new approach towards mixture modeling and accelerated product development. *Chemometrics and Intelligent Laboratory Systems,* 114**,** 116-121.

GAYDOU, V., KISTER, J. & DUPUY, N. 2011. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. *Chemometrics and Intelligent Laboratory Systems,* 106**,** 190-197.

GAYNANOVA, I. & LI, G. 2019. Structural learning and integrative decomposition of multi-view data. *Biometrics,* 75**,** 1121-1132.

GELADI, P. & KOWALSKI, B. R. 1986. Partial least-squares regression: a tutorial. *Analytica chimica acta,* 185**,** 1-17.

GELADI, P., MACDOUGALL, D. & MARTENS, H. 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied spectroscopy,* 39**,** 491-500.

GERLACH, R. W., KOWALSKI, B. R. & WOLD, H. O. 1979. Partial least-squares path modelling with latent variables. *Analytica Chimica Acta,* 112**,** 417-421.

GERRETZEN, J., SZYMAŃSKA, E., JANSEN, J. J., BART, J., VAN MANEN, H.-J., VAN DEN HEUVEL, E. R. & BUYDENS, L. M. 2015. Simple and effective way for data preprocessing selection based on design of experiments. *Analytical chemistry,* 87**,** 12096-12103.

GUO, Q., WU, W. & MASSART, D. 1999. The robust normal variate transform for pattern recognition with near-infrared data. *Analytica chimica acta,* 382**,** 87-103.

HANAFI, M., KOHLER, A. & QANNARI, E. M. 2010. Shedding new light on hierarchical principal component analysis. *Journal of chemometrics,* 24**,** 703-709.

HANAFI, M., MAZEROLLES, G., DUFOUR, E. & QANNARI, E. 2006. Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 20**,** 172-183.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. Unsupervised learning. *The elements of statistical learning.* Springer.

HELLTON, K. H. & THORESEN, M. 2016. Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics,* 17**,** 537-548.

HENDRIKS, M. M., VAN EEUWIJK, F. A., JELLEMA, R. H., WESTERHUIS, J. A., REIJMERS, T. H., HOEFSLOOT, H. C. & SMILDE, A. K. 2011. Data-processing strategies for metabolomics studies. *TrAC Trends in Analytical Chemistry,* 30**,** 1685-1698.

HERTRAMPF, A., SOUSA, R., MENEZES, J. & HERDLING, T. 2016. Semi-quantitative prediction of a multiple API solid dosage form with a combination of vibrational spectroscopy methods. *Journal of Pharmaceutical and Biomedical Analysis,* 124**,** 246-253.

JACKSON, J. 1991. A User's Guide to Principal Components; John & Wiley. *New York.*

JACKSON, J. E. 2005. *A user's guide to principal components*, John Wiley & Sons.

JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2013. *An introduction to statistical learning*, Springer.

JOE QIN, S. 2003. Statistical process monitoring: basics and beyond. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 17**,** 480-502.

JOLLIFFE, I. 2005. Principal component analysis. *Encyclopedia of statistics in behavioral science.*

JOLLIFFE, I. T. 2002. Graphical representation of data using principal components. *Principal component analysis***,** 78-110.

JØRGENSEN, K., MEVIK, B.-H. & NÆS, T. 2007. Combining designed experiments with several blocks of spectroscopic data. *Chemometrics and intelligent laboratory systems,* 88**,** 154-166.

JØRGENSEN, K., SEGTNAN, V., THYHOLT, K. & NÆS, T. 2004. A comparison of methods for analysing regression models with both spectral and designed variables. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 18**,** 451-464.

KEUN, H. C., EBBELS, T. M., ANTTI, H., BOLLARD, M. E., BECKONERT, O., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2003. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Analytica chimica acta,* 490**,** 265-276.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai, 1995. Montreal, Canada, 1137-1145.

KOURTI, T. & MACGREGOR, J. F. 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and intelligent laboratory systems,* 28**,** 3-21.

KROONENBERG, P. M. & DE LEEUW, J. 1980. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika,* 45**,** 69-97.

KULIGOWSKI, J., PÉREZ-GUAITA, D., SÁNCHEZ-ILLANA, Á., LEÓN-GONZÁLEZ, Z., DE LA GUARDIA, M., VENTO, M., LOCK, E. F. & QUINTÁS, G. 2015. Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE). *Analyst,* 140**,** 4521-4529.

KVALHEIM, O. M., BRAKSTAD, F. & LIANG, Y. 1994. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry,* 66**,** 43-51.

LAHAT, D., ADALI, T. & JUTTEN, C. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE,* 103**,** 1449-1477.

LEE, D. S. & VANROLLEGHEM, P. A. 2003. Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. *Biotechnology and Bioengineering,* 82**,** 489-497.

LI, W. 2014. *Risk assessment of power systems: models, methods, and applications*, John Wiley & Sons.

LIEBER, C. A. & MAHADEVAN-JANSEN, A. 2003. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied spectroscopy,* 57**,** 1363-1367.

LILAND, K. H., NÆS, T. & INDAHL, U. G. 2016. ROSA—a fast extension of partial least squares regression for multiblock data analysis. *Journal of Chemometrics,* 30**,** 651-662.

LOCK, E. F., HOADLEY, K. A., MARRON, J. S. & NOBEL, A. B. 2013. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics,* 7**,** 523.

LÖFSTEDT, T. 2012. *OnPLS: Orthogonal projections to latent structures in multiblock and path model data analysis.* Umeå universitet.

LOPES, J., MENEZES, J., WESTERHUIS, J. & SMILDE, A. 2002. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnology and bioengineering,* 80**,** 419-427.

MÅGE, I., MENICHELLI, E. & NÆS, T. 2012. Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food quality and preference,* 24**,** 8-16.

MÅGE, I., MEVIK, B. H. & NÆS, T. 2008. Regression models with process variables and parallel blocks of raw material measurements. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 22**,** 443-456.

MÅGE, I., SMILDE, A. K. & VAN DER KLOET, F. M. 2019. Performance of methods that separate common and distinct variation in multiple data blocks. *Journal of Chemometrics,* 33**,** e3085.

MARTENS, H. & RUSSWURM JR, H. 1982. Food research and data analysis; proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway.

MARTENS, H. & STARK, E. 1991. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *Journal of pharmaceutical and biomedical analysis,* 9**,** 625-635.

MISHRA, P., ROGER, J.-M., JOUAN-RIMBAUD-BOUVERESSE, D., BIANCOLILLO, A., MARINI, F., NORDON, A. & RUTLEDGE, D. N. 2021. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends in Analytical Chemistry,* 137**,** 116206.

MØLLER, S. F., VON FRESE, J. & BRO, R. 2005. Robust methods for multivariate data analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 19**,** 549-563.

MOSIER, C. I. 1951. The need and means of cross validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement.*

MUTEKI, K., MACGREGOR, J. F. & UEDA, T. 2006. Rapid development of new polymer blends: the optimal selection of materials and blend ratios. *Industrial & engineering chemistry research,* 45**,** 4653-4660.

NÆS, T., BROCKHOFF, P. B. & TOMIC, O. 2011a. *Statistics for sensory and consumer science*, John Wiley & Sons.

NAES, T., TOMIC, O., AFSETH, N. K., SEGTNAN, V. & MÅGE, I. 2013. Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems,* 124**,** 32-42.

NÆS, T., TOMIC, O., MEVIK, B. H. & MARTENS, H. 2011b. Path modelling by sequential PLS regression. *Journal of Chemometrics,* 25**,** 28-40.

NELSON, P. R., TAYLOR, P. A. & MACGREGOR, J. F. 1996. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and intelligent laboratory systems,* 35**,** 45-65.

NOGUCHI, Y., ZHANG, Q.-W., SUGIMOTO, T., FURUHATA, Y., SAKAI, R., MORI, M., TAKAHASHI, M. & KIMURA, T. 2006. Network analysis of plasma and tissue amino acids and the generation of an amino index for potential diagnostic use. *The American Journal of Clinical Nutrition,* 83**,** 513S-519S.

NOMIKOS, P. & MACGREGOR, J. F. 1994. Monitoring batch processes using multiway principal component analysis. *AIChE Journal,* 40**,** 1361-1375.

NORRIS, G. & JUX, U. 1984. Fine wall structure of selected Upper Jurassic gonyaulacystinean dinoflagellate cysts from southern England. *Palaeontographica Abteilung B Paläophytologie,* 190**,** 158-168.

P. CAMPOS, M., SOUSA, R. & S. REIS, M. 2018. Establishing the optimal blocks' order in SO-PLS: Stepwise SO-PLS and alternative formulations. *Journal of Chemometrics,* 32**,** e3032.

PEREIRA, A. C., CARVALHO, M. J., MIRANDA, A., LEÇA, J. M., PEREIRA, V., ALBUQUERQUE, F., MARQUES, J. C. & REIS, M. S. 2016. Modelling the ageing process: A novel strategy to analyze the wine evolution towards the expected features. *Chemometrics and Intelligent Laboratory Systems,* 154**,** 176-184.

PEREIRA, A. C., REIS, M. S., SARAIVA, P. M. & MARQUES, J. C. 2010a. Analysis and assessment of Madeira wine ageing over an extended time period through GC–MS and chemometric analysis. *Analytica Chimica Acta,* 660**,** 8-21.

PEREIRA, V., CÂMARA, J. S., CACHO, J. & MARQUES, J. C. 2010b. HPLC-DAD methodology for the quantification of organic acids, furans and polyphenols by direct injection of wine samples. *Journal of separation science,* 33**,** 1204-1215.

POLIZZI, M. A. & GARCÍA-MUÑOZ, S. 2011. A framework for in-silico formulation design using multivariate latent variable regression methods. *International journal of pharmaceutics,* 418**,** 235-242.

QANNARI, E. M., WAKELING, I., COURCOUX, P. & MACFIE, H. J. 2000. Defining the underlying sensory dimensions. *Food Quality and Preference,* 11**,** 151-154.

QIN, S. J., VALLE, S. & PIOVOSO, M. J. 2001. On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 15**,** 715-742.

RAMAKER, H.-J., VAN SPRANG, E. N., WESTERHUIS, J. A. & SMILDE, A. K. 2003. Dynamic time warping of spectroscopic BATCH data. *Analytica Chimica Acta,* 498**,** 133-153.

RAMAKER, H.-J., VAN SPRANG, E. N., WESTERHUIS, J. A. & SMILDE, A. K. 2005. Fault detection properties of global, local and time evolving models for batch process monitoring. *Journal of Process control,* 15**,** 799-805.

RÄNNAR, S., LINDGREN, F., GELADI, P. & WOLD, S. 1994. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Journal of Chemometrics,* 8**,** 111-125.

RÄNNAR, S., MACGREGOR, J. F. & WOLD, S. 1998. Adaptive batch monitoring using hierarchical PCA. *Chemometrics and intelligent laboratory systems,* 41**,** 73-81.

REIS, M., SARAIVA, P. & BAKSHI, B. 2009. Denoising and signal-to-noise ratio enhancement: wavelet transform and Fourier transform.

REIS, M. S. 2013a. Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables. *Chemometrics and Intelligent Laboratory Systems,* 127**,** 7-16.

REIS, M. S. 2013b. Network-induced supervised learning: Network-induced classification (NI-C) and network-induced regression (NI-R). *AIChE Journal,* 59**,** 1570-1587.

REIS, M. S. 2019. Multiscale and multi-granularity process analytics: A review. *Processes,* 7**,** 61.

REIS, M. S. & GAO, F. 2021. Special Issue "Advanced Process Monitoring for Industry 4.0". MDPI.

REIS, M. S. & GINS, G. 2017. Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. *Processes,* 5**,** 35.

REIS, M. S., GINS, G. & RATO, T. J. 2019. Incorporation of process-specific structure in statistical process monitoring: A review. *Journal of Quality Technology,* 51**,** 407-421.

REIS, M. S. & KENETT, R. 2018. Assessing the value of information of data-centric activities in the chemical processing industry 4.0. *AIChE Journal,* 64**,** 3868-3881.

REIS, M. S., RENDALL, R., RATO, T. J., MARTINS, C. & DELGADO, P. 2021. Improving the sensitivity of statistical process monitoring of manifolds embedded in high-dimensional spaces: The truncated-Q statistic. *Chemometrics and Intelligent Laboratory Systems,* 215**,** 104369.

REIS, M. S. & SARAIVA, P. M. 2019. Data-Centric Process Systems Engineering for the Chemical Industry 4.0. *Systems Engineering in the Fourth Industrial Revolution***,** 137-159.

RENDALL, R., CHIANG, L. H. & REIS, M. S. 2019. Data-driven methods for batch data analysis–A critical overview and mapping on the complexity scale. *Computers & Chemical Engineering,* 124**,** 1-13.

RENDALL, R., PEREIRA, A. C. & REIS, M. S. 2017. Advanced predictive methods for wine age prediction: Part I–A comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods. *Talanta,* 171**,** 341-350.

RENDALL, R. R. & REIS, M. S. 2014. A Comparison Study of Single-Scale and Multiscale Approaches for Data-Driven and Model-Based Online Denoising. *Quality and Reliability Engineering International,* 30**,** 935-950.

RINNAN, Å., VAN DEN BERG, F. & ENGELSEN, S. B. 2009. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry,* 28**,** 1201-1222.

ROGER, J.-M., CHAUCHARD, F. & BELLON-MAUREL, V. 2003. EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemometrics and Intelligent Laboratory Systems,* 66**,** 191-204.

RUDNITSKAYA, A., ROCHA, S., LEGIN, A., PEREIRA, V. & MARQUES, J. C. 2010. Evaluation of the feasibility of the electronic tongue as a rapid analytical tool for wine age prediction and quantification of the organic acids and phenolic compounds. The case-study of Madeira wine. *Analytica Chimica Acta,* 662**,** 82-89.

SARRAGUÇA, M. C., CRUZ, A. V., AMARAL, H. R., COSTA, P. C. & LOPES, J. A. 2011. Comparison of different chemometric and analytical methods for the prediction of particle size distribution in pharmaceutical powders. *Analytical and bioanalytical chemistry,* 399**,** 2137-2147.

SAVITZKY, A. & GOLAY, M. J. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry,* 36**,** 1627-1639.

SAVORANI, F., TOMASI, G. & ENGELSEN, S. B. 2010. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of magnetic resonance,* 202**,** 190-202.

SCHMIDT, S., SCHINDLER, M. & ERIKSSON, L. 2021. Block-wise exploration of molecular descriptors with Multi-block Orthogonal Component Analysis (MOCA). *Molecular Informatics***,** 2100165.

SCHOUTEDEN, M., VAN DEUN, K., PATTYN, S. & VAN MECHELEN, I. 2013. SCA with rotation to distinguish common and distinctive information in linked data. *Behavior research methods,* 45**,** 822-833.

SCHOUTEDEN, M., VAN DEUN, K., WILDERJANS, T. F. & VAN MECHELEN, I. 2014. Performing DISCO-SCA to search for distinctive and common information in linked data. *Behavior research methods,* 46**,** 576-587.

SKOV, T., BALLABIO, D. & BRO, R. 2008. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Analytica chimica acta,* 615**,** 18-29.

SKOV, T., VAN DEN BERG, F., TOMASI, G. & BRO, R. 2006. Automated alignment of chromatographic data. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 20**,** 484-497.

SLAMA, F. C. 1991. *Multivariate statistical analysis of data from an industrial fluidized catalytic cracking process using PCA and PLS.*

SLIŠKOVIĆ, D., GRBIĆ, R. & NYARKO, E. K. 2009. Data preprocessing in data based process modeling. *IFAC Proceedings Volumes,* 42**,** 559-564.

SMILDE, A. K., MÅGE, I., NAES, T., HANKEMEIER, T., LIPS, M. A., KIERS, H. A., ACAR, E. & BRO, R. 2017. Common and distinct components in data fusion. *Journal of Chemometrics,* 31**,** e2900.

SMILDE, A. K., VAN DER WERF, M. J., BIJLSMA, S., VAN DER WERFF-VAN DER VAT, B. J. & JELLEMA, R. H. 2005. Fusion of mass spectrometry-based metabolomics data. *Analytical chemistry,* 77**,** 6729-6736.

SMILDE, A. K., WESTERHUIS, J. A. & BOQUE, R. 2000. Multiway multiblock component and covariates regression models. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 14**,** 301-331.

SONG, Y., WESTERHUIS, J. A. & SMILDE, A. K. 2020. Separating common (global and local) and distinct variation in multiple mixed types data sets. *Journal of Chemometrics,* 34**,** e3197.

TAULER, R., KOWALSKI, B. & FLEMING, S. 1993. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical chemistry,* 65**,** 2040-2047.

TORGRIP, R., ÅBERG, K., ALM, E., SCHUPPE-KOISTINEN, I. & LINDBERG, J. 2008. A note on normalization of biofluid 1D 1 H-NMR data. *Metabolomics,* 4**,** 114-121.

TRYGG, J. 2002. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 16**,** 283-293.

VAN DEN BERG, R. A., HOEFSLOOT, H. C., WESTERHUIS, J. A., SMILDE, A. K. & VAN DER WERF, M. J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics,* 7**,** 1-15.

VAN DER KLOET, F. M., SEBASTIÁN-LEÓN, P., CONESA, A., SMILDE, A. K. & WESTERHUIS, J. A. 2016. Separating common from distinctive variation. *BMC bioinformatics,* 17**,** 271-286.

VAN DEUN, K., VAN MECHELEN, I., THORREZ, L., SCHOUTEDEN, M., DE MOOR, B., VAN DER WERF, M. J., DE LATHAUWER, L., SMILDE, A. K. & KIERS, H. A. 2012. DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes. *PloS one,* 7**,** e37840.

VAN MECHELEN, I. & SMILDE, A. K. 2010. A generic linked-mode decomposition model for data fusion. *Chemometrics and Intelligent Laboratory Systems,* 104**,** 83-94.

WALCZAK, B. & MASSART, D. L. 2001. Dealing with missing data: Part II. *Chemometrics and Intelligent Laboratory Systems,* 58**,** 29-42.

WANGEN, L. & KOWALSKI, B. 1989. A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of chemometrics,* 3**,** 3-20.

WESTERHUIS, J. A. 1997. *Multivariate statistical modelling of the pharmaceutical process of wet granulation and tableting*, [University Library Groningen][Host].

WESTERHUIS, J. A. & COENEGRACHT, P. M. 1997. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 11**,** 379-392.

WESTERHUIS, J. A., DE JONG, S. & SMILDE, A. K. 2001. Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems,* 56**,** 13-25.

WESTERHUIS, J. A., KOURTI, T. & MACGREGOR, J. F. 1998. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 12**,** 301-321.

WESTERHUIS, J. A. & SMILDE, A. K. 2001. Deflation in multiblock PLS. *Journal of Chemometrics: A Journal of the Chemometrics Society,* 15**,** 485-493.

WOLD, H. 1982. Soft modeling: the basic design and some extensions. *Systems under indirect observation,* 2**,** 343.

WOLD, S. 1976. Pattern recognition by means of disjoint principal components models. *Pattern recognition,* 8**,** 127-139.

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

WOLD, S. 1987. *PLS modeling with latent variables in two or more dimensions*, Verlag nicht ermittelbar.

WOLD, S., ALBANO, C., DUNN, W., EDLUND, U., ESBENSEN, K., GELADI, P., HELLBERG, S., JOHANSSON, E., LINDBERG, W. & SJÖSTRÖM, M. 1984. Multivariate data analysis in chemistry. *Chemometrics.* Springer.

WOLD, S., ANTTI, H., LINDGREN, F. & ÖHMAN, J. 1998a. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent laboratory systems,* 44**,** 175-185.

WOLD, S., ESBENSEN, K. & GELADI, P. 1987a. Principal component analysis. *Chemometrics and intelligent laboratory systems,* 2**,** 37-52.

WOLD, S., HELLBERG, S., LUNDSTEDT, T., SJOSTROM, M. & WOLD, H. 1987b. Proc. Symp. on PLS model building: theory and application. *Frankfurt am Main.*

WOLD, S., KETTANEH, N., FRIDÉN, H. & HOLMBERG, A. 1998b. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and intelligent laboratory systems,* 44**,** 331-340.

WOLD, S., KETTANEH, N. & TJESSEM, K. 1996. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics,* 10**,** 463-482.

WOLD, S., MARTENS, H. & WOLD, H. 1983. The multivariate calibration problem in chemistry solved by the PLS method. *Matrix pencils.* Springer.

WOLD, S., SJÖSTRÖM, M. & ERIKSSON, L. 2001a. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems,* 58**,** 109-130.

WOLD, S., TRYGG, J., BERGLUND, A. & ANTTI, H. 2001b. Some recent developments in PLS modeling. *Chemometrics and intelligent laboratory systems,* 58**,** 131-150.

WU, W., DASZYKOWSKI, M., WALCZAK, B., SWEATMAN, B. C., CONNOR, S. C., HASELDEN, J. N., CROWTHER, D. J., GILL, R. W. & LUTZ, M. W. 2006. Peak alignment of urine NMR spectra using fuzzy warping. *Journal of Chemical Information and Modeling,* 46**,** 863-875.

YI, L., DONG, N., YUN, Y., DENG, B., REN, D., LIU, S. & LIANG, Y. 2016. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. *Analytica chimica acta,* 914**,** 17-34.

YIN, S., DING, S. X., XIE, X. & LUO, H. 2014. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial electronics,* 61**,** 6418-6428.

ZHU, J., GE, Z., SONG, Z. & GAO, F. 2018. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control,* 46**,** 107-133.